



Some results from Open Call #2: Genomic prediction for strawberry powdery mildew resistance/susceptibility

Beneficiaries:

Njos Fruit and Berry Center, Norway

Fresh Forward, The Netherlands





Experiment Setups at Njøs and Fresh Forward

Njøs Fruit and Berry Centre – August 2021

- A **3 × 5 North Carolina II crossing design** was established (factorial scheme). The field was set up with raised beds, plastic mulch, and drip irrigation.
- A total of **356 accessions**, each represented by 10 clonal runner plants, were planted in a **720-plot randomized complete block design**. The remaining 8 plots were filled with parent cultivars.

Fresh Forward – Fall 2022 and 2023

- The phenotyping experiments were conducted in a **greenhouse setting**.
- Each year, six plants per accession were planted from **760 advanced breeding lines and cultivars** in an **unreplicated layout**, with control accessions regularly spaced throughout the design.

Factorial Field Design and Trait Evaluation at Njøs Fruit and Berry Center

- **Factorial design (NCII)** with 349 clones, tested in two replicates.

Example of clone distribution across parent cultivars:

	Murano	Nobel	Rondo	Rumba	Saga
Camarosa	25	25	6	25	25
Glima	25	25	25	25	25
Honeoye	25	25	18	25	25

- **Powdery mildew scored 5 times per season over two years.**
- Additional traits evaluated: **yield, fruit quality**, and a range of morphological characteristics.



Photo: Kristina A G



Photo: Fresh Forward

Disease scoring and aggregate calculations

The disease scoring was done with the ordinal five-point scale defined by Simpson (1987):

Powdery Mildew Scoring Scale:

1= No symptoms – Leaves appear healthy with no visible signs of infection.

2= Mild symptoms – Slight leaf curling; no visible mycelial growth.

3= Moderate symptoms – Noticeable leaf curling and mottling.

4= Severe symptoms – Pronounced curling, reddening, and visible damage on the underside of leaves.

5= Very severe symptoms – Extensive necrosis and some leaf death.

- Each plot in the experiments were scored bi-weekly from the end of May until Mid-July at **Njos** and weekly from Mid-August to Mid-September at **Fresh Forward**.
- Five consecutive scores were used to calculate the aggregates **Area Under The Disease Progress Curve (AUDPC)** and **Area Under The Disease Progress Slope (AUDPS)** as defined by Simko and Piepho (2012).

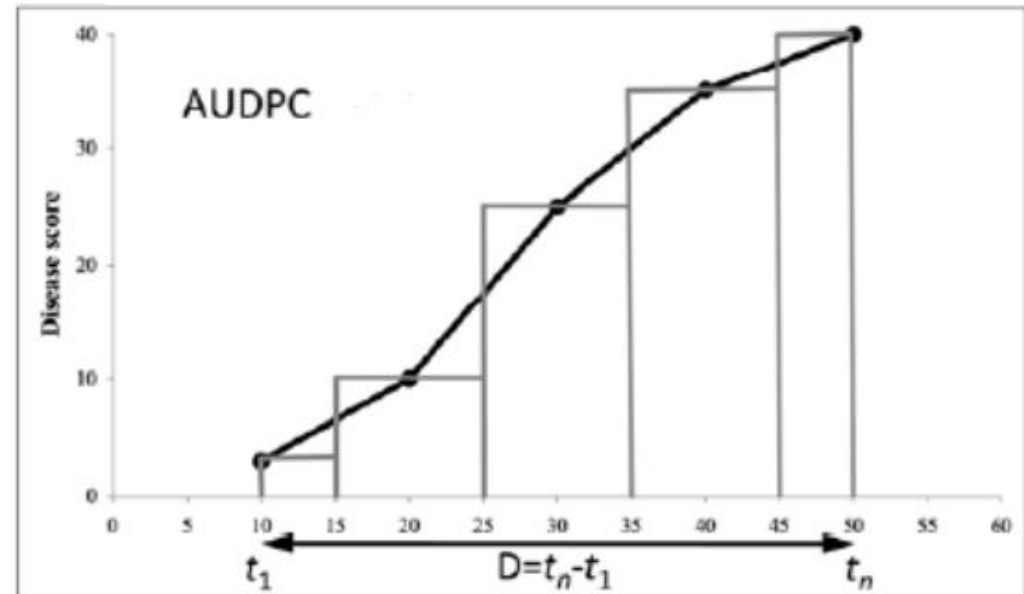


Calculating AUDPC: Area Under the Disease Progress Curve

- **AUDPC** summarizes the total disease intensity over time by integrating disease scores across multiple observations.
- It is calculated using the trapezoidal method, combining both severity and timing.
- The formula estimates **area under the curve** by summing trapezoids between each time point
- Useful for comparing disease progression between genotypes or treatments

$$AUDPC = \sum_{i=1}^{n-1} \frac{y_i + y_{i+1}}{2} \times (t_{i+1} - t_i)$$

- y_i : disease score at time t_i
- t_i : time of observation



From BLUEs to GWAS and Genomic Prediction: a two-step approach

•Step 1:

- Estimated **BLUEs** (Best Linear Unbiased Estimators) for all entries using **ASReml-R**.
- Entries treated as **fixed effects**, while factors like year, replicate, and interactions were treated as **random effects**.

•Step 2 – GWAS:

- Used **BLUEs** as input.
- Applied **one single-locus model**: SUPER.
- Applied **two multiple-loci models**: FarmCPU and BLINK.

•Step 2 – Genomic Prediction (GP):

- BLUEs used to predict phenotypes using **ASReml-R** (baseline model).
- Combined **phenotypes (BLUPs)** with **genotypic data (SNPs)**.
- Included **pedigree information** when available (e.g., Njøs dataset).

Genomic Selection Model – The Prediction Equation


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

- y_{ijk} : Observed phenotype
- μ : Overall mean
- $PC_{hi} \gamma_h$: Fixed effect of principal components (population structure)
- b_j : Block effect
- g_i : Genomic effect
- ge_{ij} : Genotype-by-environment interaction
- ε_{ijk} : Residual error

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


Comparing Broad Sense Heritability of Powdery Mildew Resistance: Fresh Forward vs. Njø

Broad sense heritabilities of broad sense heritability (H^2) and their standard errors from greenhouse experiments with natural infection in three experiments at **Fresh Forward** (2022 jb: junebearers in 2022, 2023 jb: junebearers in 2023, and 2023 eb: everbearers in 2023).



Trait	All	2022	2023 <u>eb</u>	2023 <u>jb</u>
<u>Audpc</u>	0.53 ± 0.04	0.51 ± 0.04	0.58 ± 0.10	0.77 ± 0.03
<u>Audps</u>	0.53 ± 0.04	0.52 ± 0.04	0.56 ± 0.10	0.77 ± 0.03
Endpoint	0.42 ± 0.04	0.50 ± 0.05	0.47 ± 0.14	0.53 ± 0.06
Mean	0.50 ± 0.04	0.56 ± 0.04	0.58 ± 0.10	0.78 ± 0.03

Broad sense heritabilities of broad sense heritability (H^2) and their standard errors from a field experiment at **Njos Fruit** and Berry Center in the fall 2021, scored for powdery mildew during the summer of 2022 and again in the summer of 2023.

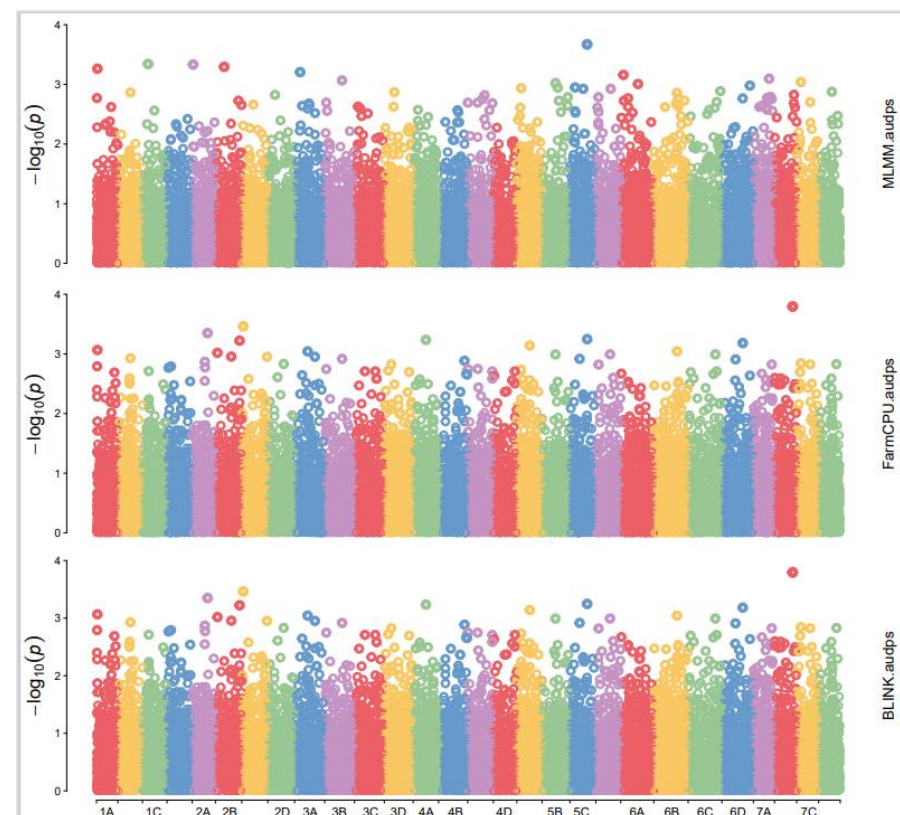
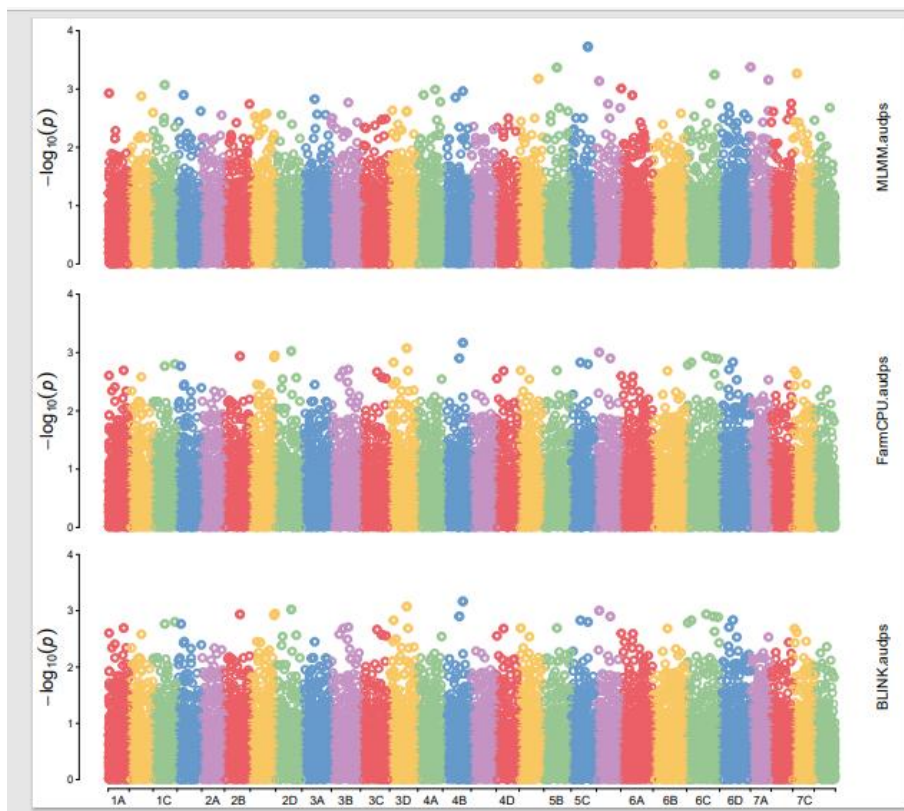


Trait	Both years	2022	2023
<u>Audpc</u>	0.74 ± 0.02	0.82 ± 0.02	0.81 ± 0.02
<u>Audps</u>	0.75 ± 0.02	0.84 ± 0.02	0.82 ± 0.02
Endpoint	0.54 ± 0.04	0.64 ± 0.04	0.69 ± 0.03
Mean	0.74 ± 0.03	0.83 ± 0.02	0.82 ± 0.02

GWAS using various statistical approaches - Njøs data. (MLMM, FarmCPU, BLINK).

2022

2023



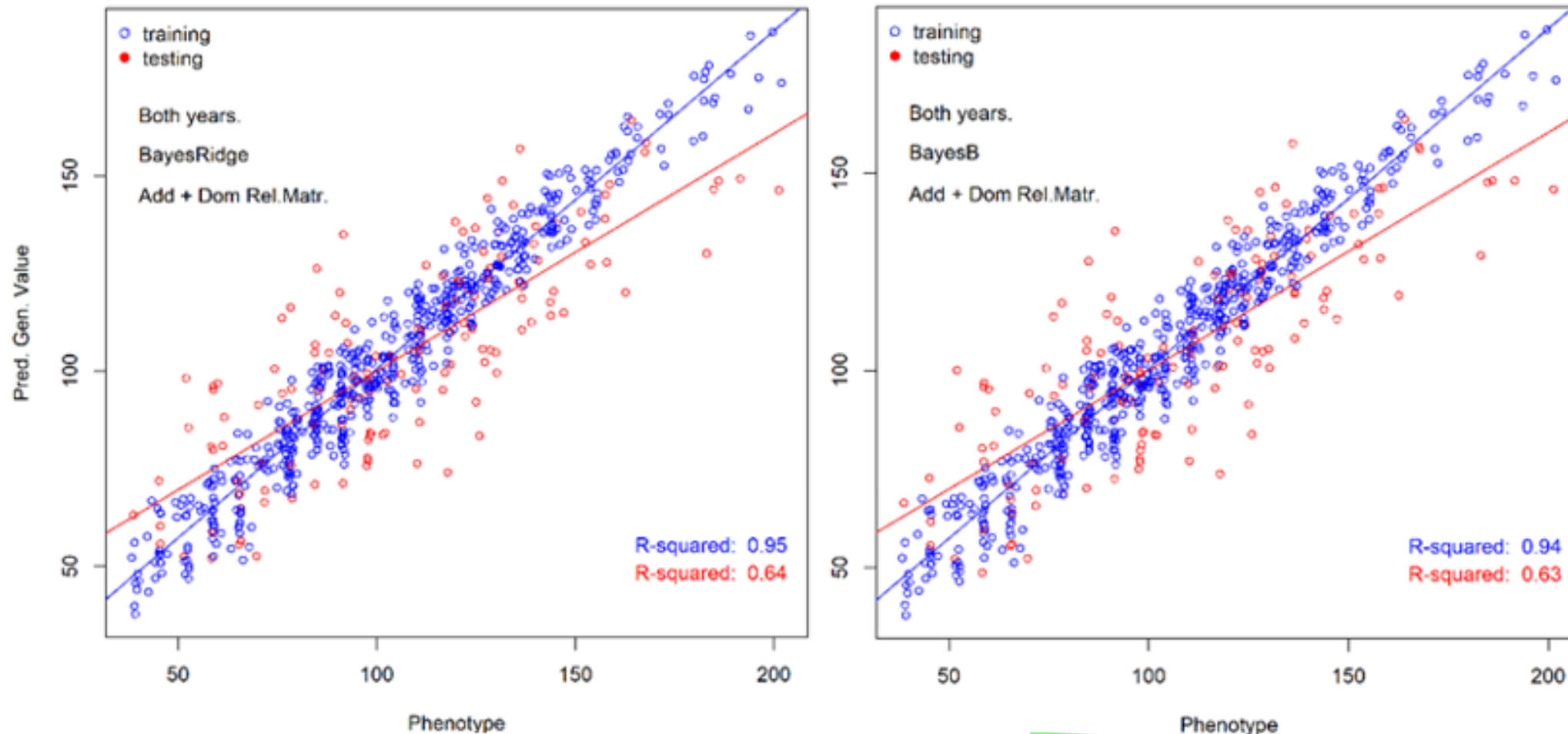
9 GWAS: No signals observed

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!

Comparing Genomic Prediction Models for Powdery Mildew Resistance at Njøs

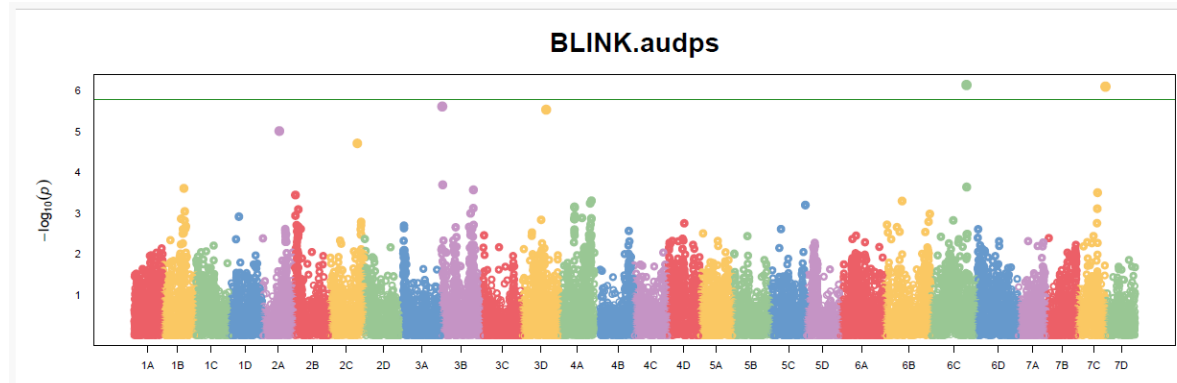


Scatter plots of a random fold in the cross-validation of the Njøs powdery mildew scores over the two experimental years. Two methods – Ridge regression (left) and BayesB (right) were researched. The additive and dominance relationships matrices were included in the prediction modelling.

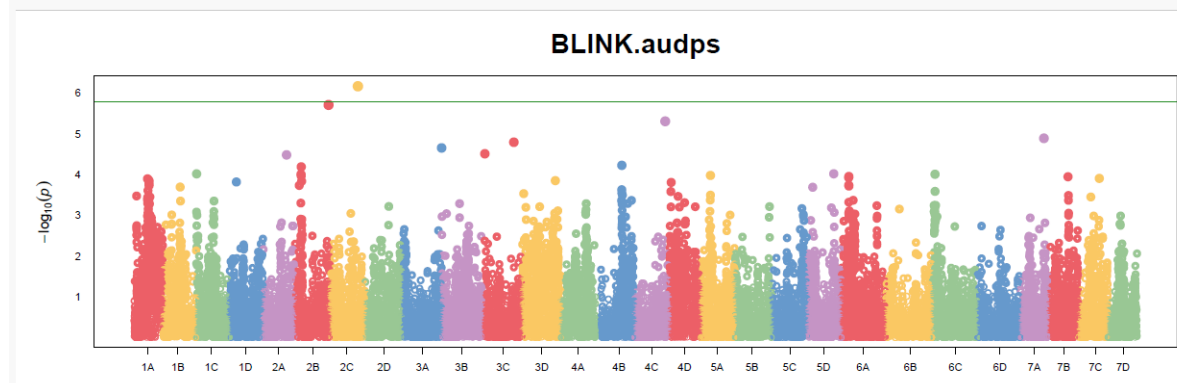
Manhattan plots of AUDPS (2022, 2023, and Combined) – Fresh Forward data (June-bearers only).



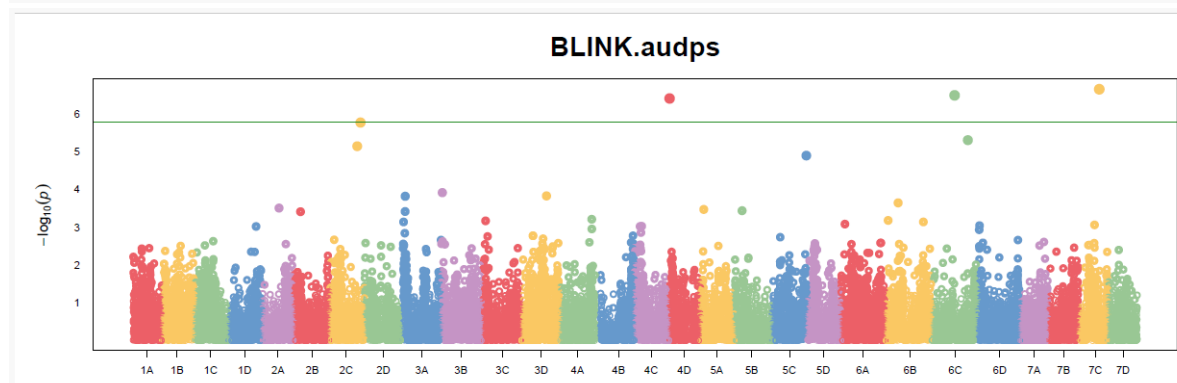
2022



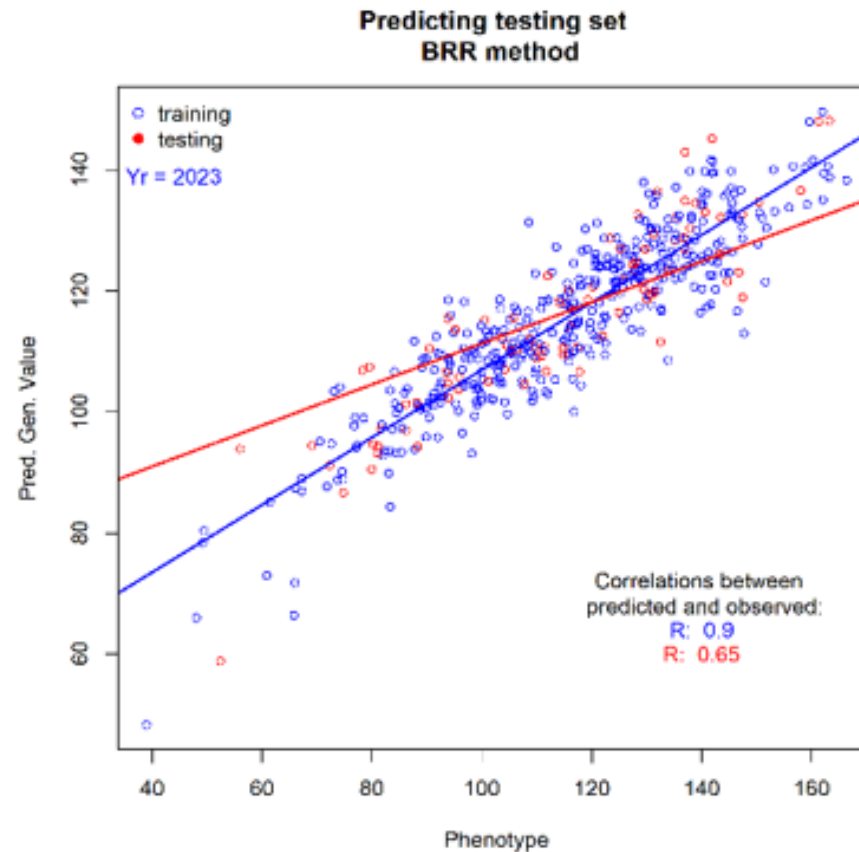
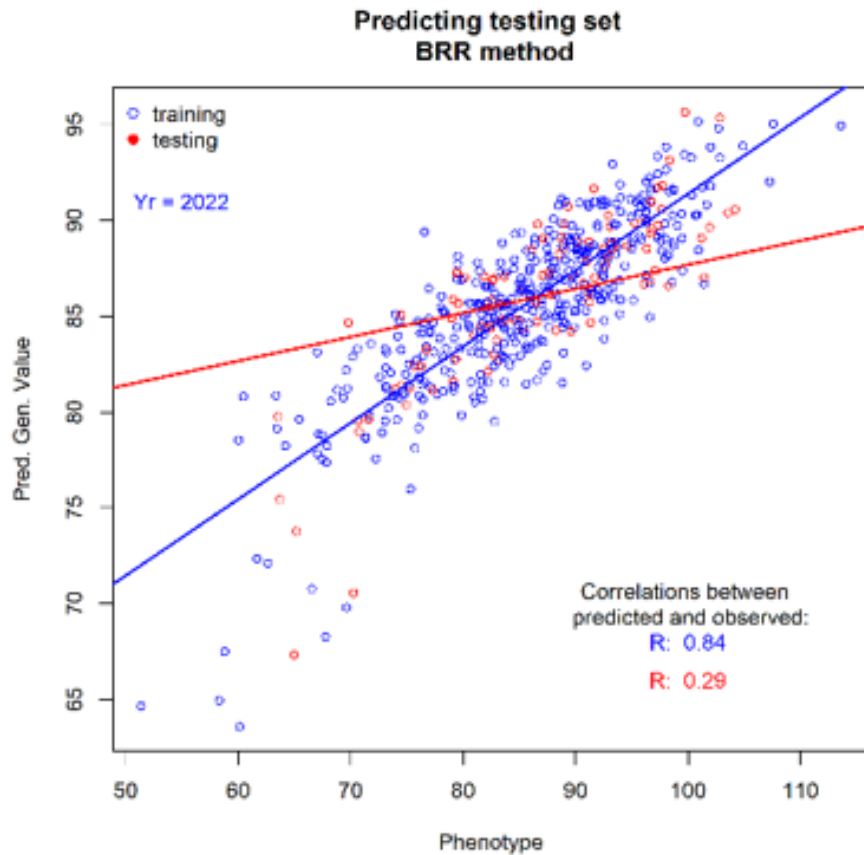
2023



Both Yrs



Predictive Ability of Genomic Selection for Powdery Mildew Resistance - Fresh Forward data (2022 vs 2023)



Two random fold-plots from the cross-validation of powdery mildew scores at Fresh Forward. The 2022 experiment (left) showed a low predictive ability while a significant improvement was observed in the 2023 experiment.



Cross-Validation: Evaluating Prediction Accuracy

- NB! The previous figures represent only one selection of training-testing sets.
- Question: Are these valid results for the entire data set?
- One way to answer this is to do cross-validation.

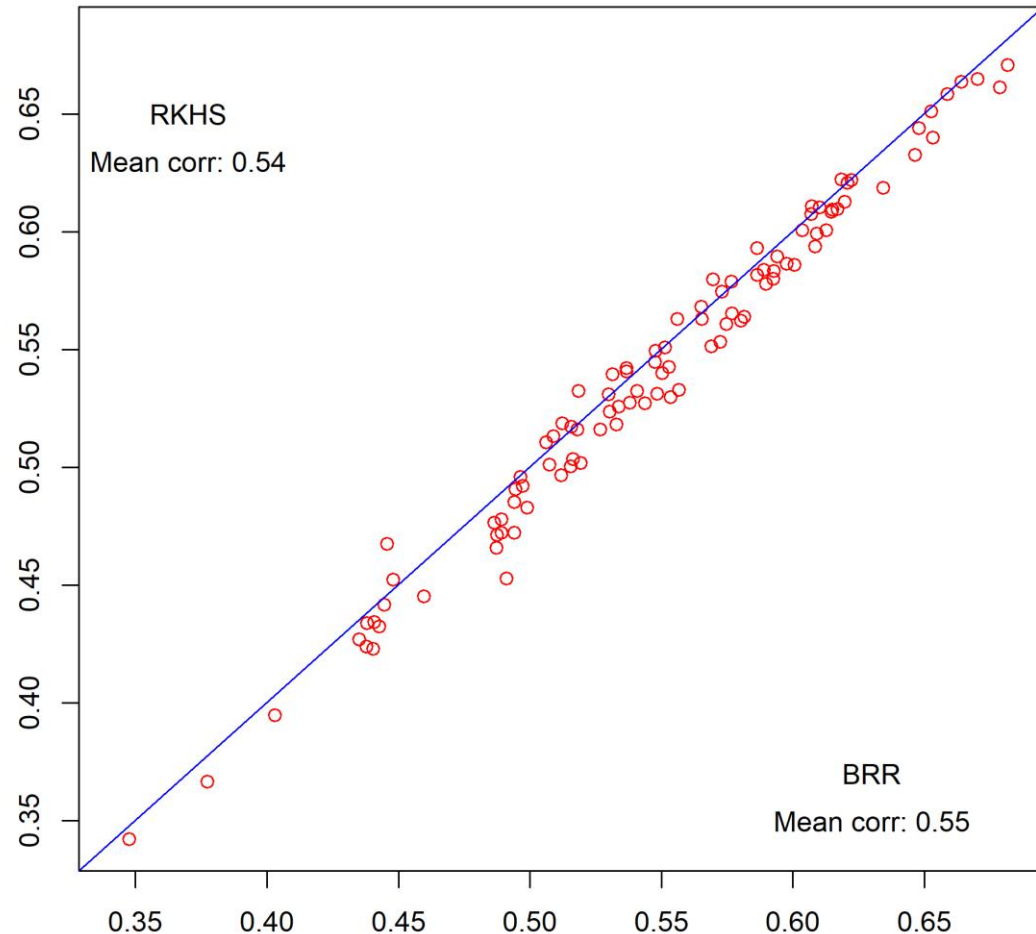
4-fold (k=4) cross validation, one rep:

Fold 1	Testing set		Training set	
Fold 2	Training set	Testing set	Training set	
Fold 3	Training set	Training set	Testing set	Training set
Fold 4	Training set	Training set	Training set	Testing set

Prediction Accuracy from 100 Cross-Validation Runs – Fresh Forward data (2023)

- 100 K=5 training-testing splits used to calculate prediction accuracy as correlation between observed and predicted values.
- BRR and RKHS showed comparable accuracy, with mean correlations of 0.55 and 0.54, respectively.
- BRR slightly outperformed RKHS, suggesting a small advantage for linear models in this dataset.
- Results indicate reliable predictions not driven by random partitioning.

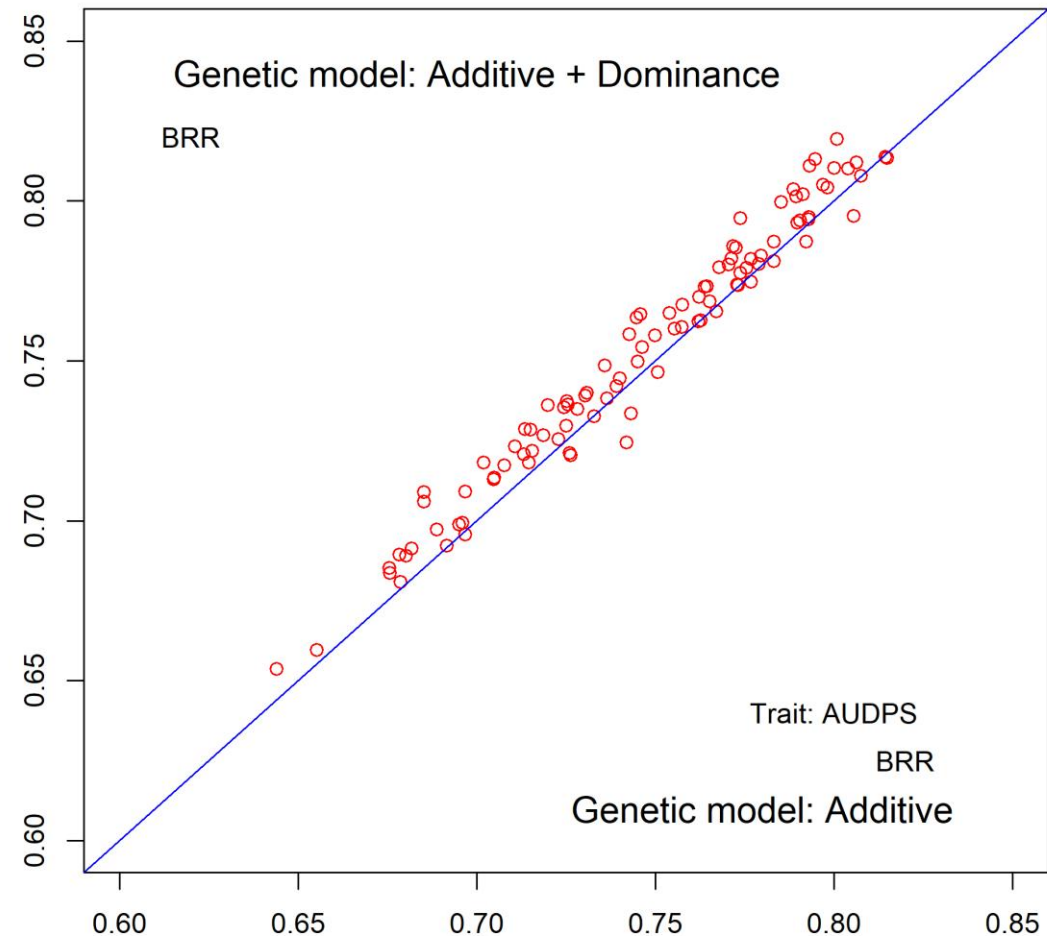
Phenotype - prediction correlations from 100 training-testing partitions.



Comparing genetic models: Additive vs Additive + Dominance relationship matrices - Njos data

Phenotype - predictions correlations
from 100 training-testing partitions.

- **Additive + Dominance model** slightly outperforms the pure **Additive model** in prediction accuracy.
- Results based on 100 cross-validation partitions using the **BRR method**.
- **BayesB approach not necessary** — no significant gain over simpler models.
- **Additive model remains efficient and effective**, especially considering computational cost.



Conclusions:

- From both a **computational and practical standpoint**, the prospects for implementing genomic prediction in applied breeding are promising.
- However, **genotyping costs remain a major barrier** for SMEs in the horticultural breeding sector. For instance, **Werner et al. (2023)** based their simulations on genotyping costs of €25 and phenotyping costs of €20 per sample.
- In our study, we used the **50K strawberry SNP array** for genotyping. However, this density is not required for effective genomic prediction.
- A reduced **5K SNP panel would be sufficient** for prediction purposes and would **significantly lower genotyping costs**. The development of such a **low-cost genotyping assay is urgently needed** to make genomic prediction broadly accessible.

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





Conclusions:

- * From a computational and practical point of view the prospects for using a genomic prediction approach in applied breeding are good.
- The **costs** involved, regarding the genotyping, are currently prohibitive for SMEs in the horticultural breeding industry. (Werner et al 2023 did their simulation studies based on costs of €25 for genotyping and €20 for phenotyping.)
- We used the 50K strawberry array for the genotyping. This number of markers is not necessary for genomic prediction. A 5K markers would do for genomic prediction purposes and an assay of this size would reduce the genotyping costs significantly. Such a low-cost assay is in dire need.

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

Great question! You're referring to the use of a **correlation coefficient** r —typically the **accuracy of prediction** in genomic selection—between the **true breeding value (TBV)** and the **estimated breeding value (EBV)** based on genomic data.

Let's break it down:

🔴 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.).

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 to ensure **efficient selection**.

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

Great follow-up — and you're absolutely right to ask this! It's one of the most interesting and *counterintuitive* aspects of genomic selection. Let's unpack why **genetic gain can still happen**, even when the prediction accuracy r is low (like 0.2–0.4).

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.



For testing the prediction models' robustness, a 5-fold **cross validation** was performed.


There are many versions of **cross-validation, but** we performed it like this:

The sample of phenotypes were randomly partitioned in into 5 sets of roughly equal size.

The model was fit using all samples except the first partition (the first fold).

The held-out samples will then be predicted by this model.

Then the first subset (fold) is re-assigned with its known observational data and the second subset set to missing. Another prediction of the second set's missing data points was then performed. This was repeated all the way to the fifth fold. This would complete the first replicate. Then the entire phenotype data would be re-randomized, and the second replicate would continue in the same manner as the first. The observed and the predicted phenotypes would be visualized in scatter plots and the correlation between them computed as an indication of the model's predictive ability. Moreover, this correlation (r_{TI} in the genetic gain formula below) is important since it is directly related to the expected genetic gain during selection.

 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.

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

