Analysis of Genomic Selection Variying Markers

28 de mayo de 2025

1. Introduction

2. Material and Methods

2.1. Genotypic data

2.2. Phenotypic data

2.3. GWAS analysis

2.4. GP using marker densities

~~To identify the optimal number of markers for predicting GEBVs for each LCH component of the color traits, genomic prediction was performed using varying SNP densities. Markers were selected based on GWAS results and ranked by importance using our custom GSCORE scoring function (see Section 2.3).~~

2.5. Heritabiliy

ChatGPT:
Narrow-sense heritability (h^2) of each color-related trait was estimated using a Bayesian linear mixed model implemented in the BGLR R package, with a realized genomic relationship matrix computed using AGHmatrix.
The genomic relationship matrix (G) was constructed from tetraploid marker data using the VanRaden method, with ploidy correction enabled. We fit a single-kernel RKHS (Reproducing Kernel Hilbert Space) model in BGLR, treating the G matrix as a random effect and estimating variance components via MCMC sampling (nIter = 15,000, burnIn = 5,000, thin = 5).
Narrow-sense heritability was then computed as:

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}$$

h2=σG2+σE2σG2
Where σ_G^2 is the additive genomic variance and σ_E^2 is the residual variance, extracted from the BGLR model output files.

3. Results

3.1. GP using marker densities

Genomic prediction was performed using marker subsets representing 5 % to 100 % of the total 4,641 SNPs. Predictive ability, assessed by the Pearson correlation between observed phenotypes and genomic estimated breeding values (GEBVs), varied by trait and marker density (see Figure S4).

In general, predictive ability increased rapidly with marker density up to ∼40–50 % of the total set, after which gains plateaued—indicating diminishing returns. This pattern held across most color traits and LCH components (Hue, Chroma, Lightness), suggesting that a moderate number of well-distributed, informative SNPs can achieve predictive performance comparable to that of the full marker set.

Stem color and primary flower color traits consistently showed higher predictive abilities, with correlations nearing 0.75 at full marker density. In contrast, secondary traits—such as secondary tuber flesh or sprout color—showed lower predictive values, even with more markers. This variability reflects differences in genetic architecture, with some traits likely controlled by major loci, and others influenced by more polygenic and complex effects.

These results support the feasibility of cost-effective genomic selection in tetraploid *Andigenum* potatoes using an optimized SNP subset. The top 50 % of markers, selected via the GSCORE metric (Section 2.3), effectively captured relevant genetic variation, enabling accurate predictions while reducing computational demand.

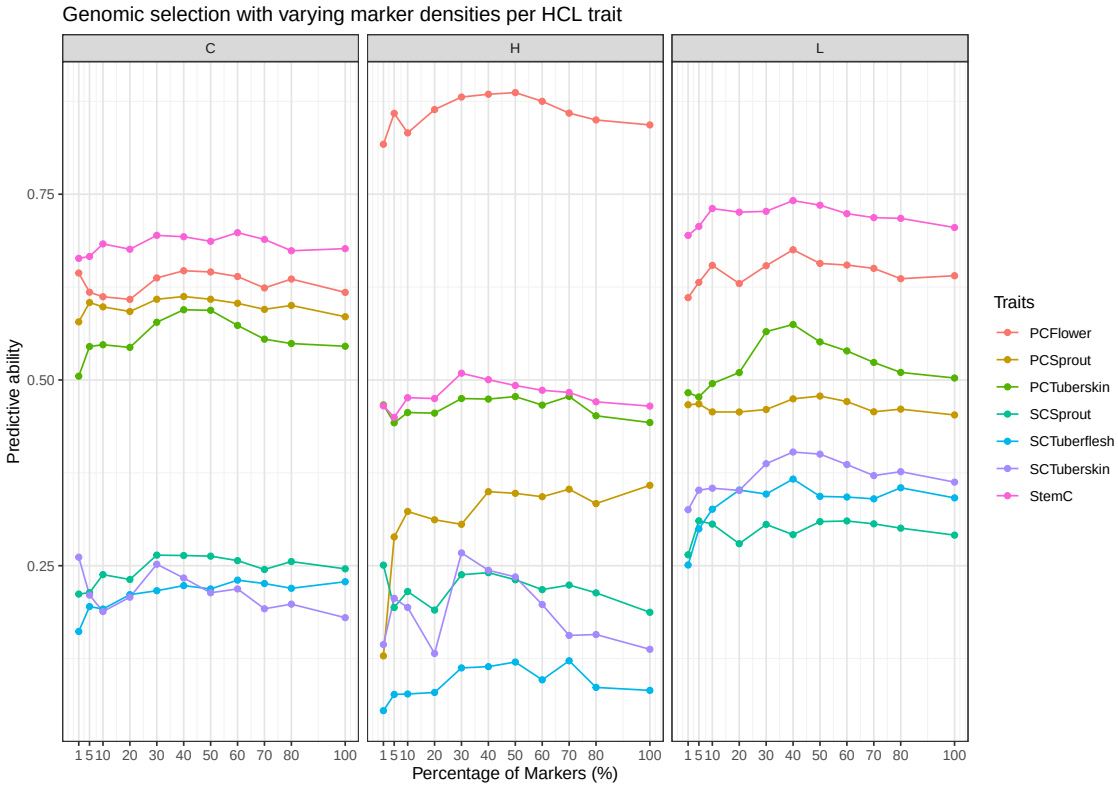

Figura 1: **GP comparison for LCH traits with subsets of markers resulting from the GWAS process**. Each division corresponds to one of LCH component and within each division are the values corresponding to each trait. On the vertical axis is the value of the prediction ability for each LCH component, while on the horizontal axis are the number of markers of each subset with which the GEBV prediction was performed.

4. ChatGpt Editting:

To evaluate the efficiency of genomic prediction models across different SNP densities, genomic selection (GS) was performed using subsets of markers representing 5 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, and 100 % of the total 4,641 markers. Predictive ability, measured as the Pearson correlation between observed phenotypes and genomic estimated breeding values (GEBVs), varied depending on the trait and the number of markers used.

In general, predictive ability increased rapidly with the inclusion of additional SNPs up to approximately 40–50 % of the marker set. Beyond this threshold, the gains in predictive ability plateaued, indicating diminishing returns with additional markers. This trend was consistent across most color traits and LCH components (Hue, Chroma, and Lightness), suggesting that a moderate number of well-distributed and informative SNPs can yield prediction accuracies comparable to those obtained using the complete marker set.

Among the evaluated traits, components of stem color and primary flower color consistently demonstrated higher predictive abilities, with correlations reaching values close to 0.75 when using 100 % of the markers. In contrast, secondary traits, such as the secondary tuber flesh or sprout colors, exhibited lower predictive values even at higher marker densities. These results indicate variability in the genetic architecture of the color traits, where

some traits are likely controlled by fewer loci with larger effects, while others may be more polygenic and influenced by complex interactions.

Overall, these findings highlight the potential for cost-effective genomic selection in tetraploid Andigenum potatoes by utilizing an optimized number of markers. The top 50 SNPs, selected via our GSCORE metric (Section 2.3), captured sufficient genetic architecture. This approach allows for improved selection accuracy without the computational burden of using the entire marker set.

5. DeepSeek Edditing:

We evaluated the predictive ability of genomic selection (GS) for 27 Hue-Chroma-Lightness (HCL) traits derived from nine color characteristics in tetraploid Andigenum potatoes. Using the BWGS pipeline (Charmet et al., 2020), we implemented 12 GS models spanning parametric (GBLUP, EGBLUP, RR-BLUP, LASSO), semiparametric (RKHS, Random Forest, SVM), and Bayesian approaches (BRR, BL, BA, BB, BC). Five-fold cross-validation repeated five times was employed to assess model performance.

Key findings:

1. **Full Marker Set Performance (4,641 SNPs):**
    - Predictive abilities (correlation between observed and predicted values) varied substantially among traits (Figure 5B). Primary flower color (PCFlower) and stem color (StemC) showed the highest predictive abilities (0.86 and 0.75, respectively), while secondary tuber flesh color (SCTuberflesh) had the lowest (0.18).
    - The RKHS model outperformed others for 11/27 HCL components, followed by EGBLUP (9/27) and SVM (5/27). Bayesian models showed particular effectiveness for traits with complex inheritance patterns.
2. Optimization of Marker Number:
    - Analysis of progressively reduced marker sets (100, 50, 35, 15, and 5 top-ranked SNPs) revealed that predictive ability plateaued at ∼50 markers for most traits (Supplementary Figure S3). Notably:
    - Primary tuber skin color (PCTuberskin) maintained 92 % of full-set predictive ability with just 50 markers
    - Secondary sprout color (SCSprout) required 100 markers to reach maximum performance
    - The top 50 SNPs, selected via our GSCORE metric (Section 2.3), captured sufficient genetic architecture while reducing computational load by 98.9 %.
3. Trait Architecture Insights:
    - Traits with highest predictive abilities (PCFlower, StemC) also exhibited the highest narrow-sense heritabilities ($h^2 > 0.8$, Figure 5A), suggesting predominantly additive genetic control.
    - Poorly predicted traits (e.g., SCTuberflesh) showed lower heritability ($h^2 = 0.42$) and higher environmental influence, consistent with their weaker GWAS signals.
4. Component-Specific Patterns:
    - Hue (H) components were generally more predictable than Chroma (C) or Lightness (L), with mean predictive abilities of 0.61 vs. 0.49 and 0.47, respectively. This aligns with known biochemical pathways where hue variation strongly reflects anthocyanin and carotenoid content.

Breeding Implications:

The demonstrated efficacy of GS with reduced marker sets (50-100 SNPs) enables cost-effective implementation in Andigenum potato breeding programs. Our results suggest:

- Early-generation selection for highly heritable traits (flower/stem color) can achieve accuracy >0.75
- Secondary color traits may require phenotypic validation despite genomic predictions
- The RKHS model is recommended as default for HCL trait prediction, with EGBLUP for hue-related components