

Genetic Insights into Color Traits in Tetraploid Andigenum Potatoes

¹CI Tibaitatá, Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), Km 14 via Mosquera, Bogotá, 250047, Cundinamarca, Colombia.

²CI Sede Central, Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), Km 14 via Mosquera, Bogotá, 250047, Cundinamarca, Colombia.

³Current address: Grupo de Investigación Bioinformática y Biocomputación , Universidad del Valle, Cali, 760042, Valle del Cauca, Colombia.

Corresponding author(s). E-mail(s): phreyes@agrosavia.co;

1 Introduction

Potato (*Solanum tuberosum L.*) is the most important non-cereal crop in the world (worldwide) and is essential for food security (Reyes-Herrera et al., 2024). The origin of cultivated potatoes includes the Andean and Chilean landraces, with the Andean landraces being the most widely grown. In addition, Andean landraces have high morphological and genetic diversity (Spooner et al., 2005). Several studies from the Andean region have reported color and nutritional diversity in Peru (Bellumori et al., 2020), Ecuador (Balladares and Ramos, 2018), Argentina (Calliope et al., 2018), and Colombia (Berdugo-Cely et al., 2017, 2023).

Plant color measurements have evolved over time. Initially, it relied on human perception, followed by color matching with predefined charts. Currently, more accurate and objective methods involve spectrometric and photographic color measurements (Kasajima, 2019). However, characterizations using color descriptors must be consistent, and for this purpose, it is necessary to relate characterizations with different descriptors.



Fig. 1 Color variation in tuber skin and flesh among CCC accessions.

2 Materials and Methods

The Colombian Central Collection of potatoes (CCC) is one of the most diverse collections in Colombia and the most important source of genetic variability for the improvement of this crop in Colombia ([Manrique-Carpintero et al., 2023](#)). Currently, the CCC clonal collection conserved in the field preserves 1225 accessions, with 68.8% consisting of *Solanum tuberosum* subsp. *andigenum* landraces. Most accessions were collected before 1985.

The regeneration of the CCC is conducted annually in the municipality of Zipaquirá (See Figure 2), located in the department of Cundinamarca, Colombia. This area is situated at an altitude of 2,950 m, with an average temperature of 15°C and a relative humidity of 75% ([Berdugo-Cely et al., 2017](#)).

2.1 Genotypic data

A set of 657 tetraploid accessions from the CCC, consisting of the *S. tuberosum* group Andigenum, was previously genotyped ([Berdugo-Cely et al., 2017](#)) using the Illumina Infinium SolCAP SNP array (8303 SNP). The array was processed on the Illumina HiScan SQ system (Illumina, San Diego, CA) at AGROSAVIA, and the ClusterCall R package ([Schmitz Carley et al., 2017](#)) was used to obtain the dosage genotype calls from the Theta values and(?) raw data values, applying default parameters, and calibration from F1 populations supported by the software. This resulted in a genotype call matrix (0: AAAA, 1: AAAB, 2: AABB, 3: ABBB, and 4: BBBB) for each SNP across all accessions.

To ensure data quality and reliability for subsequent GWAS analyses, standard filtering criteria were applied. SNPs with a minor allele frequency (MAF) below 1% were excluded to eliminate rare variants that may lead to spurious associations due



Fig. 2 Aerial image of the CCC field collection in 2022, with each furrow representing an accession, consisting of 20 plants. The orthophoto was created using images captured by a P4 UAV at an altitude of 12 meters in September 2022.

to insufficient statistical power ([Anderson et al., 2010](#)). Individuals with a genotype call missing rate (MIND) greater than 10% were removed to avoid bias introduced by low-quality samples. SNPs with a missing rate (GENO) of $\geq 10\%$ were filtered to maintain marker integrity and reduce imputation noise. Additionally, SNPs with Hardy-Weinberg equilibrium (HWE) p-values below 1e-10 were excluded because extreme deviations may indicate genotyping errors, population structure artifacts, or selection ([Wigginton et al., 2005](#)). These thresholds are commonly used in polyploid GWAS studies and are consistent with established quality control practices ([Lu et al., 2013](#); [Pavan et al., 2020](#)). Following this filtering process, 4,641 SNPs were retained for downstream genomic analyses.

2.2 Phenotypic data

Annual regeneration of CCC was used to characterize morphological traits. The field consists of seven square meters allocated for planting 20 seed tubers per accession. Due to the large size of the *S. tuberosum Andigenum* group collection and limited human resources, 600 accessions were evaluated over three years from 2015 to 2017. Color traits were characterized for stem (7 codes), berry (7 codes), and primary and secondary colors of the flower (8 and 9 codes, respectively), tuber skin (9 and 10 codes, respectively), tuber flesh (8 and 9 codes, respectively), and sprout (5 and 6 codes, respectively), using the descriptors proposed by ([Gómez, 2000](#)).

Additionally, in 2019, color characterization was performed using the fifth edition of the Royal Horticultural Society (RHS) color chart ([Voss, 2002](#)) to provide a more detailed color evaluation. This change expanded the codes from a maximum of ten options to a set of 884 codes. The chart was used to characterize the following nine traits: stem color (StemC), berry color (BerryC), primary flower color (PCFlower), primary tuber skin color (PCTuberskin), secondary tuber skin color (SCTuberskin), primary tuber flesh color (PCTuberflesh), secondary tuber flesh color (SCTuberflesh), primary sprout color (PCSprout), and secondary sprout color (SCSprout).

Moreover, the color characterization using the RHS color chart was converted to the Hue, Chroma, and Lightness (HCL or LCH) color space, as this model is closely aligned with human color perception and has been used in previous studies on potato color ([Caraza-Harter and Endelman, 2020](#)). The values of these LCH traits were derived from the transformation of the values of the potato color traits, measured using the Royal Horticultural Society (RHS) fifth edition color chart (<http://rhscf.orgfree.com/>) ([Voss, 2002](#)), into the three components: L, C, and H, of the Commission Internationale de l'Éclairage (CIE) LCH or CIE*L*C*h color space. Consequently, three components were obtained for each potato color trait, which will be referred to as LCH traits. Thus, the nine potato color traits were transformed into 27 LCH traits, which were used for genomic analyses (Fig. S1). As no outliers were detected, all available data were included in the analysis.

Correlation between two color descriptors

Two descriptors with different numbers of codes were used to characterize the color traits in CCC in different years. The correlation between the evaluations of both descriptors was calculated for each of the nine color traits using Cramer's V statistic ([Lyman et al., 1986](#)). Cramer's V measures the strength of the association between two qualitative variables, with values ranging from 0 to 1. Values below 0.6 indicate a weak association, while values of 0.6 or higher indicate a medium to strong association. To study the genetics underlying color, we selected traits with a medium-to-strong association, indicating genetic influence.

Linking Color to Nutritional and Nutraceutical Variables

This study also examined the association between color traits and nutritional and nutraceutical values using data from a previous study [Berdugo-Cely et al. \(2023\)](#).

The nutritional traits of the potatoes were represented by the Ascorbic Acid Content (AAC), while the nutraceutical traits were: Total Phenol Content (TPC) and Antioxidant Activity (measured by 2,2-Diphenyl-1-picrylhydrazyl Radical Scavenging Activity - DPPH and Ferric Reducing Antioxidant Power - FRAP assays). Given that the nutritional and nutraceutical variables were continuous, and the color traits were categorical, the determination coefficient was applied as a measure of association Nagelkerke et al. (1991). To evaluate these associations, data from 282 Andigenum accessions in the CCC collection, which included information on both sets of traits, were analyzed.

2.3 GWAS analysis

A previous GWAS study by (Berdugo-Cely et al., 2017) employed the same genetic dataset and color descriptors as those defined by (Gómez, 2000). In contrast, the present study used the RHS color chart as a phenotype for color characterization, increasing the number of trait values from 30 to 884, thus achieving greater resolution.

We utilized MultiGWAS (Garreta et al., 2021), a tool that integrates the outputs from multiple GWAS methods. Specifically, we incorporated the results from GWASpoly (Rosyara et al., 2016) for tetraploid genotypes and GAPIT (Tang et al., 2016) for diploid genotypes. The analysis was conducted using the MultiGWAS Full Model (Q+K), which controls for population structure (Q, typically derived from principal component analysis) and kinship (K, based on a genetic relatedness matrix), thereby reducing the likelihood of false-positive associations.

In MultiGWAS, each tool implements a mixed linear model (MLM) (Phenotype Genotype + Q + K), but with method-specific adjustments. GWASpoly adapts the MLM for polyploid data, incorporating population structure and kinship as fixed and random effects, respectively. GAPIT employs a standard MLM, treating the population structure (Q) as a fixed effect (e.g., from PCA) and kinship (K) as a random effect via its kinship matrix. By unifying these approaches, MultiGWAS robustly controls for population stratification and relatedness while detecting marker-trait associations across ploidy levels (diploid and tetraploid).

To address the issue of multiple testing, the MultiGWAS tool applies the Bonferroni correction. However, instead of adjusting the *p*-values, MultiGWAS adjusts the threshold to determine a significant *p*-value. Specifically, this threshold is set to α/m , where α represents the significance level and m is the number of tested markers from the genotype matrix. The CMPlot library (Yin et al., 2021) was used to generate Manhattan and Circos plots to visualize significant SNP markers.

Candidate Marker Identification

Given the evaluation of multiple traits and the integration of results from the four GWAS tools in MultiGWAS, we developed a proprietary scoring function called GSCORE. This function was designed to rank markers by incorporating the outputs from the MultiGWAS. Specifically, GSCORE selected the top 50 markers with the lowest *p*-values from each tool, resulting in 100 markers (2 tools \times 50 markers). GSCORE

is calculated as the sum of three weighted terms: inflation factor (I), contributing 70% of the score. Replicability (R) accounts for 10%, and significance (S) accounts for the remaining 20%. The scoring function is represented by the following equation:

$$GSCORE(M) = 0.7 * I + 0.1 * R + 0.2 * S$$

I is the inflation factor score, defined as $I = 1 - |1 - \lambda(M)|$, where $\lambda(M)$ is the inflation factor for marker M. This score is the highest when $\lambda(M)$ is close to 1. In addition, R is the number of SNPs shared among the three GWAS tools. Finally, S is a binary value (1 or 0) that indicates whether the SNP is significant (p -value < threshold). Marker M achieves a high $GSCORE$ when it has an inflation factor $\lambda(M)$ close to one, identifies a large number of shared SNPs across tools, and is statistically significant. In contrast, the score is low if $\lambda(M)$ is either too low (close to 0) or excessively high, identifies few shared SNPs, or is not significant. In other scenarios, the score is determined by a balance between the inflation factor, shared SNP count, and significance. This approach prioritizes markers that are statistically robust, consistent across tools, and highly significant.

Marker annotation

Markers were annotated by taking information from the Potato Genome Sequencing Consortium (PGSC) public data based on the double monoploid *S. tuberosum* Group Phureja from assembly DMv6.1([Pham et al., 2020](#)) obtained from the SpudDB website. Annotation was performed by associating marker identifiers with information from both the high confidence gene models annotation, the InterProScan assigned GO terms for the working gene models, the InterProScan search results for the working gene models, and the SolCAP 69 K SNP positions on the DMv6.1 assembly available in SpudDB. Furthermore, for each trait, the amino acid sequences were analyzed using BlastKOALA ([Kanehisa et al., 2016](#)) for KEGG mapping to identify common functional categories among the significant markers.

2.4 Heritability Estimation

Narrow-sense heritability (h^2), defined as the proportion of phenotypic variance explained by additive genetic effects ([de los Campos et al., 2015](#)), was estimated for each of the 21 LCH components derived from color traits. We used Bayesian Ridge Regression (BRR), implemented in the BGLR R package ([Pérez and de los Campos, 2014](#)), because the color-related traits analyzed in this study are expected to follow a polygenic architecture, with many loci of small effect contributing additively to the phenotype.

BRR assumes normally distributed genetic effects across markers and estimates heritability by repeatedly sampling the additive genetic variance (σ_a^2) and residual variance (σ_e^2) through a Bayesian process. The model was run for 15,000 iterations, with a burn-in of 5,000 and thinning every 5 steps. Heritability was computed at each iteration as:

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

This procedure yields a posterior distribution of heritability values for each trait, rather than a single point estimate. From this distribution, we report the posterior mean and the 95% credible interval. Unlike confidence intervals from frequentist methods, a Bayesian 95% credible interval indicates the range within which the true h^2 value lies with 95% probability, given the data and model.

2.5 Genomic Prediction (GP)

We applied Genomic Prediction (GP) to estimate the genetic potential of individuals for each of the 21 LCH color trait components. A total of seven statistical models were used, spanning parametric, semiparametric, and Bayesian approaches:

- **Parametric:** Genomic Best Linear Unbiased Prediction (GBLUP)
- **Semiparametric:** Reproducing Kernel Hilbert Spaces (RKHS)
- **Bayesian:** Bayesian Ridge Regression (BRR), Bayesian LASSO (BL), Bayes A, Bayes B, and Bayes C

All models were implemented using the BWGS R package ([Charmet et al., 2020](#)). Predictive performance was assessed via 5-fold cross-validation, repeated five times for each trait. Predictive ability was measured as the Pearson correlation between observed phenotypes and genomic estimated breeding values (GEBVs) across validation folds. All model parameters were set to the default values provided by the BWGS library.

2.5.1 GP using marker subsets

To determine the optimal number of markers for genomic prediction (GP) of each LCH color trait component, GP was conducted using subsets containing varying numbers of markers. These markers were selected based on GWAS results and prioritized according to their relevance using our custom GSCORE scoring function (see Section [2.3](#)).

3 Results

The distribution for the three components of the color traits can be seen in Figure S1 in the supplementary information.

3.1 Correlation between descriptors

The correlations between the values of the color traits measured with the two color descriptors over different years are presented in Figure ?? A. Among the traits, only berry color and primary tuber flesh color exhibited weak correlations, leading to their exclusion from the genetic study. In contrast, the remaining seven color traits showed moderate to strong correlations, ranging from 0.65 to 0.97. Although some color traits were correlated with others, all moderate to strong correlations, as expected, were concentrated along the diagonal of the correlation matrix. In particular, there were correlations between primary and secondary tuber skin colors, between primary tuber

skin color and both primary and secondary tuber flesh colors, and between primary and secondary sprout colors.

3.1.1 Linking Color to Nutritional and Nutraceutical Variables

In Figure ?? B, the heatmap presents the determination coefficient R^2 between the color traits measured using the RHS color scale and nutritional/nutraceutical variables. R^2 below 0.3 reflect a weak association (blue cells), while R^2 between 0.3 and 0.51 reflects a moderate association between color traits and nutritional variables (yellow and red). Three specific traits—the secondary color of the tuber flesh and the primary and secondary colors of the tuber skin—exhibited notable associations with nutritional variables. Antioxidant activity (measured using FRAP and DPPH assays) showed the strongest correlation with these traits. Additionally, Total Phenol Content and Ascorbic Acid Content demonstrated the highest association with the primary color of the tuber skin. Stem color, meanwhile, showed a moderate association with antioxidant activity and Total Phenol Content.

3.2 GWAS Analysis

GWAS analysis identified 21 significant SNPs associated with five traits: ten related to the secondary color of the tuber flesh, four to the primary color of the flower, three to the primary color of the sprout, two to the secondary color of the sprout, and two to the stem color. Manhattan plots for all seven traits are presented in Figure 3 (A-G). The most significant SNPs were associated with the Hue component and secondary color of the tuber flesh trait.

Notably, chromosome 2 accounted for 28.5% of the significant SNPs, followed by chromosomes 1 and 11, each contributing 19% and 14.29%, respectively. A Circos plot (Figure S2) combines the data for all seven traits and presents the density of SNP markers for chromosomes, allowing the visualization of SNP co-occurrence across different traits.

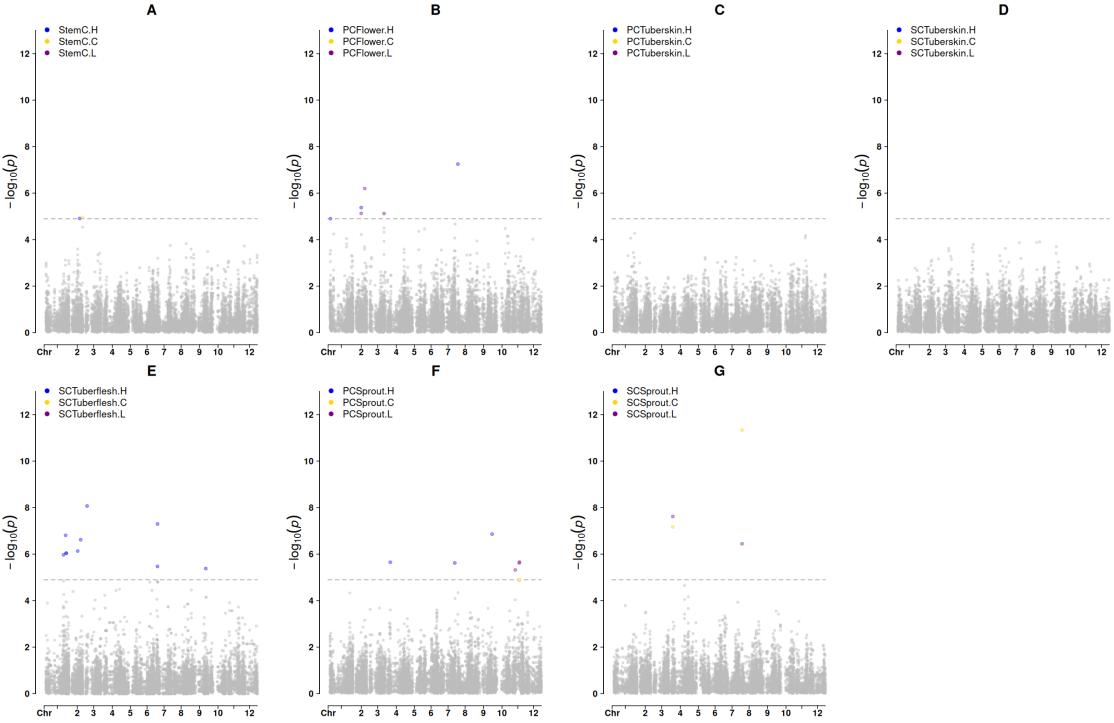


Fig. 3 (A-G) Manhattan plots show the markers linked to LCH traits, with colors indicating the significant SNPs associated with each LCH color component.

A total of 21 significant associations were detected across the nine HCL components that corresponded to five traits. These associations involved 21 SNP markers annotated with 21 genes, represented by a total of 33 transcripts. A total of eight functional pathway categories were identified based on genes annotated with significant SNPs: seven belonged to the metabolism category and one to the genetic information processing category. Information on the primary annotations linked to these markers, as retrieved from the SPUD database, is provided in Supplementary Table 1, and the specific functional categories per trait are provided in Supplementary Table 2.

Moreover, GWASpoly and GAPIT identified 50% of the significant SNPs, with both tools jointly predicting nine SNPs, representing 42.85% of the total. Although no common SNPs were found across the different traits, four SNPs were associated with distinct HCL components of the same trait.

3.3 Heritability of Color Traits

Figure 4.A summarizes the posterior distributions of narrow-sense heritability (h^2) for 21 LCH components across seven color traits in the Andigenum group of *Solanum tuberosum*. Each boxplot represents the distribution of h^2 values obtained from Bayesian Ridge Regression (BRR), showing the median, interquartile range, and 95% credible interval.

Most primary color traits—particularly PCFlower.H, StemC.L, and PCSprout.C—exhibited moderate to high heritability, with posterior means exceeding 0.55. The highest heritability was observed in the hue component of secondary tuber flesh color (SCTuberflesh.H), with a mean near 0.98, indicating strong genetic control.

In contrast, several secondary color traits such as SCSprout.C, SCSprout.H/L, and SCTuberskin.C displayed lower heritability values (posterior means between 0.20 and 0.35), suggesting a greater influence of environmental or non-additive factors.

Overall, StemC, PCFlower, and PCSprout traits consistently showed moderate to strong heritability across their Chroma, Hue, and Lightness components, supporting their potential as targets for genomic selection.

3.4 Genomic Prediction Results

Figure 4.B shows the predictive abilities of the seven genomic prediction models for each of the 21 LCH trait components. Predictive ability is defined as the correlation between observed phenotypic values and GEBVs across cross-validation folds.

The highest predictive abilities were observed in traits such as PCFlower.H, StemC.L, and PCSprout.C, with mean correlations above 0.70. Other primary color traits—including PCTuberskin.C and PCSprout.L—showed moderate predictive performance.

Traits with the lowest predictive abilities included several secondary color traits such as SCSprout.H, SCTuberskin.C, and SCTuberflesh.L, with mean correlations below 0.40. These results mirror the patterns observed in the heritability estimates (panel A), with high heritability traits generally showing higher predictive accuracy.

Model performance varied across traits, with BRR, GBLUP, and RKHS performing best for different subsets of traits. The top-performing model per trait is annotated in Figure 4.B.

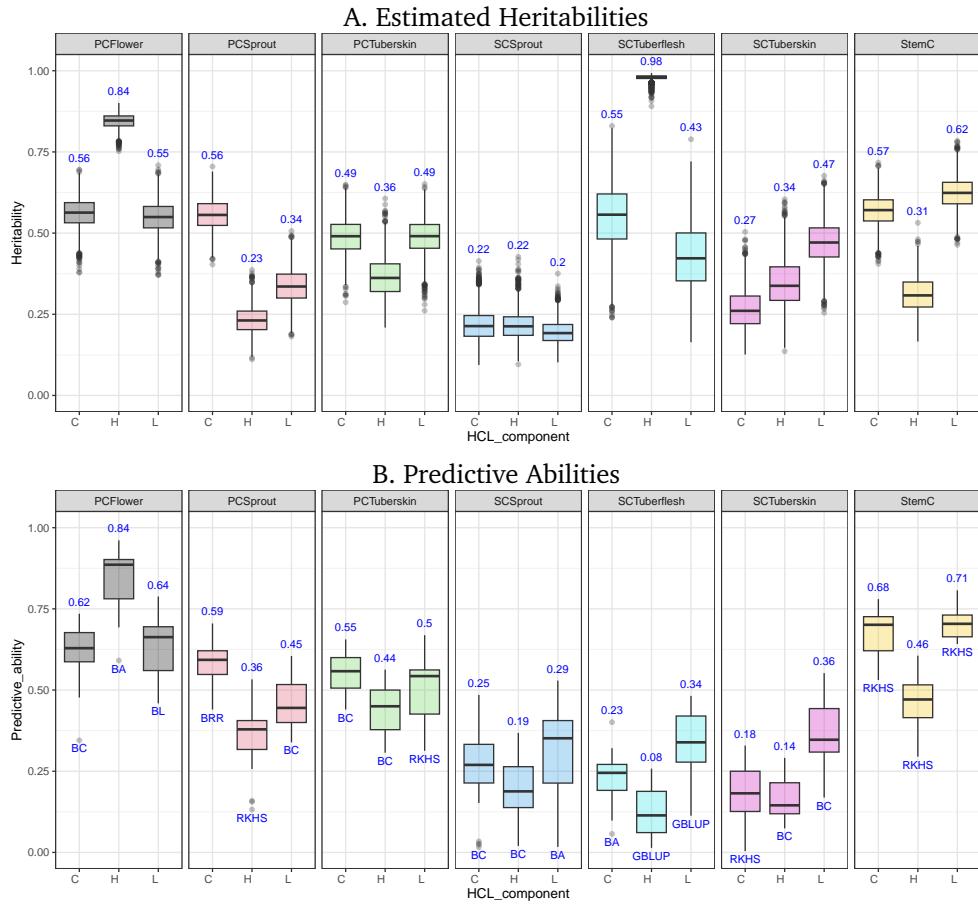


Fig. 4 Heritability and Predictive Ability for LCH Traits. Each boxplot corresponds to one of the 21 LCH color trait components in *Solanum tuberosum* Andigenum group. Traits are grouped by color feature (e.g., flower, sprout, skin, stem) and by component (Chroma: C, Hue: H, Lightness: L). **A. Posterior distributions of narrow-sense heritability (h^2)**, estimated using Bayesian Ridge Regression (BRR) with 15,000 iterations, 5,000 burn-in, and thinning every 5 steps. Posterior medians are annotated above each box. **B. Predictive abilities of seven genomic prediction models.** Values represent the correlation between observed phenotypes and GEBVs across repeated 5-fold cross-validation. The mean predictive ability is shown above each boxplot, and the model achieving the best performance per trait is indicated below.

3.4.1 GP using marker subsets

Genomic prediction was performed using marker subsets representing 5%–100% of the 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 S3).

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 for tetraploid *Andigenum* potatoes by using an optimized SNP subset. The top 50% of markers, selected via the GSORE metric (Section 2.3), effectively captured the relevant genetic variation, enabling accurate predictions while reducing the computational demand.

4 Acknowledgments and Funding

References

- Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., Zondervan, K.T.: Data quality control in genetic case-control association studies. *Nature Protocols* **5**(9), 1564–1573 (2010) <https://doi.org/10.1038/nprot.2010.116>
- Berdugo-Cely, J.A., Céron-Lasso, M.d.S., Yockteng, R.: Phenotypic and molecular analyses in diploid and tetraploid genotypes of *solanum tuberosum* l. reveal promising genotypes and candidate genes associated with phenolic compounds, ascorbic acid contents, and antioxidant activity. *Frontiers in Plant Science* **13**, 1007104 (2023)
- Bellumori, M., Chasquibol Silva, N.A., Vilca, L., Andrenelli, L., Cecchi, L., Innocenti, M., Balli, D., Mulinacci, N.: A study on the biodiversity of pigmented andean potatoes: Nutritional profile and phenolic composition. *Molecules* **25**(14), 3169 (2020)
- Berdugo-Cely, J., Valbuena, R.I., Sánchez-Betancourt, E., Barrero, L.S., Yockteng, R.: Genetic diversity and association mapping in the Colombian Central Collection of *Solanum tuberosum* L. *Andigenum* group using SNPs markers. *PLOS ONE* **12**(3), 0173039 (2017) <https://doi.org/10.1371/journal.pone.0173039>
- Balladares, V., Ramos, M.: Evaluación de antocianinas y micronutrientes en papa nativa (*solanum andigena*). *Alimentos Ciencia e Ingeniería* **26**(1), 7–16 (2018)
- Caraza-Harter, M.V., Endelman, J.B.: Image-based phenotyping and genetic analysis of potato skin set and color. *Crop Science* **60**(1), 202–210 (2020)
- Calliope, S.R., Lobo, M.O., Sammán, N.C.: Biodiversity of andean potatoes: Morphological, nutritional and functional characterization. *Food chemistry* **238**, 42–50 (2018)

- Charmet, G., Tran, L.-G., Auzanneau, J., Rincent, R., Bouchet, S.: BWGS: A R package for genomic selection and its application to a wheat breeding programme. PLOS ONE **15**(4), 0222733 (2020) <https://doi.org/10.1371/journal.pone.0222733>
- Campos, G., Sorensen, D., Gianola, D.: Genomic Heritability: What Is It? PLoS Genetics **11**(5), 1–21 (2015) <https://doi.org/10.1371/journal.pgen.1005048>
- Garreta, L., Cerón-Souza, I., Palacio, M.R., Reyes-Herrera, P.H.: MultiGWAS: An integrative tool for Genome Wide Association Studies in tetraploid organisms. Ecology and Evolution **11**, 3–7572 (2021) <https://doi.org/10.1002/ece3.7572>
- Gómez, R.: Guía para las caracterizaciones morfológicas básicas en colecciones de papas nativas. Centro Internacional de la Papa (CIP), Germoplasma de Papa, Dpto. de Mejoramiento y Recursos Genéticos. CIP, Lima, Perú (2000)
- Kasajima, I.: Measuring plant colors. Plant Biotechnology **36**(2), 63–75 (2019)
- Kanehisa, M., Sato, Y., Morishima, K.: Blastkoala and ghostkoala: Kegg tools for functional characterization of genome and metagenome sequences. Journal of molecular biology **428**(4), 726–731 (2016)
- Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R.J., Cherney, J.H., Casler, M.D., Buckler, E.S., Costich, D.E.: Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based snp discovery protocol. PLoS Genetics **9**(1), 1003215 (2013) <https://doi.org/10.1371/journal.pgen.1003215>
- Lyman, O., Ruchard, L., Rexroat, C., Mendenhall, W.: Statistics: A tool for the Social Sciences. PWS-Kent Publishing Company, Boston (379pp) (1986)
- Manrique-Carpintero, N.C., Berdugo-Cely, J.A., Cerón-Souza, I., Lasso-Paredes, Z., Reyes-Herrera, P.H., Yockteng, R.: Defining a diverse core collection of the colombian central collection of potatoes: a tool to advance research and breeding. Frontiers in Plant Science **14**, 1046400 (2023)
- Nagelkerke, N.J., et al.: A note on a general definition of the coefficient of determination. biometrika **78**(3), 691–692 (1991)
- Pérez, P., Campos, G.: BGLR : A Statistical Package for Whole Genome Regression and Prediction. Genetics **198**(2), 483–495 (2014)
- Pavan, S., Delvento, C., Ricciardi, L., Lotti, C., Ciani, E., D'Agostino, N.: Recommendations for choosing the genotyping method and best practices for quality control in crop genome-wide association studies. Frontiers in Genetics **11**, 447 (2020) <https://doi.org/10.3389/fgene.2020.00447>
- Pham, G.M., Hamilton, J.P., Wood, J.C., Burke, J.T., Zhao, H., Vaillancourt, B., Ou, S., Jiang, J., Buell, C.R.: Construction of a chromosome-scale long-read reference

genome assembly for potato. *GigaScience* **9**(9), 100 (2020) <https://doi.org/10.1093/gigascience/giaa100>

Rosyara, U.R., De Jong, W.S., Douches, D.S., Endelman, J.B.: Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. *The Plant Genome* **9**(2), 1–10 (2016) <https://doi.org/10.3835/plantgenome2015.08.0073>

Reyes-Herrera, P.H., Delgadillo-Duran, D.A., Flores-Gonzalez, M., Mueller, L.A., Cristancho, M.A., Barrero, L.S.: Chromosome-scale genome assembly and annotation of the tetraploid potato cultivar diacol capiro adapted to the andean region. *G3 Genes—Genomes—Genetics* **14**(9), 139 (2024) <https://doi.org/10.1093/g3journal/jkae139> <https://academic.oup.com/g3journal/article-pdf/14/9/jkae139/60432514/jkae139.pdf>

Schmitz Carley, C.A., Coombs, J.J., Douches, D.S., Bethke, P.C., Palta, J.P., Novy, R.G., Endelman, J.B.: Automated tetraploid genotype calling by hierarchical clustering. *Theoretical and Applied Genetics* **130**(4), 717–726 (2017) <https://doi.org/10.1007/s00122-016-2845-5>

Spooner, D.M., McLean, K., Ramsay, G., Waugh, R., Bryan, G.J.: A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. *proceedings of the national academy of sciences* **102**(41), 14694–14699 (2005)

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z., Pan, Y., Liu, D., Lipka, A.E., Buckler, E.S., Zhang, Z.: GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *The Plant Genome* **9**(2), 2015–110120 (2016) <https://doi.org/10.3835/plantgenome2015.11.0120>

Voss, D.H.: The royal horticultural society colour chart 2001 (2002)

Wigginton, J.E., Cutler, D.J., Abecasis, G.R.: A note on exact tests of hardy-weinberg equilibrium. *The American Journal of Human Genetics* **76**(5), 887–893 (2005) <https://doi.org/10.1086/429864>

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., *et al.*: rmvp: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics, Proteomics and Bioinformatics* **19**(4), 619–628 (2021)