Genetic Insights into Color Traits in Tetraploid Andigenum Potatoes

Luis Garreta[1], Zahara Lasso-Paredes[2], Jhon A. Berdugo-Cely[2], Ivania Cerón-Souza[2], Paula H. Reyes-Herrera[2*]

[1]Department, Organization, Street, City, 100190, State, Country.
[2]CI Tibaitatá, Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), Km 14 via Mosquera, Bogotá, 250047, Cundinamarca, Colombia.

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

Abstract

The abstract serves both as a general introduction to the topic and as a brief, non-technical summary of the main results and their implications. Authors are advised to check the author instructions for the journal they are submitting to for word limits and if structural elements like subheadings, citations, or equations are permitted.

Keywords: keyword1, Keyword2, Keyword3, Keyword4

1 Introduction

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 improvement of this crop in Colombia (Manrique-Carpintero et al., 2023). Currently, the CCC clonal collection conserved in the field preserves 1210 accessions, with 68.8% consisting of *Solanum tuberosum subsp. andigenum* landraces. The majority of these 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 sits

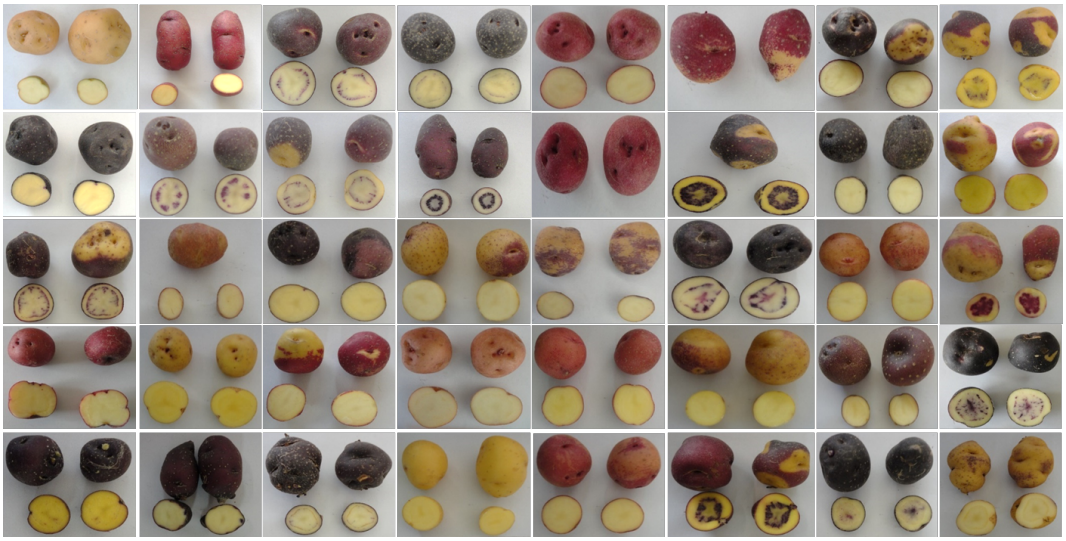

Fig. 1 Color variation in tuber skin and flesh among CCC accessions.

at an altitude of 2,950 meters, 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 XY raw data, applying default parameters and calibration from F1 populations. This resulted in a genotype call matrix (0: AAAA, 1: AAAB, 2: AABB, 3: ABBB, 4: BBBB) for each SNP across all accessions.

To ensure data quality, the following filters were applied: minor allele frequency (MAF) to exclude SNPs with MAF below 1%; individual missing rate (MIND) to filter out individuals with a MIND greater than 10%; SNP missing rate (GENO) to remove SNPs with missing values exceeding 10%; and Hardy-Weinberg equilibrium (HWE) to exclude SNPs with a p-value below the 1e-10 threshold in the Hardy-Weinberg exact test.

2.2 Phenotypic data

The annual regeneration of the CCC is utilized 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 the limited human resources, a total of 600 accessions were evaluated over three different years, from 2015 to 2017. Color traits were characterized for stem (7 codes), berry (7

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.

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 carried out 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 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. The distribution of 27 color traits is presented in Figure S1 of the supplementary information.

Correlation between two color descriptors

Two descriptors with different numbers of codes were used to characterize the color traits in the CCC during 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 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 the color, we selected traits with a medium to strong association, indicating a genetic influence.

Color and nutrition variables association

This study also examined the association between color traits and nutritional values using data from a previous study by Berdugo-Cely et al. (2023). The nutritional content of the potatoes was represented by four variables: Total Phenol Content (TPC), Antioxidant Activity (measured by DPPH and FRAP assays), and Ascorbic Acid Content (AAC). Given that the nutritional variables were continuous and the color traits 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

An analysis was conducted to explore associations between potato LCH traits and genomic regions for both additive and dominant effects. We used MultiGWAS software (Garreta et al., 2021), which runs and integrates the results of three GWAS tools: GWASpoly (Rosyara et al., 2016) for tetraploid genotypes, and GAPIT (Tang et al., 2016) and TASSEL (Bradbury et al., 2007) for diploid genotypes.

The analysis employed a full model, or Q+K model, which accounts for population structure and the relationships between samples by using built-in algorithms to calculate both principal components as covariates and kinship among pairs of individuals. The GWAS tools utilize a mixed linear model approach (MLM: Phenotype + Genotype + Structure + Kinship), considering population structure and kinship information. This allows for the detection of marker-trait associations while controlling for confounding factors.

To address the issue of multiple testing, the MultiGWAS tool applies the Bonferroni correction. However, instead of adjusting the *p-values*, MultiGWAS adjust the threshold for determining 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 for visualizing significant SNP markers.

Candidate Marker Identification

Given the evaluation of multiple traits and the integration of results from four GWAS tools in MultiGWAS, we developed a proprietary scoring function called GSCORE. This function is designed to rank markers by incorporating the outputs from MultiGWAS. Specifically, GSCORE selects the top 50 markers with the lowest p-values from each tool, resulting in 200 markers (3 tools $\times$ 50 markers). The GSCORE is calculated as the sum of three weighted terms: Inflation factor (I), contributing 70% of the score. Replicability (R), accounting for 10%, and Significance (S), making up 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 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), indicating whether the SNP is significant (p-value ¡ threshold). A marker M achieves a high $GSCORE$ when it has an inflation factor $\lambda(M)$ close to 1, 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, the shared SNP count, and the 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 SpudDB. 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 69K SNP positions on the DMv6. 1 assembly. Moreover, for each trait, the amino acid sequences were analyzed using BlastKOALA (Kanehisa et al., 2016) for KEGG mapping to identify functional categories common among significant markers.

2.4 Heritability

Narrow-sense heritability h^2, defined as the proportion of phenotypic variance explained by additive effects (de los Campos et al., 2015), was estimated for all LCH components by the following equation:

$$h^2 = \frac{\sigma_a^2}{\sigma_y^2}$$

where σ_a^2 is the additive genetic variance and σ_y^2 is the phenotypic variance. These variances were obtained after fitting a whole genome regression model using Bayesian regression implemented in the BGLR package of R (Pérez and de los Campos, 2014).

2.5 Genomic Prediction

We used 12 Genomic Prediction (GP) models to use a large number of genetic markers to generate genomic estimated breeding values (GEBVs) for each of the 27 LCH traits. The GP models included parametric (GBLUP, EGBLUP, RR, and LASSO), semiparametric (RKHS, RF, and SVM), and Bayesian (BRR, BL, BA, BB, and BC) models. Additionally, to identify the optimal number and type of markers for estimating these GEBVs, various marker subsets generated through the GWAS process were tested.

The R package for the Breed Wheat Genomic Selection Pipeline (BWGS) (Charmet et al., 2020) was used. Additionally, 5-fold cross-validation, repeated five times, was employed to assess the performance of the 12 models for each LCH trait. The parameter values for running each model were set to the default settings provided by the BWGS library.

2.5.1 GP using marker subsets

Marker subsets derived from the GWAS process were created and utilized in genomic prediction (GP) to determine the optimal number of markers for predicting GEBVs for each LCH component of the color traits. The selected markers were ranked by importance based on our custom GSCORE scoring function(see Section 2.3).

3 Results

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

3.1 Correlation between descriptors

The correlations between the values of the color traits measured with two color descriptors over different years are presented in Figure 3 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, are concentrated along the diagonal of the correlation matrix. In particular, there are correlations

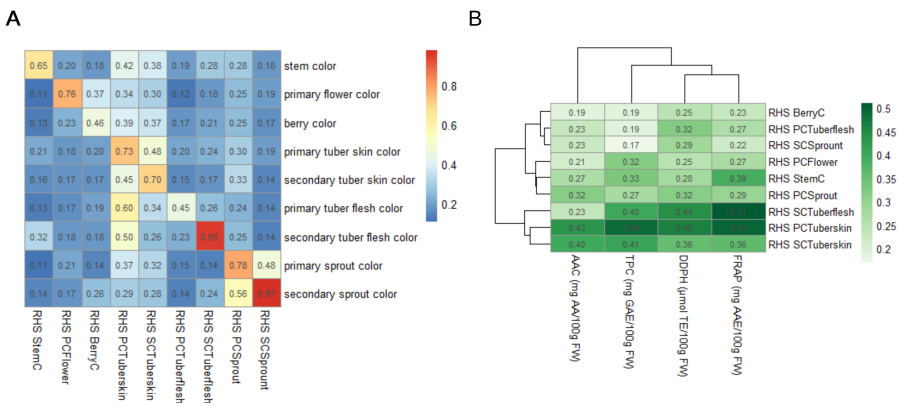

Fig. 3 **A.** Heatmap showing the correlation matrix of potato color traits in CCC accessions, derived form descriptors by Gómez (2000) and measured with RHS (Voss, 2002) color chart. The color bar is located in the top right, with blue indicating low correlation values and red representing high correlation values. **B.** Heatmap showing the determination coefficient between nutritional components and potato RHS color traits. The color bar is located in the top right, with the intensity of green representing the determination coefficient.

.

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

3.1.1 Color and Nutrition

In Figure 3 B, the heatmap presents the determination coefficient, R^2, between the color traits measured using the RHS color scale and nutritional variables. R^2 below 0.3 reflect a weak association (blue cells), while R^2 between 0.3 and 0.51 reflect 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 the nutritional variables. Antioxidant activity (measured by FRAP and DPPH assays) showed the strongest correlations with these traits. Additionally, Total Phenol Content and Ascorbic Acid Content demonstrated the highest association with the primary color of the tuber skin. The stem color, meanwhile, showed a moderate association with antioxidant activity and Total Phenol Content.

3.2 GWAS Analysis

The GWAS analysis identified significant SNPs associated with five traits: the primary color of the flower, the secondary color of the tuber flesh, and the primary and secondary colors of the sprout, and stem color. Manhattan plots for all seven traits are presented in Figure 4 (A-G). The majority of significant SNPs were associated with the Hue component. However, for the secondary color of the sprout, significant SNPs were also identified for the Chroma and Lightness components.

Notably, chromosome 6 accounted for 32% of the significant SNPs, followed by chromosomes 7, 8, and 12, each contributing more than 10% of the significant SNPs. For more details, a Circos plot (Figure S2) combines the data for all seven traits and presents the density of SNP markers for chromosomes, allowing visualization of SNP co-occurrence across different traits.

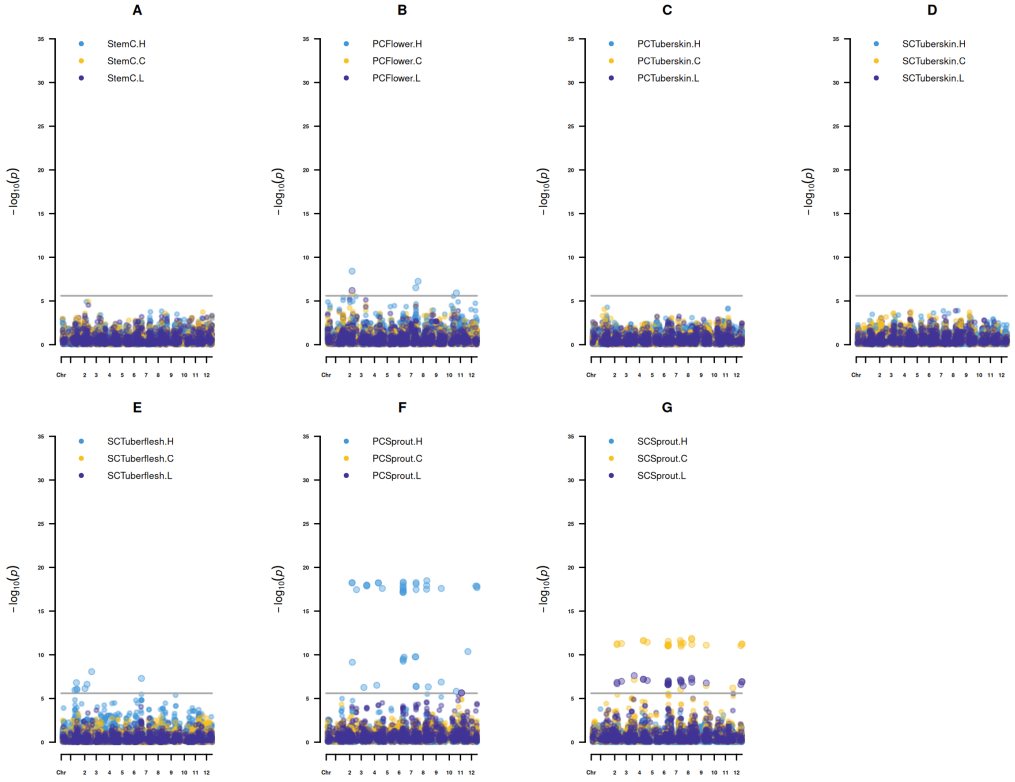

Fig. 4 (A-G) Manhattan plots display the markers associated with LCH traits, with colors representing the different LCH color components.

A total of 175 significant associations were identified in 21 HCL components for seven traits, represented by 87 SNP markers. Of these 175 associations, 166 were annotated to 242 genes. Information on the primary annotations linked to these markers, as retrieved from the SPUD database, is provided in Supplementary Table 1. A total of 18 functional categories were identified for the pathways associated with genes annotated to the significant SNPs. Four of these functional categories were present in more than 70% of the traits with significant annotations: *Protein families: genetic information processing, Carbohydrate metabolism, Protein families: metabolism,* and *Environmental information processing.* The proportions of specific functional categories per trait are provided in Supplementary Table 2.

3.3 Heritability

Overall, most traits exhibit high average heritability values (greater than 0.5), with the exception of the secondary color of the sprout. Figure 5 A. shows the heritability of selected LCH traits. The primary color of the flower and the secondary color of the tuber flesh exhibited high heritability values (greater than 0.9) for two LCH traits. Furthermore, the color of the stem showed a heritability that exceeded 0.8.

3.4 Genomic Prediction

The results for genomic prediction (GP) using all 4,641 markers from the potato dataset are in Figure 5 B. The predictive ability corresponds to the correlation between the observed phenotype values and the predicted genomic estimated breeding values (GEBV) in the validation set. The primary flower color and stem color exhibited the highest predictive abilities, followed by the primary color of the tuber skin and the primary color of the tuber sprout. However, the secondary colors of the sprout, tuber flesh, and tuber skin showed low predictive abilities.

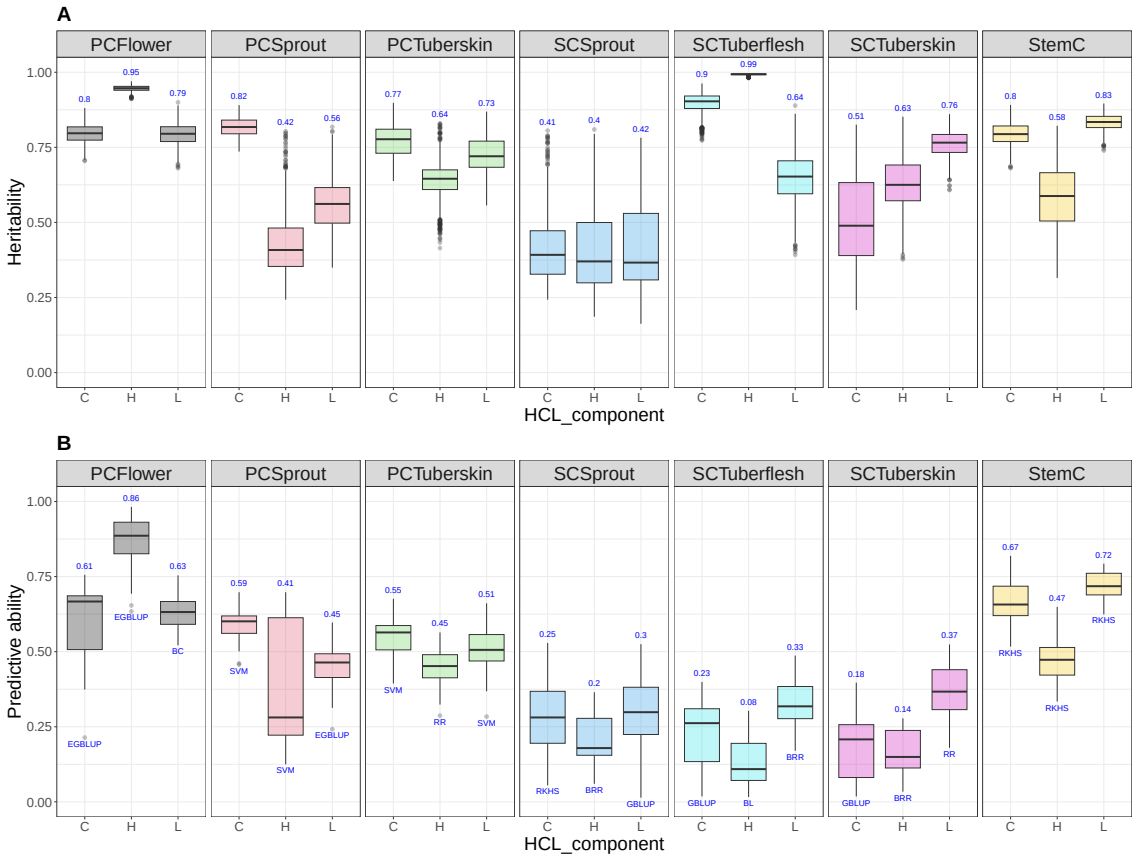

Fig. 5 A . Estimated heritabilities for LCH traits. At the top of each division is the name of the potato color trait, while at the bottom is the LCH component of that trait being evaluated. At the top of each boxplot, in blue, is the mean value of the heritability.. **B. Predictive abilities of GS models for LCH traits.** Comparison of genomic predictions for LCH components using the full set of markers. The horizontal axis represents the three LCH components for each of the seven potato color traits, while the vertical axis displays the predictive ability values, ranging from 0 to 1. The mean predictive ability is displayed at the top of each boxplot, while the model that achieved the best prediction is indicated at the bottom.

3.4.1 GP using marker subsets

Evaluations were performed on five subsets containing the top 5, 15, 35, 50, and 100 markers, as shown in Figure S3. Predictive ability increases with the number of markers; however, it tends to decline beyond 50 markers. Based on the GP results obtained from marker subsets, the top 50 markers generally produced predictive ability values comparable to those achieved with the full set of 4,641 markers. Refer to Figure S4 for the evaluation results of GEBV prediction using the top 50 markers identified through GWAS.

References

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)

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

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S.: TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23**(19), 2633–2635 (2007) https://doi.org/10.1093/bioinformatics/btm308

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)

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)

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)

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)

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

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

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)

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)