frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 March 2024
DOI 10.3389/fpls.2024.1340189

OPEN ACCESS

EDITED BY
Dilip R. Panthee,
North Carolina State University, United States

REVIEWED BY
Ram Kumar Basnet,
Rijk Zwaan, Netherlands
Rodomiro Ortiz,
Swedish University of Agricultural Sciences, Sweden

*CORRESPONDENCE
Trine Aalborg
traa@bio.aau.dk

†PRESENT ADDRESS
Arla Foods Ingredients,
Videbæk, Denmark

RECEIVED 17 November 2023
ACCEPTED 14 February 2024
PUBLISHED 08 March 2024

CITATION
Aalborg T, Sverrisdóttir E, Kristensen HT and Nielsen KL (2024) The effect of marker types and density on genomic prediction and GWAS of key performance traits in tetraploid potato.
Front. Plant Sci. 15:1340189.
doi: 10.3389/fpls.2024.1340189

The effect of marker types and density on genomic prediction and GWAS of key performance traits in tetraploid potato

Trine Aalborg*, Elsa Sverrisdóttir, Heidi Thorgaard Kristensen† and Kåre Lehmann Nielsen

Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark

Genomic prediction and genome-wide association studies are becoming widely employed in potato key performance trait QTL identifications and to support potato breeding using genomic selection. Elite cultivars are tetraploid and highly heterozygous but also share many common ancestors and generation-spanning inbreeding events, resulting from the clonal propagation of potatoes through seed potatoes. Consequentially, many SNP markers are not in a 1:1 relationship with a single allele variant but shared over several alleles that might exert varying effects on a given trait. The impact of such redundant "diluted" predictors on the statistical models underpinning genome-wide association studies (GWAS) and genomic prediction has scarcely been evaluated despite the potential impact on model accuracy and performance. We evaluated the impact of marker location, marker type, and marker density on the genomic prediction and GWAS of five key performance traits in tetraploid potato (chipping quality, dry matter content, length/width ratio, senescence, and yield). A 762-offspring panel of a diallel cross of 18 elite cultivars was genotyped by sequencing, and markers were annotated according to a reference genome. Genomic prediction models (GBLUP) were trained on four marker subsets [non-synonymous (29,553 SNPs), synonymous (31,229), non-coding (32,388), and a combination], and robustness to marker reduction was investigated. Single-marker regression GWAS was performed for each trait and marker subset. The best cross-validated prediction correlation coefficients of 0.54, 0.75, 0.49, 0.35, and 0.28 were obtained for chipping quality, dry matter content, length/width ratio, senescence, and yield, respectively. The trait prediction abilities were similar across all marker types, with only non-synonymous variants improving yield predictive ability by 16%. Marker reduction response did not depend on marker type but rather on trait. Traits with high predictive abilities, e.g., dry matter content, reached a plateau using fewer markers than traits with intermediate-low correlations, such as yield. The predictions were unbiased across all traits, marker types, and all marker densities >100 SNPs. Our results suggest that using non-synonymous variants does not enhance

the performance of genomic prediction of most traits. The major known QTLs were identified by GWAS and were reproducible across exonic and whole-genome variant sets for dry matter content, length/width ratio, and senescence. In contrast, minor QTL detection was marker type dependent.

KEYWORDS

***Solanum tuberosum*, genomic prediction, GBLUP, tetraploid potato breeding, GWAS, marker density, marker type**

1 Introduction

Since its original domestication from Peruvian wild species progenitors into Andean and Chilean landraces (Spooner et al., 2005, 2014), the cultivated potato has been globally disseminated, and *Solanum tuberosum L.* is currently the world's third most important food crop (FAOSTAT, 2023). Its high efficiency in energy yield per cultivated area and high nutritional value compared to cereals have resulted in a cosmopolitan growth distribution (Wilson et al., 2021), and potato remains a crop of key interest for future global food security. However, to accommodate the expected 35%–56% increase in food demand by 2050 compared to 2010 (Van Dijk et al., 2021), as well as the exhaustion of fertile farmlands and imminent climate changes, it is paramount to intensify and accelerate the development of more sustainable crop strains (Lenaerts et al., 2019) with improved yield, pest resistance, and space/nutrient/water use efficiency.

The rapid progression of genome sequencing technologies and the availability of reference genomes for potato (The Potato Genome Sequencing Consortium, 2011; Sharma et al., 2013; Pham et al., 2020) have paved the way for the implementation of genome-assisted breeding methods, such as genomic selection (GS), in potato crop breeding (Hickey et al., 2017). Genomic selection breeding relies on a set of thousands of genome-wide markers, obtained by, e.g., single-nucleotide polymorphism (SNP) microarrays or genotyping by sequencing (GBS), and it is then assumed that all quantitative trait loci (QTL) are in linkage disequilibrium with at least one marker (Meuwissen et al., 2001). The marker effects on traits are then estimated using a model trained on a panel of genotypes and phenotypes, and the model is used to calculate the genomic estimated breeding values (GEBVs) from genotype data on a breeding population, facilitating directed parent crossing and early selection of breeding candidates without the need for direct phenotyping (Heffner et al., 2009). GS is particularly useful for traits with complex, multigenic inheritance patterns, such as chipping quality and dry matter content, where optimization has been largely ineffective in traditional breeding by phenotype-directed sexual crosses (Wilson et al., 2021). Several studies have already evaluated the performance of different statistical methods and machine learning approaches for the reliable prediction of genomic estimated breeding values for a host of agronomic performance traits, training on a vast collection of elite breeding materials (Slater et al., 2016; Sverrisdóttir et al., 2017, 2018; Stich and Van Inghelandt, 2018; Byrne et al., 2020; Selga et al., 2021; Wilson et al., 2021; Ortiz et al., 2022, 2023; Pandey et al., 2023). However, the development of strategies for enhancing the power of GS is still highly relevant for potato breeding.

While the crop fitness traits can have highly different genetic characteristics and heritabilities, all phenotypic variations are underpinned by causal alleles with either deleterious or advantageous trait effects under selection pressure, the former dominating in frequency in domesticated lineages (Zhu et al., 2022). In contrast to diploid crops, deleterious alleles that, e.g., induce frameshift mutations, create non-synonymous base changes, reform splice sites, or generate alternative stop codons, are ineffectively purged from breeding population gene pools during purifying selection due to the autotetraploid status of crop potato (Pham et al., 2017). In addition, insufficient recombination events in clonal propagation (Zhu et al., 2022) and the generally recessive nature of those alleles (Dwivedi et al., 2023) have led to a high abundance of deleterious alleles during domestication and clonal propagation, which is why elite potato, despite their high heterozygosity, exhibits signs of acute inbreeding depression (The Potato Genome Sequencing Consortium, 2011; Zhang et al., 2019). At the same time, the fixation of a few targeted desirable, recessive alleles in a single breeding line or population is extremely challenging in the case of tetrasomic inheritance (Muthoni et al., 2015). For breeding purposes, the identification of the subset of deleterious and beneficial alleles that manipulate phenotypic variation would, in principle, allow more accurate genomic prediction modeling based on the high information-carrying diagnostic markers alone (Zhang et al., 2018; Ramstein and Buckler, 2022). However, the complex multigenic signature and/or low heritability of several tuber quality traits and yield combined with the considerable environmental component of the observed phenotypes of such traits have complicated full genetic trait characterization (van Eck, 2007). GS models are instead traditionally based on a high density of anonymous genome-wide markers (Heffner et al., 2009), but the introduction of large amounts of redundant data has unknown consequences for model performance. An issue of similar concern arises for genome-wide

association studies (GWAS). GWAS can be used to identify the subset of the total genetic variation that underpins target tuber and plant quality traits (Naeem et al., 2021). Elucidation of trait genetic architectures and trait-associated markers by this method as well as QTL mapping can be used to direct breeding efforts and selection models toward high-impact alleles, and several markers for the selection of tuber traits have already been identified (Fischer et al., 2013; D'hoop et al., 2014; Schreiber et al., 2014; Li et al., 2019; Byrne et al., 2020; Zia et al., 2020; Park et al., 2021; Ahmad et al., 2022). However, the sheer quantity of genetic variation in the potato genomes with, on average, one SNP per 29 bp (The Potato Genome Sequencing Consortium, 2011) represents a challenge. Though most variants will have a neutral effect and not impact trait phenotype, the inclusion of these observations results in reduced power of association due to increased stringency of correction for multiple testing but might not necessarily contribute to proportional amplifications of the QTL signal intensities.

Intuitively, reducing the background of redundant marker density by selecting only non-synonymous variants, which are expected to underpin the largest genetic contribution to phenotypic variance, can be expected to increase genomic prediction accuracy. While marker reduction to a set of non-synonymous variants is certain to reduce the stringency of correction in GWAS, it is also possible that including only functional effect variants will improve the signal power of additional minor effect QTLs by concentrating high-information variants. Following this hypothesis, we evaluated how GS model prediction accuracy and the resolution of GWAS were impacted by using different single-nucleotide polymorphism marker subsets: (i) amino acid-changing SNPs within protein-coding genes, non-synonymous SNPs (nsSNPs), (ii) amino acid-conserving SNPs within protein-coding genes, synonymous SNPs (sSNPs), (iii) SNPs located outside exons, non-coding SNPs (ncSNPs), and (iv) a combination of all types, as well as assessing the effect of marker reduction on GS models. We genotyped a panel of 762 clones called MASPOT by GBS, used in previous studies (Sverrisdóttir et al., 2017, 2018), and trained GBLUP models to predict GEBVs for five agronomic traits with different heritabilities and modes of inheritance, namely, chipping quality, dry matter content, yield, length/width ratio, and senescence using each of the filtered marker subsets in order to evaluate model response to marker type and robustness to marker reduction. We used single-trait GWAS to identify the associated loci for each of the five traits and assessed whether the information level of the trait genetic architecture carried by different functional variants differed.

2 Materials and methods

All statistical analyses and graphics were performed using R Statistical Software (v4.3.1) (R Core Team, 2023) in RStudio (v2023.6.2.561) (Posit team, 2023). Graphics were generated using the ggplot2 package (v3.4.3) in R (Wickham, 2016) unless otherwise stated.

2.1 Plant material

A mapping population called the MASPOT population, consisting of circa 5,000 offspring, was established at the LKF Vandel breeding station (presently Danespo A/S) in Vandel, Denmark. The MASPOT population was generated by the systematic cross-pollination of 18 elite potato cultivars in a full-diallel crossing design, the parents being either established cultivars or advanced breeding clones (Sverrisdóttir et al., 2017). The design was, however, limited by low fertility in specific crosses and male sterility in some of the parents. Male sterility is not an unusual trait in elite potato cultivars (Sanetomo and Gebhardt, 2015). A total of 762 clones were chosen randomly from the full mapping population and is henceforth referred to as the MASPOT panel in this paper (Supplementary Figure 1). The 762 offspring were planted in field trials in Vandel, Denmark, in 2013 and 2014 as described in Sverrisdóttir et al. (2017, 2018). The plants were grown to a plant density of approximately 40,000 plants/hectare with 30 cm between plants and 75 cm between rows. In 2013, the tuber seedlings were planted in April 24 and 25 (no replicates) in 24-parcel blocks and harvested in August 11–29 (109–128 days after planting). The plants were desiccated 1 to 2 weeks before harvest. No checks were used. In 2014, the clones were divided into four groups based on parent earliness and planted in a randomized 28-parcel block design with two replicates. The groups were planted in April 24, 25, 28, or 29 and harvested in August 11–29 (109–129 days after planting), also with 1 to 2 weeks of desiccation. The groups were harvested in chronological order. A total of 19 checks were planted in two replicates, 18 of which were the MASPOT panel parents. The checks were inspected manually for signs of unusual development/disease infection. No abnormalities were observed, taken as an indication of credible plant material for all clones. As the population was highly diverse, not all plants had fully matured at harvest. The soil type was sandy loam. Fertilization was performed with 1,000 kg/hectare NPK 14-3-15. Pests and diseases were controlled with Fenix and Titus (weed) before and immediately after sprouting, Mospilan (insects) ultimo June and again ultimo July, and alternating Ranman and Revus (late blight) from approximately June 23 until desiccation as needed, depending on the weather. The fields were irrigated as needed. Additional details about the population can be found in Sverrisdóttir et al. (2017). Since the propagation of the population happened simultaneously with the trials, the number of seed tubers available in 2013 were lower than in 2014. Therefore, we have not used a formally established incomplete block design. Furthermore, the trial data is not corrected for soil heterogeneity due to the following reasons: (i) Danish regulations for crop rotation do not allow growing potatoes at the same site in two consecutive years and (ii) since significant senescence variation was observed in 2013, the plants were

grouped by senescence in 2014 to minimize "neighbor vigor effects" in the trials. As a consequence, the data could be corrected reliably for soil heterogeneity.

2.2 Phenotyping and adjustment for environmental effects

Phenotyping of dry matter content and chipping quality is described in Sverrisdóttir et al. (2017). In short, dry matter content [%] was determined for the MASPOT panel and parent clones harvested in 2013 (one replicate) and 2014 (two replicates). The tubers were washed, and a basket holding 1.5–10 kg of tubers was weighed above and under water shortly after harvesting. The dry matter content was then calculated using the following empirical equation:

$$DM[\%] = 214 \cdot \left(\left(\frac{weight\ \ in\ \ air}{(weight\ \ in\ \ air) - (weight\ \ in\ \ water)} \right) - 0.988 \right)$$

Chipping quality was determined as the chip color following frying in oil after the cold storage of tubers. Phenotyping of chipping quality was performed only for the 2013 harvested clones. The tubers were stored at either 4°C for roughly 2 months, after which they were incubated at ambient temperature 2–6 h prior to frying. Four to six slices (1 to 2 mm) of each tuber were fried in sunflower oil at 180°C until the bubbles ceased to emerge (generally 2 to 3 min). The frying color was visually assessed to a standard set on an arbitrary grading scale from 1 (dark) to 9 (light).

Yield was measured for the 2014 harvested clones in the field at harvest as the total weight of five tubers from each clone in two replicates, i.e., 2 × 5 tubers each. The weight was converted into hkg/ha values, assuming 40,000 plants/hectare to account for plot variations from year to year.

Length/width ratio was determined as the length/width ratio (LW) for the 2014 harvested clones:

$$LW = \frac{length}{width}$$

Tuber length was defined as the longest measure and the width as the measure perpendicular to this and measured on a SCOUT camera (Newtec A/S, No. 0213). The measures do not consider tuber anatomy, where length is defined as the distance from the rose (apex) to the heel (attachment of stolon). The true definition will only be violated for irregular tubers, which are relatively rare and hence assumed to not significantly affect the downstream analyses. Outliers were identified on the length or width (diameter) parameter relative to the nearest neighbor and removed following Dixon's Q test (Dean and Dixon, 1951) before the calculation of length/width ratio, where the critical confidence level, Q_{crit}, was estimated for batches of up to 200 tubers by regression and used for a two-tailed test as outlined by Rorabacher (1991).

Senescence (an earliness proxy) was scored manually on a scale from 9 (late senescence) to 1 (no senescence). Phenotyping was performed for the 2013 and 2014 harvested clones. The scoring was performed temporally at three points, splitting the scale accordingly: the first scoring (when the first cultivars begin dying off) used the upper end of the scale (9-8-7), the second used the middle of the scale (6-5-4), and the third scoring, capturing cultivars that display late senescence, used the lower end of the scale (3-2-1). Each clone was only scored once.

All phenotypic data were corrected for variation across years by fitting a linear mixed model to the phenotypic data *via* restricted maximum likelihood (REML) using the following model:

$$y_{ij} = \mu + genotype_i + year_j + e_{ij}$$

where y_{ij} is the observed phenotype, μ is the overall mean, $genotype_i$ is the random effect of the ith genotype, $year_j$ is the fixed effect of the jth year, and e_{ij} is the error term (Sverrisdóttir et al., 2017). The model was made with the lme4 package in R (Bates et al., 2015). Terms for genotype-by-environment (G × E) were not included in either the GS or GWAS models since the experimental design did not produce sufficiently robust phenotyping to allow the rigorous estimation of this in the MASPOT panel (Sverrisdóttir et al., 2017, 2018), which is why the model was simplified to avoid infusion of additional error and lessen the risk of overfitting.

2.3 Genotyping

Genotyping was performed by GBS. GBS libraries were prepared according to Sverrisdóttir et al. (2017), following a protocol adapted from Elshire et al. (2011). The 5′ and 3′ adapters for Illumina sequencing were designed for a 96-multiplexing system. DNA was extracted from leaf tissue and digested with *ApeKI*. The fragments were ligated to adapters, pooled in 96-plex libraries, purified, and amplified by PCR. The MASPOT panel libraries were sequenced on a HiSeq 2000 (Illumina, San Diego, CA, USA) with single-read sequencing (100 bp), and each 96-plex library was sequenced on three channels on a flow cell.

2.4 Filtering raw sequence data, mapping, and SNP calling

Sequenced reads were processed as described in Sverrisdóttir et al. (2017). The reads were demultiplexed, trimmed, and mapped onto the *S. tuberosum* Group Phureja reference genome sequence [DM v4.03; (Sharma et al., 2013)]. SNPs were called using the Genome Analysis Toolkits (McKenna et al., 2010) UnifiedGenotyper tool with ploidy set to 4 and the minimum phred-scaled confidence threshold of 50 for variant calling and of 20 for variant omission (and filtered with LowQual if less than the calling threshold), as described in Sverrisdóttir et al. (2017). The SNPs were then filtered to a root mean squared quality of 30, including only biallelic variants. Since potatoes are not tetraploid across all loci, but rather have a mean gene copy number of 3.2 (Sun

et al., 2022), enforcing the expectation of tetraploidy across all marker sites would constitute a confounding error. As a consequence, the called tetraploid genotypes were not used explicitly, but rather variant allele frequencies estimated from the sequencing data for each variant were used directly as genotypes for statistical analyses cf. (Ashraf et al., 2016) to accommodate the gene copy number variation across the potato genome. Minor allele frequency (MAF) was calculated from read coverage, and SNPs were filtered to a MAF of 1% (average variant frequency<0.99 and >0.01), a read coverage >5, and a missing rate of maximum 50%.

2.5 Marker reduction and filtering

SNPs were annotated using SnpEff (Cingolani et al., 2012) using a custom database built from the *S. tuberosum* DM v4.03 reference genome. The SNPs were filtered into three subsets based on annotation: non-synonymous (including missense, stop codon gain/loss, start codon gain/loss, frameshift, and exon loss variants), synonymous (including synonymous variants), and non-coding (excluding all exonic variants, reduced to every third non-coding variant), in addition to a combination set of the former three subsets. For GWAS, a 1-in-3 reduction of the combination set was also analyzed. For each of the four annotated SNP sets, the SNPs were further filtered to read coverage between 5 and 60, while individuals with >70% missing data were removed.

Reduced marker sets, for the evaluation of genomic prediction model tolerance to marker reduction, were prepared for each of the four SNP sets by iteratively reducing the sets to every other position, resulting initially in two bins. This was done to ensure whole-genome dispersion of the SNPs and avoid the introduction of regional bias. Following the second iteration, resulting in four bins, only four bins were kept for each iteration. Iterations were performed this way until ~150 markers remained in each SNP set. A final reduction to every 10th marker was then performed on each set. Supplementary Figure 2 presents an overview of the marker reduction strategy. The SNP density plots were generated using the CMplot package in R (LiLin-Yin, 2023).

2.6 Statistical analyses

2.6.1 Assessment of population structure

The population structure of the MASPOT panel was ascertained by performing a principal component analysis (PCA) using the prcomp function of the built-in stats package (R Core Team, 2023) on the genomic relationship matrix (**G**) computed from the combination set of annotated SNPs (93,170 SNPs) (Supplementary Figure 3). The genomic relationship matrix was created from the genotype matrix (**Z**) based on the first VanRaden (2008) method. The **Z** matrix contains the genotypes taken as allele frequencies for each sample and SNP from sequence data (Ashraf et al., 2016). The allele frequencies were calculated as the ratio between allele counts of the alternative allele and the total allele count, producing a value between 0 and 1. This allows the genotype matrix to capture tetraploid allele dosages.

$$AF = \frac{AC_{alt}}{AC_{ref} + AC_{alt}}$$

The allele frequencies were corrected for missing data following the correction, w_i, described by VanRaden (2008):

$$w_i = \sqrt{\frac{\sum p_k(1-p_k) \quad over \quad all \quad loci}{\sum p_k(1-p_k) \quad over \quad only \quad non-missing \quad loci}}$$

where p_k is the mean allele frequency at locus k. A total of 16.26% of all markers were imputed. The genotype matrix was centered and adjusted for missing values according to Ashraf et al. (2016), whereafter means were set to zero, corresponding to mean imputation for missing data.

$$\mathbf{Z}_{ik} = (\mathbf{X}_{ik} - p_k) \cdot w_i$$

where $\boldsymbol{X}_{ik}$ is the allele frequency in family i at locus k. The genomic relationship matrix was computed from **Z** using global scaling, following method 1 of VanRaden (2008), with a modification to adjust for tetraploidy (Ashraf et al., 2014, 2016).

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{0.25\sum p_k(1-p_k)}$$

where $0.25\sum p_k(1-p_k)$ is the sum of genotypic variance and also the average diagonal of $\boldsymbol{ZZ}'$.

2.6.2 Genomic prediction models

Genomic predictions for each of the single traits, using each of the generated SNP sets, were performed using a standard additive GBLUP model, equivalent to a ridge-regression with uniform shrinkage of SNP effects, without accounting for marker effect size, i.e., assuming that each marker accounts for an equal proportion of the total genetic variance, though shrinkage is dependent on sample size and allele frequency (Gianola, 2013). A G × E term was not included due to insufficient robustness of across-year phenotyping (Sverrisdóttir et al., 2017). GEBVs are directly estimated using the genomic relationship matrix (Meuwissen et al., 2001):

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{g} + \mathbf{e}$$

where $\boldsymbol{y}$ is a vector of observed phenotypes, μ is the mean, $\boldsymbol{e}$ is a vector of residual effects with $\mathbf{e} \sim N\left(0,\ \mathbf{I}\sigma_e^2\right)$, where **I** is an identity matrix and σ_e^2 is the residual variance, and $\boldsymbol{g}$ is a vector of random genomic breeding values with distribution $\mathbf{g} \sim N\left(0, \mathbf{G}\sigma_g^2\right)$ is the genetic variance of the model. All models were computed using the BGLR package in R (Pérez and de los Campos, 2014) with default settings for priors and settings of 12,000 iterations and a burn-in of 2000. All analyses were performed using an eightfold cross-validation scheme, where clones were randomly divided into eight groups, one group being used for validation while the model was trained using the data of the seven remaining groups. This process was repeated, each time with a different group as validation set, until predictions had been calculated for all individuals. Each analysis was repeated with 10 different cross-validation groupings, and the GEBV was calculated as the average across all samplings. The accuracy of the GEBVs was determined as the Pearson correlation

coefficient between the predicted GEBVs and the observed phenotypes, described here as the prediction correlation:

$$r(\mathbf{GEBV}:\mathbf{y})$$

Correlation coefficients for each trait, when using ~30k markers, were compared pairwise by Welch two-sample *t*-test with Bonferroni correction of significance level to 0.05/*N*, where *N* is the total number of tests for each trait. A linear regression of the observed phenotypes on the predicted values was used as a measure of bias of the GEBVs, where a regression slope of β=1 indicates no bias, β<1 implies that extremely high (low) GEBVs over-(under) estimate the observed phenotype and *vice versa* for β>1 (Luan et al., 2009). The prediction correlation and bias summary statistics were evaluated for all modeled SNP sets.

2.6.3 Heritability

The pedigree narrow sense heritabilities (h_p^2) were estimated for each trait as the linear regression coefficient of the mid-parent phenotypic value (i.e., mean parental phenotype) against the offspring value. The offspring of one or more parent with missing phenotypic data was not included. Genomic narrow sense heritability was estimated as the ratio of genomic to phenotypic variance using the genomic relationship matrix of the full combination data set in a REML analysis (de los Campos et al., 2015).

$$h_g^2 = \frac{\sigma_g^2}{\sigma_y^2}$$

2.6.4 Genome-wide association studies

Genome-wide association studies were conducted for each of the four fully annotated SNP sets. Additionally, for the combination SNP set, a reduced dataset was prepared by taking every third SNP, generating a reduced set of 31,032 SNPs, for which GWAS was also conducted. GWAS was performed by single marker regression with the regress package in R (Clifford and McCullagh, 2006, 2020) using the following model for each SNP in the respective data subsets, cumulating the marker effects:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X_i}\beta_i + \boldsymbol{g} + \mathbf{e}$$

where $\boldsymbol{y}$ is a vector of observed phenotypes, μ is the mean, $\boldsymbol{X_i}$ is the vector of SNP genotype values taken as allele frequency at the *i*th position, β_i is the corresponding additive genetic effect of the *i*th SNP, $\boldsymbol{g}$ is a vector of random genomic breeding values with distribution $\mathbf{g} \sim N\left(0, \mathbf{G}\sigma_g^2\right)$, where $\boldsymbol{G}$ is the genomic relationship matrix of Sverrisdóttir et al. (2017) computed for the MASPOT panel, and each SNP set, σ_g^2, is the genetic variance of the model, and $\boldsymbol{e}$ is a vector of residual effects with $\mathbf{e} \sim N\left(0, \mathbf{I}\sigma_e^2\right)$. The genomic relationship matrix was included as a fixed effect in the model to correct for the genomic relationship between the MASPOT offspring. For each chromosome, a **G**-matrix was calculated based on the SNPs mapped to the remaining chromosomes, excluding the target chromosome. This **G**-matrix was then used to correct for population structure for the SNPs of the excluded chromosome to ensure that SNPs were not included in the model twice (Kristensen et al., 2018).

Potatoes are clonally propagated, and modern cultivars can be expected to have a significant proportion of recurrent genetic variation in their pedigree. This leads to a population structure which, in turn, has a tendency to cause overdispersion of the test statistics in association analyses (Devlin et al., 2001). This would result in an increased number of false positive associations. To adjust for this, we calculated genomic inflation factors, λ_{gc}, for each trait and SNP subset and used to correct *p*-values for inflation due to systematic effects not captured by the model according to Hinrichs et al., (2009) and Kristensen et al. (2018). Inflation factors were computed as the median value of the chi-squared statistic of the SNPs divided by the expected median value, i.e., assuming no association between the SNPs and the trait.

$$\chi^2 = Q_{\chi^2}^{-1}(P,\ 1)$$

$$\lambda_{gc} = \frac{median\left(\chi^2\right)}{Q_{\chi^2}^{-1}(0.5,\ 1)}$$

Each *p*-value (*P*) is converted to a χ^2 quantile using the quantile function of the chi-squared distribution, i.e., the inverse of the cumulative distribution function (CDF) Q_{χ^2}, with one degree of freedom. To determine the genomic inflation factor, λ_{gc}, the median of the χ^2-quantiles is then divided with the chi-square of the 50th percentile with one degree of freedom, i.e., the expected median χ^2, assuming no SNP–trait association.

For λ_{gc}>1, the chi-squared quantile of the *p*-values was divided by the inflation factor and used to calculate corrected *p*-values using the CDF of the chi-squared distribution with one degree of freedom.

$$P_{corrected} = 1 - Q_{\chi^2}\left(\left(\frac{\chi^2}{\lambda_{gc}}\right), 1\right)$$

To control false positive associations, Bonferroni correction was used with a false discovery rate of $p < 0.05/N$, where 0.05 is the overall significance threshold and *N* is the total number of markers tested in the analysis. The proportion of phenotypic variance explained by markers was calculated using the formula from Shim et al. (2015). QQ and Manhattan plots were plotted using the qqman package in R (Turner, 2018).

3 Results

3.1 Genotyping statistics

Sequencing yielded an average of four million trimmed and filtered reads per sample for the MASPOT panel of 762 clones. A total of 3.4 million variant sites were found. Following filtering for MAF > 1%, minimum coverage of five, and a missing rate of maximum 50%, 182,757 variants remained with sequence positions as in Sverrisdóttir et al. (2017).

3.1.1 Filtering and reduction of markers according to SNP annotation

Four subsets of SNP data were created from the annotation filter: (1) non-synonymous variants (32,352 nsSNPs), synonymous (34,695 sSNPs), non-coding (33,743 ncSNPs), and a combination set (100,790 of all SNPs). After filtering each data set to read coverage between 5 and 60 and individuals with >70% missing data, the data sets were reduced as presented in Table 1 (Supplementary Files 1–4).

Markers were well distributed over each of the 12 chromosomes in each of the four annotation-based data sets (Figure 1, Supplementary File 5), consistent with marker density distributions found in another study of tetraploid potato (Wilson et al., 2021). The markers most densely populated the apocentromeric regions (The Potato Genome Sequencing Consortium, 2011) in all sets, with the two exonic variant sets presenting with Mb-sized windows of SNP sparsity in and around the centromeres. This is consistent with low gene density (Sun et al., 2022) and repressed meiotic recombination (Marand et al., 2017) in the pericentromeric regions of potato genomes, leading to reduced polymorphisms, and hence marker density as well as enrichment of fixed deleterious alleles (Zhang et al., 2019), and tight genetic linkage in these genomic areas.

3.2 Phenotypes and trait heritability estimates

Phenotypes for yield, dry matter content, chipping quality, length/width ratio, and senescence (a proxy of earliness) were assessed for the 755 individuals retained in the MASPOT panel after full SNP data filtering (Figure 2, Supplementary File 6). The phenotypes were corrected for yearly effects between seasons using a linear mixed model only, as our data did not allow the rigorous estimation of G × E effects (Sverrisdóttir et al., 2017). Chipping quality phenotypes were missing for 31% of the MASPOT panel. Following correction, the phenotypic data appeared approximately normally distributed for all traits, except for length/width ratio, following a quotient distribution, and chipping quality (Supplementary Figure 4), presenting with a right skew. This was most pronounced for length/width ratio and was likely a result of this being a ratio distribution of two normally distributed variables (length and width, respectively) (Díaz-Francés and Rubio, 2013). Regardless, all phenotypes were used for GS and GWAS models without transformation, and the significance threshold indicator lines used in Manhattan plots were based on Bonferroni correction for all phenotypes.

TABLE 1 Number of single-nucleotide polymorphisms (SNPs) and individuals in each data set.

SNP set	SNPs after first filtering	SNPs after second filtering	Individuals after filtering
	MAF > 1%, coverage > 5, missing rate<50%	5< coverage <60, missing rate<70%	MASPOT panel
Full	182,757	171,859	755
Non-synonymous	32,352	29,553	755
Synonymous	34,695	31,229	755
Non-coding	33,743	32,388	751
Combination (non-synonymous, synonymous, non-coding)	100,790	93,170	755

Total number of individuals before filtering: 762 in MASPOT panel.

Pedigree heritabilities (Supplementary Figure 5) were generally estimated as higher compared to genomic narrow sense heritability (Table 2). Yield exhibited the lowest heritability, with genomic and pedigree heritabilities of 22% and 30%, respectively, while dry matter content had the highest pedigree heritability of 91%, but only 41% genomic heritability. The difference in genomic narrow sense heritabilities may be a result of insufficient sampling of the true genomic diversity in the 93,170 marker combination set to accurately estimate relatedness as suggested in Sverrisdóttir et al. (2017). However, the specific large difference for dry matter content cannot be explained by such a general effect. We speculated that this may be related to the fact that the cytoplasmic type of potato. cytoplasmic type W/γ, is positively correlated with starch content (Sanetomo and Gebhardt, 2015) and therefore dry matter content. The cytoplasmic type is captured in the pedigree heritability estimates, but since genomic markers are derived from nuclear DNA, this is likely not captured in the genomic heritability estimate. The means and median dry matter contents of offspring of W/γ-type cytoplasm mother (male-sterile) were compared with the remaining panel. Indeed they were significantly different ($p = 2.2*10^{-16}$), indicating that the cytoplasmic markers related to male sterility, not captured in the genomic markers, could constitute some/all of the unaccounted genetic diversity underpinning the trait heritability. However, removing W/γ individuals and parents from the analysis, the pedigree heritability was 95% and the genomic estimates 40%, in essence the same values as obtained with the entire panel.

3.3 Population structure characterization

Figure 3 shows the first three principal components from the PCA of the genomic relationship matrix of the 93,170 marker annotated combination data set. The offspring of the full diallel cross showed no clear separation into distinct genetic groupings based on PC1 and PC2, though same-father siblings generally congregated closer within the span of the full panel. This trend was less pronounced for same-mother siblings. Plotting PC1 against PC3 showed some offspring diverging genetically from the main group. These individuals were generally progeny of 93-CAQ-14 (father), Agria (mother), or 96-BYM-8 (mother) [Agria grandmother], in addition to a few clones with miscellaneous parentage. Beyond 96-BYM-8 being a descendent of Agria, their pedigrees do not allude to any distinct features of these MASPOT

10.3389/fpls.2024.1340189

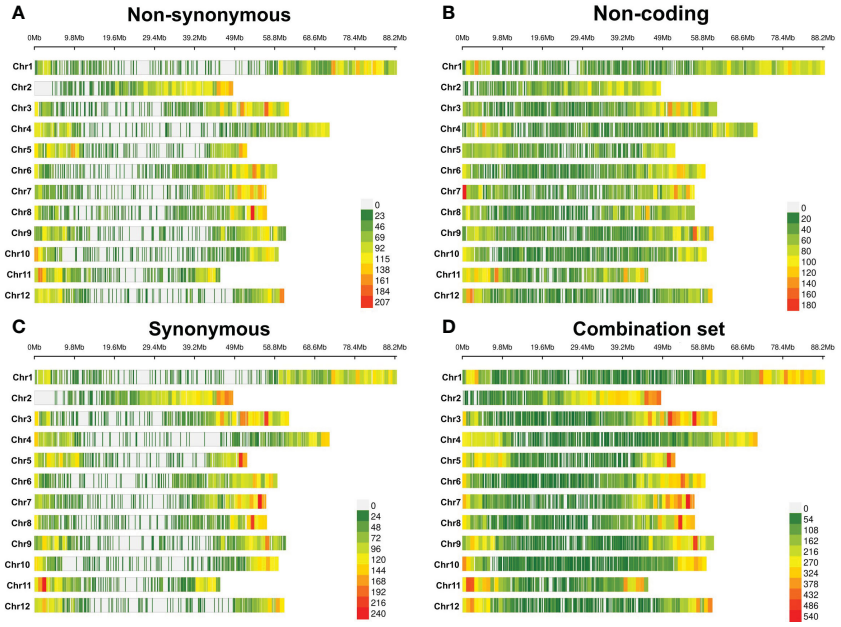

FIGURE 1
Heat map of marker density in 1-Mb windows for each chromosome. **(A)** Non-synonymous variants, **(B)** non-coding variants, **(C)** synonymous variants, and **(D)** combination set of annotated variants. The color gradient denotes marker count.

parents' heritages (data not shown). The population structure is corrected for in both GS and GWAS through the **G**-matrix.

3.4 Genomic prediction

3.4.1 Effect of marker filtering and density

Genomic predictions with marker reduction were conducted based on the four annotated subsets, non-synonymous, synonymous, non-coding, and the combination set. The correlations between observed phenotypes and GEBVs calculated for each individual using GBLUP models with eightfold cross-validation and 10 repeats are shown in Figure 4A. The highest correlation was obtained for dry matter content of 0.75 using the ncSNPs (Table 3). However, performance was not significantly different from sSNPs and the combined SNP set. Only the nsSNPs produced a significantly poorer prediction accuracy than the other three sets ($p = 7.9*10^{-5}$ vs. sSNP, $5.0*10^{-6}$ vs. ncSNPs, and $2.0*10^{-5}$ vs. all SNPs), but still yielding a mean correlation coefficient of 0.74. In all dry matter content cases, the prediction accuracies plateaued at around 1,000 markers. Chipping quality and length/width ratio could be modeled to intermediate correlations of maximum 0.54 and 0.50, respectively. For chipping quality, filtering the markers to nsSNPs and sSNPs produced significantly lowered prediction accuracies compared to using ncSNPs alone ($p = 6.2*10^{-4}$ and $4.1*10^{-5}$, respectively) or a combination of all SNPs for over 25,000 markers ($p = 6.3*10^{-3}$ and $3.5*10^{-4}$), while the correlation coefficients plateaued at ~10,000 markers. The numerical difference in performance was, however, only slight. For length/width ratio, only sSNPs performed significantly better than nsSNPs (0.50 compared to 0.48, $p = 3.3*10^{-3}$), while no statistically significant difference could be observed between the remaining data sets for over 25,000 markers. Traits senescence and yield had the lowest correlation coefficients of maximum 0.35 and 0.28, respectively, and the correlation coefficients did not reach a plateau for other than the combination set when using ~50,000

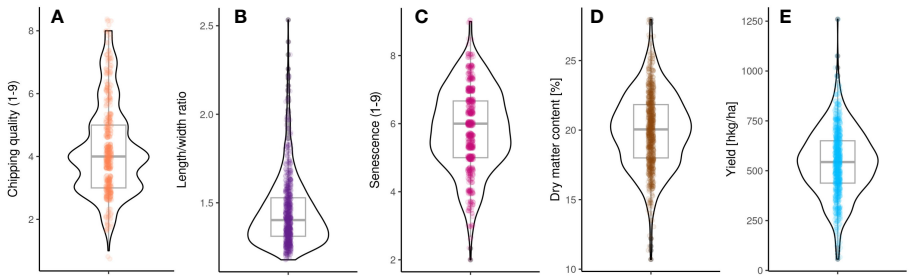

FIGURE 2
Distribution of phenotypes in the MASPOT panel. **(A)** Chipping quality, **(B)** length/width ratio, **(C)** senescence, **(D)** dry matter content, and **(E)** yield.

TABLE 2 Mean, range, phenotypic and genetic variance, coefficient of variance (CV), genomic narrow sense heritability, and pedigree narrow sense heritability.

Phenotype	Mean	Range	Phenotypic variance	Genetic variance	Phenotypic CV (%)	Genetic CV (%)	h_p^2 (%)	h_g^2 (%)
Chipping quality	4.24	1–8	2.02	1.25	98.44	77.34	74.02	61.73
Length/ width ratio	1.45	1.18–2.53	0.04	0.02	3.33	2.25	59.01	45.56
Senescence	5.83	2–9	1.41	0.53	20.33	12.47	60.55	37.61
Dry matter content	19.98	10.7–27.95	6.89	2.80	13.14	8.37	91.17	40.55
Yield	540.81	56.73–1,259.68	26,528.26	5,710.41	30.12	13.97	29.71	21.53

markers. For senescence, there was not a statistically significant best-performing annotation set—only the ncSNP models gave a statistically significant worse prediction accuracy >25,000 SNPs, with a mean correlation coefficient of 0.32 ($p = 4.9*10^{-6}$ vs. nsSNPs, $5.1*10^{-5}$ vs. sSNPs, and $7.3*10^{-7}$ vs. all SNPs). For modeling yield, the nsSNPs produced significantly best results, improving correlation coefficients by 0.03–0.04 compared to the other SNP types when using >25,000 markers ($p = 2.8*10^{-6}$ vs. sSNP, $9.1*10^{-7}$ vs. ncSNPs, and $9.0*10^{-5}$ vs. all SNPs).

In general, reducing the number of markers used to model GEBVs disintegrated the model performance at low marker counts, with the highest obtained correlation coefficient of the trait being proportional to the number of reductions tolerated before model collapse. Model collapse was indicated by reduced mean correlation coefficients and substantial widening of variance. For traits with robust, high prediction accuracy, like dry matter content, as few as 1,000 markers were able to produce correlation coefficients close to the best obtained performance, while 10,000 markers are required to model chipping quality and length/width ratio to optimum prediction accuracies, closest approximating the trait pedigree narrow sense heritabilities. For senescence and yield, with generally low prediction accuracy

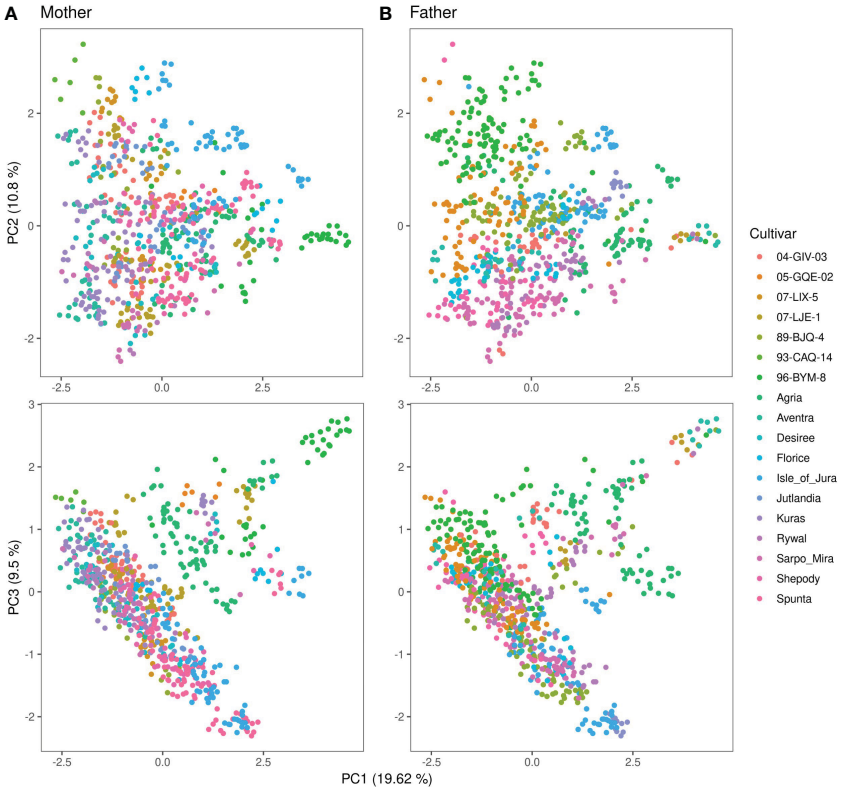

FIGURE 3
Principal component analysis of the genomic relationship matrix of the annotated combination set of 93,170 SNP markers for 755 MASPOT panel clones colored by **(A)** mother and **(B)** father. Principal component 1 is plotted against principal component 2 in the top and against principal component 3 in the bottom. The principal components explain 19.62%, 10.8%, and 9.5% of the total explained variance, respectively.

10.3389/fpls.2024.1340189

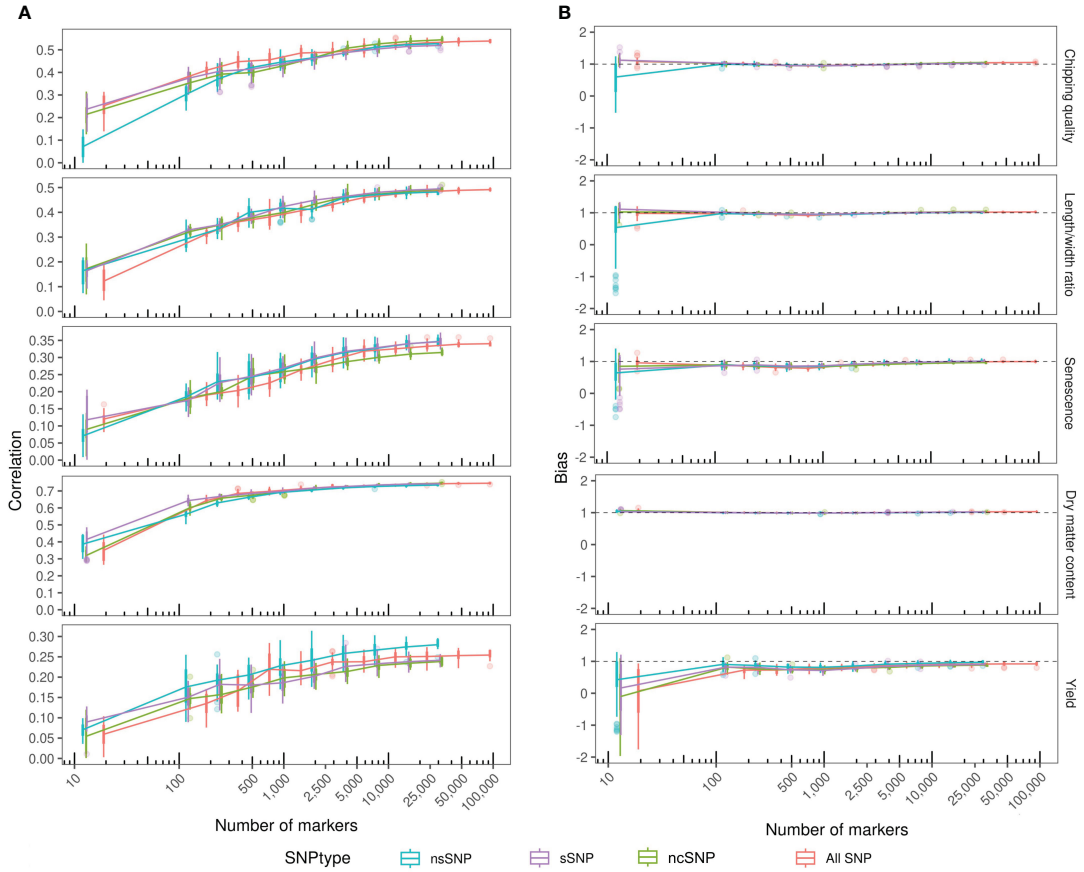

FIGURE 4
GBLUP prediction correlation coefficient and bias within the MASPOT panel over a number of annotated markers. The markers were reduced from full annotated SNP data sets by iterative reduction to every other marker (every 10th for the final reduction) to avoid the introduction of positional bias. Each reduction was repeated up to four times for each iteration. Color by marker annotation (SNPtype). **(A)** Prediction correlation coefficient between observed and predicted phenotypic values. Boxplots of correlation coefficients determined for each marker count with connecting lines through the mean. **(B)** Bias of GEBVs estimated as the slope of the linear regression line between observed and predicted phenotypic values. Boxplots of biases determined for each marker count with connecting lines through the mean bias.

models, increasing the number of markers will improve model performance.

Filtering the markers according to annotation did not notably improve or worsen model performance for any of the analyzed traits, in either top performance (with a high number of markers)—except for yield—or model collapse during marker reduction.

The ability to predict the amplitude of the phenotypic variance, prediction bias, was evaluated for each model as the slope (β) of the regression line between the predicted (x) and observed (y) phenotypic values (Figure 4B) and was quite robust. For all traits and across all annotation subsets, the biases approximate 1 regardless of the number of markers included, except for<100

TABLE 3 Mean prediction correlation ± standard deviation observed for each trait across all annotation datasets using ~30,000 markers for modeling.

Trait	Non-synonymous SNPs (29,553)	Synonymous SNPs (31,229)	Non-coding SNPs (32,388)	Combination (46,585)
Chipping quality	0.527 ± 0.010 (1.027)	0.519 ± 0.010 (1.024)	**0.544 ± 0.011 (1.052)**	0.536 ± 0.011 (1.046)
Length/ width ratio	0.482 ± 0.010 (1.012)	**0.495 ± 0.005 (1.025)**	0.493 ± 0.009 (1.042)	0.489 ± 0.010 (1.022)
Senescence	**0.347 ± 0.012 (1.011)**	0.346 ± 0.015 (1.006)	0.315 ± 0.008 (0.967)	0.339 ± 0.008 (0.997)
Dry matter content	0.735 ± 0.004 (1.016)	0.743 ± 0.002 (1.023)	**0.746 ± 0.003 (1.028)**	0.744 ± 0.003 (1.029)
Yield	**0.280 ± 0.014 (0.970)**	0.242 ± 0.011 (0.901)	0.239 ± 0.010 (0.885)	0.253 ± 0.013 (0.920)

Highest obtained correlation for each trait in bold. Bias in brackets.

markers. However, a slight dip in bias below 1 and subsequent re-stabilization at ~1 were observed for intermediate marker counts for both senescence and yield—increasing axis resolution reveals that the same trend can be observed for all traits (not shown), indicating that the GEBVs became less biased with increasing model fitting using ~10,000 markers. Using as many markers as possible produced the most reliable prediction accuracies, while a reduced set can still yield models of similar performance for some traits.

3.5 Genome-wide association studies

Single-marker regression GWAS was conducted chromosome-wise for the full combination set of 93,170 SNPs for each trait (Supplementary File 7) as well as for non-synonymous, synonymous, non-coding, and a 1-in-3 reduction of the combination set (Supplementary Figures 6-10, Supplementary File 8). Population structure was accounted for by including **G**-matrices based on all chromosomes, except the one encoding the marker being tested for association. The Q–Q plots of the observed *versus* expected -$\log_{10}$(*p*-value) for each analysis showed some inflation of the *p*-values from the expected, assuming no association (Figure 5A) for most traits, which was why the genomic inflation factors (λ_{gc}), ranging from 1.03 to 1.17, were used to correct the *p*-values for traits where λ_{gc}>1 (Figure 5B). Deviation in the tail after correction indicated that significant marker effects were found.

The Manhattan plots of the corrected -$\log_{10}$(*p*-values) are shown in Figure 6 for the full combination set. For chipping quality, a single significant marker was identified on chromosome X (*p*-value = $3.6*10^{-7}$), explaining 3.9% of the total phenotypic variance.

The length/width ratio phenotypes follow ratio distribution, but regardless we have used Bonferroni correction for correction for multiple testing, and the weak QTL peaks for the trait might be false positive associations from lenient correction. Two regions with significant SNPs were found for length/width ratio—a single SNP on chromosome II (*p*-value = $1.1*10^{-7}$, Chr2;33033331, 3.7% explained variance), while significant associations were detected almost chromosome-wide for chromosome X (142 in total), but most densely around 48 Mb, with a distinct peak of *p*-value = $1.9*10^{-30}$, explaining 16% of the phenotypic variance. A region of three same-gene SNPs in the unmapped pseudomolecules (chromosome 0) (*p*-value = $2.3*10^{-9}$, PGSC0003DMG400026855) was also found but has been mapped to chromosome X (48.6 Mb) in the DMv6.1 reference genome (Pham et al., 2020) and is hence part of the 48-Mb major QTL.

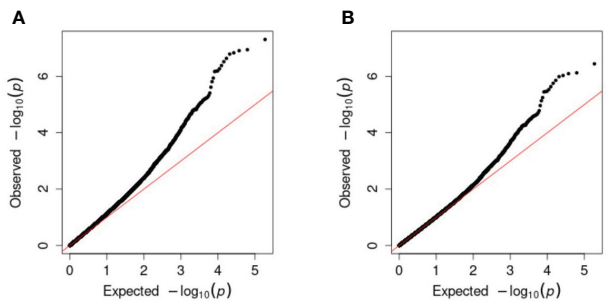

FIGURE 5
Q–Q plots of expected -$\log_{10}$(*p*-values), assuming no marker–trait association, against observed (chipping quality and the full combination set in this example). **(A)** Before correction with the genomic inflation factor and **(B)** after correction.

Dry matter content also displayed a region of significant SNPs on chromosome X between 48 and 58 Mb (peak *p*-value = $3.7*10^{-10}$, explaining 5.7% phenotypic variance).

For senescence, two regions of significant SNPs were identified: two same-gene SNPs on chromosome II (*p*-value = $1.4*10^{-10}$, PGSC0003DMG400012642, 5.5% explained phenotypic variance) and a major peak on chromosome V from 10 to 60 Mb with peak *p*-value = $3.9*10^{-25}$ that explained 13.7% phenotypic variance. No significant associations were found for yield.

GWAS was also performed for four of the marker sets, namely, the non-synonymous, synonymous, non-coding, and 1-in-3 reduced combination SNP sets, each being ~30,000 SNPs. Generally, all data sets yielded similar results; however, a positional shift of the major QTL could be observed for all traits using the ncSNP sets (with the lowest SNP density in genic regions and highest centromeric SNP density) either from a distal arm position to the pericentromeric region of that chromosome (e.g., for senescence) or to a distal position in another chromosome (e.g., for length/width ratio and dry matter content). In addition, associations on different chromosomes outside the signals could be observed for all traits and were found to differ in position across datasets.

The observed QTLs are in close agreement with those previously observed for the traits analyzed. Chipping quality is dependent on the genetically controlled mechanism of cold-induced sweetening (CIS) that is caused by reducing sugar accumulation in the cold storage of tubers (Fischer et al., 2013; Xiao et al., 2018). Maillard reaction during frying affects the quality of chips and French fries (D'hoop et al., 2014). The reducing sugars glucose and fructose are dissimilated from amyloplastic storage starch granules to serve as osmo- and cryoprotectants as part of the tuber starch metabolism during CIS (Schreiber et al., 2014; Van Harsselaar et al., 2017), which explains why chipping quality and starch content are correlated traits. Studies have identified loci on all chromosomes associated with frying color, and indeed several of the genes affecting the trait are related to starch and sucrose metabolism (Li et al., 2008; Werij et al., 2012; Fischer et al., 2013; D'hoop et al., 2014; Xiao et al., 2018; Byrne et al., 2020). Using nsSNPs, a SNP on chromosome III at 38.4 Mb was associated with chipping quality, located proximal to the *Pain-1* invertase (PGSC0003DMG400013856, 39255023-9538), involved in enzymatic sucrose conversion into reducing sugars (Draffehn et al., 2012), which has previously been associated with chip quality and tuber starch content (Li et al., 2008; Draffehn et al., 2010; Schreiber et al., 2014). Chipping quality-associated SNPs on chromosome X at 55.4 Mb (combination set) and 58–59 Mb (nsSNP, sSNP) were located within the starch metabolism gene-rich region at 50–60 Mb (Sharma et al., 2013). Previous studies have also found fry color and starch content associations in this region of cell wall invertases (e.g., in the *Inv-ap-a* locus of *InvCD111* and *InvCD141* at 55.8 MB), invertase inhibitors (e.g., *InvInh-10/4*), a sucrose phosphatase, a

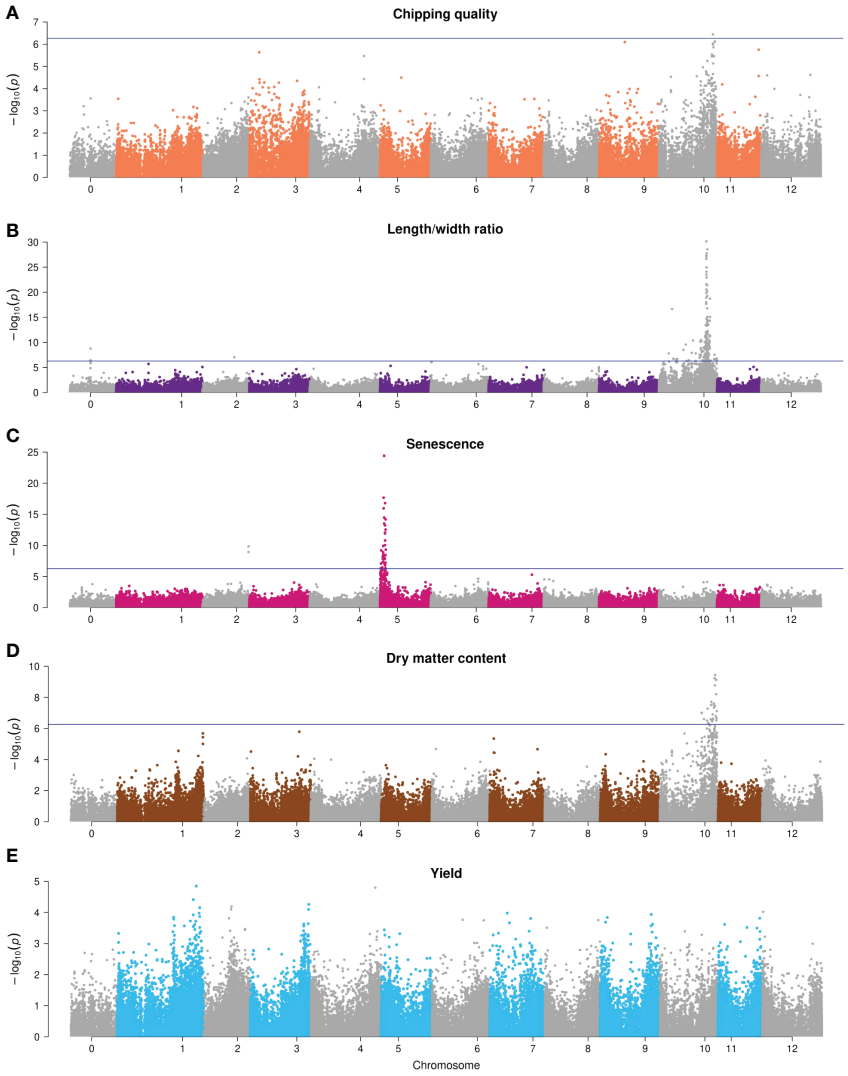

FIGURE 6
Manhattan plots of -$\log_{10}$(p-value) GWAS results of the full 93,170 marker annotated combination set. **(A)** Chipping quality, **(B)** length/width ratio, **(C)** senescence, **(D)** dry matter content, and **(E)** yield. Chromosome 0 is pseudomolecules. The significance threshold with Bonferroni correction is indicated with a horizontal line.

fructose-1,6-biphosphatase, a fructose-biphosphate aldolase, and patatins (the main tuber storage proteins) (Li et al., 2008; Draffehn et al., 2010; Schreiber et al., 2014; Byrne et al., 2020). We found numerous associations for dry matter content in this region, consistent with the traits being correlated through the shared metabolic pathway. Additionally, single-SNP chipping quality associations on chromosome XI at 42.7 Mb (1-in-3 combination) and chromosome XII at 1.5 Mb (nsSNP) were proximal to genes functional in plant starch interconversion (Schreiber et al., 2014), namely, invertase *Inv-n-11/3* (PGSC0003DMG400026530, 39907597-13829) and *AGPaseB-12* (PGSC0003DMG400046891, 1226599-30218), respectively. A selection of single SNP associations for dry matter content could also be found in using the different annotation data sets. On chromosome III, a dry matter content association was found (50.3 Mb, 1-in-3 combination) downstream of the *SEX4* phosphoglucan phosphatase (PGSC0003DMG400015246, 50875724-85587)—a region at 50.8 Mb where Wilson et al. (2021) have also found associations to dry matter content. The nsSNPs and ncSNPs gave significant associations on the north arm of chromosome XI (0.5–4 Mb and 7.1 Mb). For the ncSNPs alone, no association was seen on chromosome X, indicating missing linkage on that chromosome when using only ncSNPs. The upstream region of chromosome XI is scattered with starch and sucrose conversion genes, e.g., *UGPase-11* (PGSC0003DMG401013333, 808268-14810), sucrose transporter *Sut1* (PGSC3000DMG400009213, 9052433-7333), invertase inhibitors *InvInh-11* (PGSC0003DMG400038811, 7433845-4381), invertases *INV-11/1* (PGSC0003DMG400019494, 5046813-52945), and debranching enzyme *DBE-11* (start 3945240,

not annotated). *UGPase-11* was screened by Schreiber et al. (2014), who did not find an association in a 208-genotype tetraploid population. While not reproducible in more than two of the annotated data sets, the associations on chromosome XI do coincide with genes involved in trait-related processes.

Length/width ratio has a major QTL peak at ~48 Mb on chromosome X. The major QTL controlling overall tuber shape was identified as the R_0 locus (Van Eck et al., 1994), and trait effect has been mapped to this region in previous studies (Endelman and Jansky, 2016; Manrique-Carpintero et al., 2018; Zia et al., 2020). Most recently, the R_0 locus was fine-mapped to a 200-kb region from 49.5 to 49.7 Mb in the DMv6.1 reference genome (Pham et al., 2020) spanning 18 candidate genes. RNA sequencing indicated that five genes had differential gene expression, including genes of a lipid transfer protein and a HSI2-like protein, both with roles in plant growth hormone response, regulating plant growth and development (Fan et al., 2022). It is uncertain whether the remaining associations for length/width ratio on chromosome X are spurious hits because of linkage to the R_0 locus or true associations without additional fine-mapping. Using the ncSNPs alone reveals a major QTL on chromosome XI from 0.1-4.1 Mb and 9.6 Mb rather than on X. Tuber shape QTLs have, however, previously been identified on the north arm of chromosome XI (D'hoop et al., 2014; Manrique-Carpintero et al., 2018), which is why it cannot be refuted as a true positive data-dependent association. Additional tuber shape associations were found using the four SNP sets on chromosomes I, II, VI, and VII. While D'hoop et al. (2014) found tuber shape associations in the upstream region of chromosome II, the downstream end of chromosome VI (Manrique-Carpintero et al., 2018), and Park et al. (2021) identified a QTL on chromosome VII, the associations found in chromosome I (sSNP) appear previously undetected.

Senescence is a major QTL effect trait. Deletion alleles in the *StCDF1* gene (PGSC3000DMG400018408, 4538880-41736) in the north end of chromosome V result in early maturing plants. The gene encodes a transcription factor, mediating between the circadian clock and the tuberization signal (Kloosterman et al., 2013). We identified this major QTL, with a total of four within-gene SNPs having a significant association in the four analyses. Additional associations were also found on chromosomes II, III, and VII using different SNP sets. A maturity trait association has previously been found on chromosome II (D'hoop et al., 2014), but not on III or VII. Interestingly, the associations found between 2.8 and 2.9 Mb on chromosome III (nsSNPs) are located upstream of *auxin response factor 8-1* (PGSC3000DMG401018664, 2955745-64135) and two auxin hydrogen symporters (PGSC3000DMG400018678, 2976443-2982482 and PGSC3000DMG400018668, 3084832-91586). In addition to short-day photoperiods, tuberization and plant development are under hormonal control, and high yield is correlated with late plant maturity (van Eck, 2007). Auxin affects growth, the rate of tuberization, and cell differentiation at all stages of life, but plant response is genotype dependent (Kolachevskaya et al., 2019). A confirmation of these candidate genes as senescence QTLs requires further analysis.

4 Discussion

4.1 Marker type and density consequences for genomic prediction

We found that the performance of GS was generally insensitive to marker types. It was somewhat surprising that the ncSNPs expected to have the least importance for phenotypic expression were equally capable of prediction performance as nsSNPs, which have a more direct relationship with phenotypic performance, given that each nsSNP directly causes an amino acid change in the gene product. However, other studies on the impact on genomic prediction of using rare and low-frequency variants in dairy cattle (Zhang et al., 2018) and functionally prioritized variants in maize (Ramstein and Buckler, 2022) also reported a lack of model improvement. This indicates that there is sufficient marker density to adequately describe the relevant genomic variation independently of how markers are filtered. In fact, the distance of the marker to causal polymorphisms, as well as the density of markers needed to characterize the population, is related to the LD span in the genome (Abera Desta and Ortiz, 2014) rather than to the much higher number of individual polymorphisms. Vos et al. (2017) estimated the LD block size to be 0.6–1.5 Mb and even up to 2.5 Mb in introgressed regions in an analysis of 537 tetraploid cultivars. Coarsely extrapolating these results to this study and assuming an average LD block size of 1 Mb in the MASPOT population, this would suggest that SNPs within a window of 500 kb on either side of a causal polymorphism are reliable as markers. Such a large window size includes multiple markers of any of the types analyzed and suggests that performance difference between marker types can only be expected when comparing more distantly related genotypes, where the LD block size is smaller. In summary, the linkage between the causal alleles and neutral variants is likely sufficiently strong to render the effect of concentrating the causal variants mute, with the causal variant effects becoming diluted in their linkage patterns.

It was also found that the minimum number of markers required to satiate the model was dependent on the trait considered. Traits with lower maximum prediction accuracies, e.g., yield and senescence, still showed improvement in prediction performance when using SNP densities higher than 10k markers as compared to traits of intermediate-high heritability, where training on 10k markers amply reached model optimum performance. Indeed for the high-heritability trait dry matter content only ~1k SNPs were needed to border the plateau optimum. Yield also represents an exception to the rule that marker type is unimportant. Using nsSNPs compared to the full combination set leads to a 12% improvement of model performance at 30k marker density and 16% compared to the worst-performing ncSNPs. Furthermore, the nsSNP model for yield, in contrast to all other traits, was consistently more resilient to marker reduction than models based on other marker types. We speculate that the reason for this difference in behavior is that yield is genetically the most complex of the traits and that the developed models are, in fact, only

metastable, the nsSNP model being the most stable and thus most resilient to marker reduction. An alternative strategy to filtrating markers by type alone could be identification of candidate causal variants (Zhang et al., 2018) by, e.g., bioinformatics analysis (Lee et al., 2020; Høie et al., 2022) based on homology and/or predicted structural information to drive filtration and introduce weighting according to predicted functional impact.

4.2 Dry matter content heritability estimate discrepancy

Notably, we observed a substantial difference between genomic and pedigree heritability estimates for dry matter content, a trait known to be highly heritable (Ortiz et al., 2023). This indicates that some genetic variance is not captured in the **G**-matrix-based model employed in GS. This might stem from cryptic data structures resulting from phenotypic grouping, e.g., according to starch content performance, during breeding and subsequent in-group crossing, generating a systematic SNP × group interaction that is not captured in the heritability estimation, where SNP effects on phenotype are differential across groups, and the average SNP effect across the panel does not capture the variance. However, the specific reason for the observed genomic heritability deflation remains obscure. As a consequence, we only evaluate the performance on genomic prediction models relative to the pedigree narrow sense heritabilities.

4.3 Pedigree heritability *versus* genomic prediction

Correlations of 0.54 and 0.75 were reproduced for chipping quality and dry matter content, respectively, from those reported for GBLUP models trained on the MASPOT panel in Sverrisdóttir et al. (2017, 2018). These results are also consistent with those found in other studies using different statistical models (Stich and Van Inghelandt, 2018; Byrne et al., 2020; Wilson et al., 2021; Ortiz et al., 2022; Pandey et al., 2023). However, from the pedigree narrow-sense heritabilities, higher correlations of up to relations of 0.86 and 0.95 could be observed for chipping quality and dry matter content, respectively. Similarly, the highest obtained correlation for length/width ratio of 0.49 was somewhat lower than the theoretical maximum attainable of 0.77 based on pedigree narrow-sense heritability. A rather poor model performance with correlation of 0.35 was also obtained for senescence using GBLUP, as could be expected for a single-gene trait, despite an estimated pedigree narrow-sense heritability of 61% for the trait, i.e., a theoretical maximum correlation of 0.78. In summary, this indicates that some additional additive variations are still not captured by the prediction models. For the single, large effect QTL traits length/width ratio (Fan et al., 2022) and senescence (Kloosterman et al., 2013), including the major QTL associations as fixed effects in the models, might improve prediction (Kim et al., 2022).

At best, yield correlation coefficients were low or modest, falling between 0.24 and 0.28 using GBLUP. The narrow-sense pedigree heritability indicates that 29% of yield phenotypic variance can be explained by additive genetic effects, corresponding to a maximum correlation of 0.55. The relatively poor model performance is likely attributable to the putative metastability of the model. Increasing the number of individuals in the training population is likely needed to improve this in the future. This will presumably reduce the linkage block size and hence increase the resolution of marker–phenotype relationship. Reduced linkage block sizes will, in turn, change the minimum marker density requirements to create stable models compared to those found here.

4.4 Marker type importance in GWAS

In contrast to genomic prediction, marker type was important for GWAS analysis. We observed both intra- and inter-chromosomal displacements of major and minor QTL signals when using ncSNPs compared to other marker types. For both dry matter content and length/width ratio, the major QTL signal is transferred from chromosome X to XI. However, the signal amplitude is not diminished, which suggests that a portion of the non-coding part of chromosome X is erroneously mapped to chromosome XI, still allowing full capture of the association signal but translocating it. We speculate that the risk of such errors in mapping to the genome reference model is more likely to happen for non-coding regions than for coding regions due to an enrichment of repetitive elements in intergenic regions (Mehra et al., 2015). Similarly, the inter-chromosomal shift of the chromosome V senescence signal of senescence using ncSNPs could also be a result of mis-mapped sequencing reads. Minor QTL shifts were observed across all marker types, also indicating that the finer resolution of GWAS is indeed marker type dependent. Whether the minor QTL hits produced with different SNP types are true associations or spurious hits, however, is pending verification.

4.5 Consequences for breeding

The MASPOT panel clones used to train GBLUP genomic prediction models in this study were not selected for any agronomical performance traits following the full-diallel cross from 18 parents of elite cultivars and breeding clones, representing a diverse selection of alleles from the gene pool and contributing to a broad phenotype range for training. For application in breeding schemes, it is attractive to generate models with high reliability in a broad phenotype range to facilitate both selection and deselection of progeny during breeding selection cycles (Sverrisdóttir et al., 2018). The unbiased correlations obtained for all traits indicate that the range of predicted phenotypic variation is not inflated or deflated compared to the observed, the high-accuracy genomic prediction models being suitable in evaluating both high- and low-performing clones. Using a panel of unselected clones also reduces the potential

load of fixed trait-associated alleles in the population, as can be seen in populations of elite breeding material (Kristensen et al., 2018). This can be expected to improve the detection of genome-wide trait associations. However, using a diallel cross as genotype panel also introduces family structure, as is seen in the family-clustered heat map of the genomic relationship matrix (Supplementary Figure 3), where clusters of strong genetic relationship between full and half sibs are found along the diagonal. However, the panel is highly heterogenous across families (Sverrisdóttir et al., 2017), and while some sibling-based subgroupings are seen in the PCA plot of the genomic relationship matrix (Figure 3), most overlap to form a cohesive group, where the genetic diversity between alleles dominates origin. Regardless, working with populations with family structure, which is almost always the case during potato breeding, necessitates extra caution when evaluating results from association analyses to control false positives. In this study, we have used adjustment by genomic inflation factor in combination with the genomic relationship matrix to successfully obtain robust identification of QTLs.

Based on the evaluated effect on GBLUP prediction model performance imposed by marker density and marker type, respectively, we have found that using 1–10k markers, depending on trait (and possibly population size), distributed evenly across the potato genome, but with no particular demand to location relative to genomic features, is sufficient for the prediction of single traits in tetraploid cultivars. Increasing marker density beyond this level did not notably improve performance gains, particularly for high-heritability traits. This conveniently converges with the marker densities of available SNP arrays like the 20-k SolSTW array (Vos et al., 2015) or the commercial Infinium 12K V2 Potato Array. SNP arrays can generate highly robust genotypes compared to low-depth GBS (Gentzbittel et al., 2019), with ample marker density for high-performance prediction modeling of even low heritability traits, even though for some very complex traits such as yield, large training populations are likely necessary.

In this study, we used a single-trait standard additive GBLUP model for evaluating the impact of marker type and marker density on prediction performance since using a simple model on the individual traits facilitated the interpretation of these effects. However, in breeding programs, clones are selected on multiple traits concurrently, and in such a case, multi-trait models are more suitable to use for optimizing genomic prediction models to support selection (Ortiz et al., 2023).

5 Conclusions

The aim of this study was to study the effects of marker type and density on genomic predictions and GWAS of key tuber performance traits and to elucidate prediction accuracy dependency on marker density. Overall, it was found that relatively few markers, 1k–10k, were sufficient to support genomic prediction models in tetraploid potato. This is consistent with most high-throughput marker technologies. Marker type was found to be largely unimportant for genomic prediction but could influence QTLs' placement in GWAS, where ncSNPs alone do not perform satisfactorily.

Data availability statement

The data presented in the study are deposited in the Zenodo repository "GBS and phenotype data for MASPOT population, a panel of tetraploid potato clones", accession number https://zenodo.org/doi/10.5281/zenodo.10143576.

Author contributions

TA: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review & editing, Data curation, Validation. ES: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing, Software, Validation. HK: Data curation, Writing – review & editing, Investigation, Methodology. KN: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing, Data curation.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The MASPOT population used in this study was provided from the MASPOT project (2012-2017), funded by The Danish Council for Strategic Research (Research grant # 11-116190).

Conflict of interest

Author HK is currently employed by Arla Foods Denmark.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1340189/full#supplementary-material

References

Abera Desta, Z., and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci.* 19, 592–601. doi: 10.1016/j.tplants.2014.05.006

Ahmad, D., Zhang, Z., Rasheed, H., Xu, X., and Bao, J. (2022). Recent advances in molecular improvement for potato tuber traits. *Int. J. Mol. Sci.* 23. doi: 10.3390/ijms23179982

Ashraf, B. H., Byrne, S., Fé, D., Czaban, A., Asp, T., Pedersen, M. G., et al. (2016). Estimating genomic heritabilities at the level of family-pool samples of perennial ryegrass using genotyping-by-sequencing. *Theor. Appl. Genet.* 129, 45. doi: 10.1007/s00122-015-2607-9

Ashraf, B. H., Jensen, J., Asp, T., and Janss, L. L. (2014). Association studies using family pools of outcrossing crops based on allele-frequency estimates from DNA sequencing. *Theor. Appl. Genet.* 127, 1331. doi: 10.1007/s00122-014-2300-4

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67. doi: 10.18637/jss.v067.i01

Byrne, S., Meade, F., Mesiti, F., Griffin, D., Kennedy, C., and Milbourne, D. (2020). Genome-wide association and genomic prediction for fry color in potato. *Agronomy* 10. doi: 10.3390/agronomy10010090

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* 6, 80–92. doi: 10.4161/fly.19695

Clifford, D., and McCullagh, P. (2006). The regress function. *R News* 6, 6–10.

Clifford, D., and McCullagh, P. (2020). The regress package R package. R package version 1.3-21.

Dean, R. B., and Dixon, W. J. (1951)Simplified Statistics for Small Numbers of Observations For convenience a series of observations will be arranged in ascending order of magnitude and assigned the symbols (Accessed November 3, 2023).

de los Campos, G., Sorensen, D., and Gianola, D. (2015). Genomic heritability: what is it? *PloS Genet.* 11, e1005048. doi: 10.1371/JOURNAL.PGEN.1005048

Devlin, B., Roeder, K., and Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theor. Popul. Biol.* 60, 155–166. doi: 10.1006/tpbi.2001.1542

D'hoop, B. B., Keizer, P. L. C., Paulo, M. J., Visser, R. G. F., van Eeuwijk, F. A., and van Eck, H. J. (2014). Identification of agronomically important QTL in tetraploid potato cultivars using a marker-trait association analysis. *Theor. Appl. Genet.* 127, 731–748. doi: 10.1007/s00122-013-2254-y

Díaz-Francés, E., and Rubio, F. J. (2013). On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables. *Stat. Pap.* 54, 309–323. doi: 10.1007/s00362-012-0429-2

Draffehn, A. M., Durek, P., Nunes-Nesi, A., Stich, B., Fernie, A. R., and Gebhardt, C. (2012). Tapping natural variation at functional level reveals allele specific molecular characteristics of potato invertase Pain-1. *Plant Cell Environ.* 35, 2143–2154. doi: 10.1111/j.1365-3040.2012.02544.x

Draffehn, A. M., Meller, S., Li, L., and Gebhardt, C. (2010). Natural diversity of potato (Solanum tuberosum) invertases. *BMC Plant Biol.* 10, 1–15. doi: 10.1186/1471-2229-10-271

Dwivedi, S. L., Heslop-Harrison, P., Spillane, C., Mckeown, P. C., Edwards, D., Goldman, I., et al. (2023). Evolutionary dynamics and adaptive benefits of deleterious mutations in crop gene pools. *Trends Plant Sci.* 28, 685–697. doi: 10.1016/j.tplants.2023.01.006

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6, e19379. doi: 10.1371/journal.pone.0019379

Endelman, J. B., and Jansky, S. H. (2016). Genetic mapping with an inbred line-derived F2 population in potato. *Theor. Appl. Genet.* 129, 935–943. doi: 10.1007/s00122-016-2673-7

Fan, G., Wang, Q., Xu, J., Chen, N., Zhu, W., Duan, S., et al. (2022). Fine mapping and candidate gene prediction of tuber shape controlling Ro locus based on integrating genetic and transcriptomic analyses in potato. *Int. J. Mol. Sci.* 23. doi: 10.3390/ijms23031470

FAOSTAT (2023) *Food and Agriculture Organization of the United Nations Statistics Division*. Available online at: https://www.fao.org/faostat/en/#data/QCL (Accessed March 1, 2023).

Fischer, M., Schreiber, L., Colby, T., Kuckenberg, M., Tacke, E., Hofferbert, H. R., et al. (2013). Novel candidate genes influencing natural variation in potato tuber cold sweetening identified by comparative proteomics and association mapping. *BMC Plant Biol.* 13, 1–15. doi: 10.1186/1471-2229-13-113

Gentzbittel, L., Belzile, F., M Smulders, M. J., Waugh, R., Darrier, B., Russell, J., et al. (2019). A comparison of mainstream genotyping platforms for the evaluation and use of barley genetic resources. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00544

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194, 573–596. doi: 10.1534/genetics.113.151753

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12. doi: 10.2135/cropsci2008.08.0512

Hickey, J. M., Chiurugwi, T., Mackay, I., and Powell, W. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* 49, 1297–1303. doi: 10.1038/ng.3920

Hinrichs, A. L., Larkin, E. K., and Suarez, B. K. (2009). Population stratification and patterns of linkage disequilibrium. *Genet. Epidemiol.* 33, S88–S92. doi: 10.1002/gepi.20478

Høie, M. H., Cagiada, M., Frederiksen, A. H. B., Stein, A., and Lindorff-Larsen, K. (2022). Predicting and interpreting large scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep.* 38, 110207. doi: 10.1101/2021.06.26.450037

Kim, G. W., Hong, J. P., Lee, H. Y., Kwon, J. K., Kim, D. A., and Kang, B. C. (2022). Genomic selection with fixed-effect markers improves the prediction accuracy for Capsaicinoid contents in Capsicum annuum. *Hortic. Res.* 9. doi: 10.1093/hr/uhac204

Kloosterman, B., Abelenda, J. A., Gomez, M. D. M. C., Oortwijn, M., De Boer, J. M., Kowitwanich, K., et al. (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nat. 2013 4957440* 495, 246–250. doi: 10.1038/nature11912

Kolachevskaya, O. O., Lomin, S. N., Arkhipov, D. V., and Romanov, G. A. (2019). Auxins in potato: molecular aspects and emerging roles in tuber formation and stress resistance. *Plant Cell Rep.* 38, 681–698. doi: 10.1007/s00299-019-02395-0

Kristensen, P. S., Jahoor, A., Andersen, J. R., Cericola, F., Orabi, J., Janss, L. L., et al. (2018). Genome-wide association studies and comparison of models and cross-validation strategies for genomic prediction of quality traits in advanced winter wheat breeding lines. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00069

Lee, Y.-S., Won, K., Shin, D., and Oh, J.-D. (2020). Risk prediction and marker selection in nonsynonymous single nucleotide polymorphisms using whole genome sequencing data. *Anim. Cells Syst.* 24 (6), 321–328. doi: 10.1080/19768354.2020.1860125

Lenaerts, B., Collard, B. C. Y., and Demont, M. (2019). Review: Improving global food security through accelerated plant breeding. *J. Plant Sci.* 287 (110207). doi: 10.1016/j.plantsci.2019.110207

Li, L., Paulo, M.-J., Strahwald, J., Lübeck, J., Hofferbert, H.-R., Tacke, E., et al. (2008). Natural DNA variation at candidate loci is associated with potato chip color, tuber starch content, yield and starch yield. *Theor. Appl. Genet.* 116, 1167–1181. doi: 10.1007/s00122-008-0746-y

Li, J., Wang, Y., Wen, G., Li, G., Li, Z., Zhang, R., et al. (2019). Mapping QTL underlying tuber starch content and plant maturity in tetraploid potato. *Crop J.* 7, 261–272. doi: 10.1016/j.cj.2018.12.003

LiLin-Yin, (2023). *CMplot: Circle Manhattan Plot. R package version 4.4.1.*

Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M., and Meuwissen, T. H. E. (2009). The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. *Genetics* 183, 1119–1126. doi: 10.1534/genetics.109.107391

Manrique-Carpintero, N. C., Coombs, J. J., Pham, G. M., Laimbeer, F. P. E., Braz, G. T., Jiang, J., et al. (2018). Genome reduction in tetraploid potato reveals genetic load, haplotype variation, and loci associated with agronomic traits. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00944

Marand, A. P., Jansky, S. H., Zhao, H., Leisner, C. P., Zhu, X., Zeng, Z., et al. (2017). Meiotic crossovers are associated with open chromatin and enriched with Stowaway transposons in potato. *Genome Biol.* 18, 1–16. doi: 10.1186/s13059-017-1326-8

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297. doi: 10.1101/gr.107524.110

Mehra, M., Gangwar, I., and Shankar, R. (2015). A deluge of complex repeats: the solanum genome. *PloS One* 10, e0133962. doi: 10.1371/journal.pone.0133962

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Muthoni, J., Kabira, J., Shimelis, H., and Melis, R. (2015). Tetrasomic inheritance in cultivated potato and implications in conventional breeding. *Aust. J. Crop Sci.* 9, 185–190.

Naeem, M., Demirel, U., Yousaf, M. F., Caliskan, S., Caliskan, M. E., and Wehling, P. (2021). Overview on domestication, breeding, genetic gain and improvement of tuber quality traits of potato using fast forwarding technique (GWAS): A review. *Plant Breed.* 140, 519–542. doi: 10.1111/pbr.12927

Ortiz, R., Crossa, J., Reslow, F., Perez-Rodriguez, P., and Cuevas, J. (2022). Genome-based genotype × Environment prediction enhances potato (Solanum tuberosum L.) improvement using pseudo-diploid and polysomic tetraploid modeling. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.785196

Ortiz, R., Reslow, F., Montesinos-López, A., Huicho, J., Pérez-Rodríguez, P., Montesinos-López, O. A., et al. (2023). Partial least squares enhance multi-trait genomic prediction of potato cultivars in new environments. *Sci. Rep.* 13, 1–12. doi: 10.1038/s41598-023-37169-y

Pandey, J., Scheuring, D. C., Koym, J. W., Endelman, J. B., and Vales, M. I. (2023). Genomic selection and genome-wide association studies in tetraploid chipping potatoes. *Plant Genome* 16. doi: 10.1002/tpg2.20297

Park, J., Massa, A. N., Douches, D., Coombs, J., Akdemir, D., Yencho, G. C., et al. (2021). Linkage and QTL mapping for tuber shape and specific gravity in a tetraploid mapping population of potato representing the russet market class. *BMC Plant Biol.* 21, 1–18. doi: 10.1186/s12870-021-03265-2

Pérez, P., and de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442

Pham, G. M., Hamilton, J. P., Wood, J. C., Burke, J. T., Zhao, H., Vaillancourt, B., et al. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. *Gigascience* 9, 1–11. doi: 10.1093/gigascience/giaa100

Pham, G. M., Newton, L., Wiegert-Rininger, K., Vaillancourt, B., Douches, D. S., and Buell, C. R. (2017). Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. *Plant J.* 92, 624–637. doi: 10.1111/tpj.13706

Posit team (2023) *RStudio: Integrated Development Environment for R*. Available online at: http://www.posit.co/.

Ramstein, G. P., and Buckler, E. S. (2022). Prediction of evolutionary constraint by genomic annotations improves functional prioritization of genomic variants in maize. *Genome Biol.* 23. doi: 10.1186/s13059-022-02747-2

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rorabacher, D. B. (1991). Statistical treatment for rejection of deviant values: critical value of Dixon's "Q" Parameter and related subrange ratios at the 95 % Confidence level. *Anal. Chem.* 63, 139–146.

Sanetomo, R., and Gebhardt, C. (2015). Cytoplasmic genome types of European potatoes and their effects on complex agronomic traits. *BMC Plant Biol.* 15. doi: 10.1186/s12870-015-0545-y

Schreiber, L., Nader-Nieto, A. C., Schönhals, E. M., Walkemeier, B., and Gebhardt, C. (2014). SNPs in genes functional in starch-sugar interconversion associate with natural variation of tuber starch and sugar content of potato (Solanum tuberosum L.). *G3* 4, 1797–1811. doi: 10.1534/g3.114.012377

Selga, C., Koc, A., Chawade, A., and Ortiz, R. (2021). A bioinformatics pipeline to identify a subset of SNPs for genomics-assisted potato breeding. *Plants* 10, 1–14. doi: 10.3390/PLANTS10010030

Sharma, S. K., Bolser, D., de Boer, J., Sønderkær, M., Amoros, W., Carboni, M. F., et al. (2013). Construction of reference chromosome-scale pseudomolecules for potato: Integrating the potato genome with genetic and physical maps. *G3 Genes Genomes Genet.* 3, 2031–2047. doi: 10.1534/g3.113.007153

Shim, H., Chasman, D. I., Smith, J. D., Mora, S., Ridker, P. M., Nickerson, D. A., et al. (2015). A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. *PloS One* 10, 120758. doi: 10.1371/journal.pone.0120758

Slater, A. T., Cogan, N. O. I., Forster, J. W., Hayes, B. J., and Daetwyler, H. D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. *Plant Genome* 9. doi: 10.3835/plantgenome2016.02.0021

Spooner, D. M., Ghislain, M., Simon, R., Jansky, S. H., and Gavrilenko, T. (2014). Systematics, diversity, genetics, and evolution of wild and cultivated potatoes. *Bot. Rev.* 80, 283–383. doi: 10.1007/s12229-014-9146-y

Spooner, D. M., McLean, K., Ramsay, G., Waugh, R., and Bryan, G. J. (2005). A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. *PNAS* 102, 14694–14699. doi: 10.1073/pnas.0507400102

Stich, B., and Van Inghelandt, D. (2018). Prospects and potential uses of genomic prediction of key performance traits in tetraploid potato. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00159

Sun, H., Jiao, W.-B., Krause, K., Campoy, J. A., Goel, M., Folz-Donahue, K., et al. (2022). Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat. Genet.* 54, 342–348. doi: 10.1038/s41588-022-01015-0

Sverrisdóttir, E., Byrne, S., Nielsen, E. H. R., Johnsen, H. Ø., Kirk, H. G., Asp, T., et al. (2017). Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. *Theor. Appl. Genet.* 130, 2091–2108. doi: 10.1007/s00122-017-2944-y

Sverrisdóttir, E., Nielsen, E. H. R., Johnsen, H. Ø., Kirk, H. G., Asp, T., Janss, L., et al. (2018). The value of expanding the training population to improve genomic selection models in tetraploid potato. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01118

The Potato Genome Sequencing Consortium (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475, 189–195. doi: 10.1038/nature10158

Turner, S. D. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* 3 (25). doi: 10.21105/joss.00731

Van Dijk, M., Morley, T., Rau, M. L., and Saghai, Y. (2021). A meta-analysis of projected global food demand and population at risk of hunger for the period 2010-2050. *Nat. Food* 2, 494–501. doi: 10.1038/s43016-021-00322-9

van Eck, H. J. (2007). "Genetics of morphological and tuber traits," in *Potato Biology and Biotechnology, Advances and Perspectives* (Elsevier, Amsterdam, The Netherlands), 91–115. doi: 10.1016/B978-044451018-1/50048-8

Van Eck, H. J., Jacobs, J. M. E., Stam, P., Ton, J., Stiekema, W. J., and Jacobsen, E. (1994). Multiple alleles for tuber shape in diploid potato detected by qualitative and quantitative genetic analysis using Rflps. *Genetics* 137, 303. doi: 10.1093/GENETICS/137.1.303

Van Harsselaar, J. K., Lorenz, J., Senning, M., Sonnewald, U., and Sonnewald, S. (2017). Genome-wide analysis of starch metabolism genes in potato (Solanum tuberosum L.). *BMC Genomics* 18, 1–18. doi: 10.1186/s12864-016-3381-z

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Vos, P. G., João Paulo, M., Voorrips, R. E., F Visser, R. G., van Eck, H. J., and van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theor. Appl. Genet.* 130, 123–135. doi: 10.1007/s00122-016-2798-8

Vos, P. G., Uitdewilligen, J. G. A. M. L., Voorrips, R. E., Visser, R. G. F., and van Eck, H. J. (2015). Development and analysis of a 20K SNP array for potato (Solanum tuberosum): an insight into the breeding history. *Theor. Appl. Genet.* 128, 2387–2401. doi: 10.1007/s00122-015-2593-y

Werij, J. S., Furrer, H., van Eck, H. J., Visser, R. G. F., and Bachem, C. W. B. (2012). A limited set of starch related genes explain several interrelated traits in potato. *Euphytica* 186, 501–516. doi: 10.1007/s10681-012-0651-y

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (New York: Springer-Verlag). doi: 10.1007/978-3-319-24277-4

Wilson, S., Zheng, C., Maliepaard, C., Mulder, H. A., Visser, R. G. F., van der Burgt, A., et al. (2021). Understanding the effectiveness of genomic prediction in tetraploid potato. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.672417

Xiao, G., Huang, W., Cao, H., Tu, W., Wang, H., Zheng, X., et al. (2018). Genetic loci conferring reducing sugar accumulation and conversion of cold-stored potato tubers revealed by QTL analysis in a diploid population. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00315

Zhang, Q., Sahana, G., Su, G., Guldbrandtsen, B., Lund, M. S., and Calus, M. P. L. (2018). Impact of rare and low-frequency sequence variants on reliability of genomic prediction in dairy cattle. *Genet. Sel. Evol.* 50. doi: 10.1186/s12711-018-0432-8

Zhang, C., Wang, P., Tang, D., Yang, Z., Lu, F., Qi, J., et al. (2019). The genetic basis of inbreeding depression in potato. *Nat. Genet.* 51, 374–378. doi: 10.1038/s41588-018-0319-1

Zhu, M., Cheng, Y., Wu, S., Huang, X., and Qiu, J. (2022). Deleterious mutations are characterized by higher genomic heterozygosity than other genic variants in plant genomes. *Genomics* 114, 110290. doi: 10.1016/j.ygeno.2022.110290

Zia, M. A. B., Demirel, U., Nadeem, M. A., and Çaliskan, M. E. (2020). Genome-wide association study identifies various loci underlying agronomic and morphological traits in diversified potato panel. *Physiol. Mol. Biol. Plants* 26, 1003. doi: 10.1007/s12298-020-00785-3