Targeted genotyping-by-sequencing of potato and software for imputation

Jeffrey B. Endelman*[1], Moctar Kante[2], Hannele Lindqvist-Kreuze[2], Andrzej Kilian[3], Laura M. Shannon[4], Maria V. Caraza-Harter[1], Brieanne Vaillancourt[5], Kathrine Mailloux[5], John P. Hamilton[5], C. Robin Buell[5]

[1]Department of Plant & Agroecosystem Sciences, University of Wisconsin–Madison, Madison, WI, USA.
[2]Genomics, Genetics and Crop Improvement, International Potato Center, Lima, Peru.
[3]Diversity Arrays Technology Pty Ltd, University of Canberra, Bruce, ACT, Australia.
[4]Department of Horticultural Science, University of Minnesota, Saint Paul, MN, USA.
[5]Center for Applied Genetic Technologies, University of Georgia, Athens, GA, USA

*Corresponding author (endelman@wisc.edu)

J.B.E. ORCID 0000-0003-0957-4337
M.K. ORCID 0000-0003-4669-3132
H.L-K. ORCID 0000-0002-6523-8824
L.M.S. ORCID 0000-0003-3935-4909
M.C-H. ORCID 0000-0002-5934-1526
B.V. ORCID 0000-0002-6795-5173
J.P.H. ORCID 0000-0002-8682-5526
C.R.B. ORCID 0000-0002-6727-4677

Core Ideas

1. A mid-density, targeted genotyping-by-sequencing (GBS) assay was developed for potato.
2. The GBS assay includes markers for resistance to potato virus Y, golden cyst nematode, and potato wart.
3. The GBS assay includes multi-allelic markers for potato maturity and tuber shape.
4. The polyBreedR software has functions for manipulating and imputing polyploid marker data in Variant Call Format.
5. Linkage Analysis was more accurate than the Random Forest method when imputing from 2K to 10K markers.

ABSTRACT

Mid-density targeted genotyping-by-sequencing (GBS) combines trait-specific markers with thousands of genomic markers at an attractive price for linkage mapping and genomic selection. A 2.5K targeted GBS assay for potato was developed using the DArTag^{TM} technology and later expanded to 4K targets. Genomic markers were selected from the potato Infinium^{TM} SNP array to maximize genome coverage and polymorphism rates. When sample depth was summarized by marker, the power law $\mu \sim \sigma^{0.8}$ was consistently observed between the mean (μ) and standard deviation (σ). The DArTag and SNP array platforms produced equivalent dendrograms in a test set of 298 tetraploid samples, and 83% of the common markers showed good quantitative agreement, with RMSE (root-mean-squared-error) less than 0.5. DArTag is suited for genomic selection candidates in the clonal evaluation trial, coupled with imputation to a higher density platform for the training population. Our hypothesis that linkage analysis would be highly accurate for imputing tetraploid marker data was confirmed: the RMSE was 0.15 compared to 0.95 by the Random Forest method in a half-diallel population. Regarding high-value traits, the DArTag markers for resistance to potato virus Y, golden cyst nematode, and potato wart appeared to track their targets successfully, as did multi-allelic markers for maturity and tuber shape. In summary, the potato DArTag assay is a transformative and publicly available technology for potato breeding and genetics.

1 INTRODUCTION

Targeted genotyping-by-sequencing (GBS) has become an essential technology for molecular plant breeding. As with restriction site-associated DNA (RAD) sequencing (Baird et al., 2008; Elshire et al., 2011), targeted GBS is based on sequencing a reduced representation of the genome. A key difference is that targeted GBS uses a fixed set of primer pairs or oligonucleotide baits, with the number of targets designed based on the application and price point (Campbell et al., 2015; Gasc et al., 2016; Ali et al., 2016). DArTag is a targeted GBS method based on PCR with molecular inversion probes (Hardenbol et al. 2003) and scalable to thousands of targets (Hardigan et al., 2023; Zhao et al., 2023). As part of the CGIAR Excellence in Breeding platform, DArTag panels were developed for wheat, maize, rice, cowpea, pigeon pea, common bean, groundnut (peanut), sorghum, and potato (Excellence in Breeding, 2022). This article describes the design and validation of the first potato DArTag panel, which had 2.5K targets, as well as a second design project, which extended the assay to 4K targets.

Before DArTag, there was no comparable "mid-density" genotyping service for potato. The main genotyping platform for genetic mapping and genomic selection in potato has been an Infinium^{TM} SNP array, which was originally developed with 8303 markers and then expanded to 12K (Version 2) based on the same discovery panel of 6 varieties (Hamilton et al., 2011; Felcher et al., 2012). The 22K V3 array incorporated new SNPs from a larger discovery panel of 83 tetraploid varieties (Uitdewilligen et al., 2013; Vos et al., 2015), and the 31K V4 array added markers from yet another discovery pool (Sharma and Bryan, 2017). To maintain backwards compatibility with existing marker data, the genomic markers for DArTag were selected from the potato Infinium array.

In addition to genomic markers, the DArTag design includes “essential markers” that are prioritized during the final marker selection and primer design process. For potato, our initial priority was identifying markers with high diagnostic value (i.e., haplotype-specificity) for key resistance genes in a wide variety of genetic backgrounds. When the V1 assay was being designed, in 2020, KASP markers for the Ry_{adg} (Herrera et al., 2018) and Ry_{sto} (Nie et al., 2016) resistance genes against potato virus Y (PVY) were being widely utilized through a “low-density” genotyping service of the Excellence in Breeding platform. These two markers were therefore obvious candidates to include in the V1 DArTag panel. When the V2 DArTag panel was designed in 2023, a number of additional traits were targeted with essential markers.

Our focus for the V1 DArTag assay was tetraploid potato, which is the ploidy level of global commerce. A well-known challenge of GBS in polyploids is the high read depth needed to differentiate heterozygotes with differing allele dosage (Uitdewilligen et al., 2013). The read depth needed to achieve 95% genotyping accuracy in a tetraploid ranges from 30 to 60, depending on the population structure and other assumptions (Gerard et al. 2018; Matias et al. 2019). This limitation has motivated the use of pseudo-diploid (aka diploidized) genotype calls in previous studies (Bastien et al., 2018; Matias et al., 2019), but allele dosage information is needed for the partitioning of genetic variance and breeding value prediction in tetraploids (Endelman et al., 2018; de Bem Oliveira et al., 2019). There is little information or software available to facilitate marker imputation in tetraploids, so filling this gap has been one of our research objectives.

2 MATERIALS AND METHODS

2.1 Genomic markers

SNPs were selected from the 22K V3 SNP array (Felcher et al., 2012; Vos et al., 2015) for the 2.5K V1 DArTag set, and additional SNPs were selected from the V4 31K SNP array for the 4K V2 DArTag set. Physical positions were based on the DMv6.1 reference genome (Pham et al., 2020). Genetic map positions (in cM) were interpolated from the map positions reported in Endelman and Jansky (2016). The interpolated "Marey" map of cM vs. bp was constrained to be monotone nondecreasing (Figure S1) using an I-spline basis with 12 degrees of freedom, generated with R/splines2 (Wang and Yan, 2021). Non-negative basis coefficients were computed by minimizing the mean-squared error with R/CVXR (Fu et al., 2020). The script is available as function *interpolate_cM* in R/MapRtools (Endelman, 2023a). Initially, SNPs were selected based on discretizing the genome into 1 cM bins, and within each bin, SNPs were prioritized based on minor allele frequency (MAF) in a collection of US and CIP germplasm. After saturating the genome, additional SNPs were selected sequentially based on the ad-hoc score $d + 10 \times MAF$, where d is cM distance to the closest selected SNP.

Germplasm for evaluating V1 DArTag came from the International Potato Center (CIP) and University of Wisconsin breeding programs. Data for 703 tetraploid samples are provided as Supplemental File 1 in Variant Call Format (VCF), with the year of submission (2020, 2021, or 2022) for each sample recorded in Supplemental File 2. The function *dart2vcf* in R/polyBreedR (Endelman, 2023b) generates a VCFv4.3 compliant file from the two standard DArTag CSV files ("Allele_Dose_Report" and "Allele_match_counts_collapsed"). R/polyBreedR function *gbs* was used to replace the original DArTag genotype calls (FORMAT field GT) with those based on R/updog, using the "norm" prior (Gerard et al., 2018). Three parameters of the beta-binomial model (SE = sequencing error, AB = allelic bias, OD = overdispersion) were stored for each variant. Both functions utilize R package vcfR (Knaus and Grunwald, 2017).

A single submission of tetraploid (N=323) and diploid (N=52) samples was used to evaluate the 4K V2 DArTag assay (Supplemental Files 3 and 4). Genotype calls were made separately for each ploidy group using the *gbs* function in R/polyBreedR.

A comparison of DArTag vs. SNP array genotypes was conducted using 298 clones for V1 DArTag and 78 clones for V2 DArTag. XY intensity values and genotype calls are provided in Supplemental File 5 for 15,187 markers from the V4 SNP array, based on a normal mixture model estimated with R/fitPoly (Voorrips et al., 2011; Zych et al., 2019). The parameter file for the normal mixture model, which is distributed with R/polyBreedR as "potato_V4array_model.csv", was used when converting Genome Studio Final Reports to VCF with the function *array2vcf*. This function also requires a VCF map definition file to convert from B allele dosage to ALT dosage, which is distributed as "potato_V4array.vcf" with R/polyBreedR. The common markers between DArTag and the SNP array, including matching REF/ALT, were identified using *bcftools isec* (Danecek et al. 2021).

2.2 Imputation

Two methods were compared for the accuracy of imputing SNP array markers from DArTag: Random Forest (RF) and Linkage Analysis (LA). Method RF was implemented as R/polyBreedR function *impute_L2H*, using the R/randomForest package (Liaw and Wiener, 2002). The 100 closest markers were used as prediction variables, and the number of trees was set at 100 by monitoring the out-of-bag error. Method LA was implemented as R/polyBreedR function *impute_LA*, using the software PolyOrigin (Zheng et al., 2021). Imputation error was measured using leave-one-family-out cross-validation in a five-parent half-diallel population (pedigree in Supplemental File 6). The parent codes in Table 1 are P1=W6609-3, P2=W12078-76, P3=W13NYP102-7, P4=W14NYQ4-1, P5=W14NYQ9-2. The high density (10K) phased parental genotypes are in Supplemental File 7.

2.3 Trait markers

A set of six interconnected F1 populations was used to assess the accuracy of genotype calls for the V2 DArTag trait markers Ryadg_chr11_2499502 and H1_chr05_52349069. Parental phasing and haplotype reconstruction utilized PolyOrigin (Zheng et al., 2021), and binary trait locus (BTL) analysis utilized R/diaQTL (Amadeu et al., 2021). The pedigree, genomic marker, and dominant trait marker files needed for diaQTL are Supplemental Files 8, 9, 10, respectively. Validation of trait marker Sli_chr12_2372490 was based on the Sli_898 KASP marker (Clot et al., 2020; Kaiser et al., 2021).

Trait markers CDF1.2_chr05_4488015 and CDF1.4_chr05_4488021 target two different 7 bp insertions of *CDF1* (Kloosterman et al., 2013; Gutaker et al., 2019) and contain equivalent information in the DArT MADC (missing allele discovery count) file. The 81-bp haplotypes in the MADC file were aligned using MUSCLE v3.8 (Edgar, 2004). DArTag read counts for CDF1 alleles 1, 2, and 4 were tabulated with R/polyBreedR function *madc* and validated against genotypes determined via whole-genome sequencing with NovaSeq 2x150 reads (Song and Endelman, 2023).

Genome assemblies of *S. tuberosum* dihaploids were used to validate markers for *OFP20*, a major gene affecting tuber shape (Wu et al. 2018). High molecular weight DNA was extracted from tissue culture plantlets using a CTAB isolation method and Qiagen Genomic tips (Hilden, Germany), followed by an Amicon filter (MilliporeSigma, Burlington, MA) buffer exchange (Vaillancourt et al., 2019) or Takara NucleoBond HMW DNA kit (Takara, Kusatsu, Shiga, Japan). Genome assembly used hifiasm v0.16.1-r375 (Cheng et al. 2021, 2022) with PacBio HiFi Sequel II (Menlo Park, CA) reads from the University of Minnesota Genomics Center. Contigs less than 50kb were discarded using seqkit v2.3.0 (Shen et al. 2016), followed by Ragtag v2.1.0 (Alonge et al. 2019) to scaffold with DM 1-3 516 R44 v6.1 (Pham et al., 2020).

A multiple sequence alignment of 19 *OFP20* haplotypes (Supplemental File 11) was generated using MUSCLE v3.8. Alleles 1–7 and M6_ScOFP20 were reported by van Eck et al. (2022), and the remaining haplotypes come from the dihaploids. The frequency of *OFP20.1* was approximated by ALT frequency at marker OFP20_M6_CDS_994 (994 bp in M6 CDS). For allele *OFP20.8,* which was discovered in the dihaploids (i.e., not in the FASTA file from van Eck et al. (2022)), allele frequency was approximated by REF frequency at marker OFP20_M6_CDS_24; this only works in populations without the M6_ScOFP20 allele. Marker OFP20_M6_CDS_171 was used to report allele depth for allele 2 (ALT) vs. alleles 3 and 7 combined (REF); alleles 1 and 8 were not detected by this marker. Marker OFP20_M6_CDS_75 was supposed to capture an indel at 82 bp that differentiates alleles 3 and 7, but neither haplotype was present in the MADC file (File S4).

3 RESULTS

3.1 Genomic markers

Version 1 (V1) of the potato DArTag GBS assay contained 2501 genomic SNPs, which were selected from the 22K V3 potato SNP array to maximize genome coverage and polymorphism rates (i.e., high minor allele frequency). The number of genomic markers per chromosome ranged from 176 on chr12 to 272 on chr01. The mean distance between adjacent markers was 0.35 cM, with the largest gap of 4.77 cM located on chr11 (Figure S2).

Analysis of 703 tetraploid samples, from three submissions across three years (2020-2022), revealed variability in the amount of sequencing data per sample. In 2020, the total depth (DP sum over markers) was consistent across samples, with mean 0.53M/sample and standard deviation 0.07M (Figure 1). The distribution in 2021 was bimodal, with the two modes corresponding to different plates. The lower mode was 0.63M, while the higher mode was 0.96M. The average total depth in 2022 was similar to 2020, at 0.53M/sample, but the standard deviation was higher, at 0.17M.

When sample DP was summarized by marker, the data were more consistent across years (Figure 2). The 10th percentile for mean sample DP was 32, 53, and 24 in years 2020, 2021, and 2022, respectively (Fig. 2A). Despite the observed differences in total DP per sample (Figure 1), there was a consistent relationship between the mean (μ) and standard deviation (σ) for sample DP (Fig. 2B). The relationship between these quantities in a Poisson distribution is $\mu = \sigma^{0.5}$, which is a straight line with slope 0.5 on a log-log plot (dashed line in Fig. 2B). The observed data were overdispersed (i.e., more variable) compared to the Poisson, with slope 0.79 (SE 0.00), meaning that $\mu \approx \sigma^{0.8}$.

Tetraploid genotype calls were made with R package *updog* (Gerard et al. 2018), which provides estimates of allelic bias (AB) for each marker—a parameter that measures the relative probability of observing the REF vs. ALT allele. When AB=1, or equivalently $\log_2(AB) = 0$, there is no bias. When AB=2, or equivalently $\log_2(AB)=1$, the REF allele is twice as likely to be observed in a balanced heterozygote. 10% of the markers exhibited bias $|(AB)| > 1$ (Figure S3), but many of these still appeared to have reliable clustering (Figure 3).

V1 DArTag and SNP array genotypes were compared for 1865 common markers across 298 tetraploid clones. Both platforms identified two groups of genetically identical clones, one pair and one threesome, originating from the same F1 populations (Figure S4). This is not uncommon in potato breeding due to how single plant selection is conducted in the first field year. After removing duplicates, the two marker profiles (GBS & array) for every clone were paired under hierarchical clustering (Figure S5), indicating close agreement.

For a quantitative comparison, several measures of error were computed for each marker (Supplemental File S12). Classification error (CE), which is the proportion of samples with different genotype calls, was calculated for both tetraploid (4x) and pseudo-diploid (2x) genotypes (where differences in heterozygote allele dosage are ignored). There was a sharp bend in the cumulative distribution for 2x CE at approximately 0.1 error (Figure 4), with 1647 markers below this threshold (88% of those tested). As expected, fewer markers (1302) satisfied 4x CE < 0.1 because of the difficulty discriminating between heterozygous genotypes. For 4x genotypes, the root-mean-squared-error (RMSE) of allele dosage is potentially more meaningful than CE, and 1547 markers had RMSE < 0.5 (Figure 4), a somewhat arbitrary threshold selected because it represents the midpoint between integer dosages.

Version 2 (V2) of the potato DArTag GBS assay was designed in 2023 and contains 3893 genomic SNPs, of which 2144 were included in Version 1. The additional SNPs were selected from the 31K V4 potato SNP array using the same criteria as before. GBS and SNP array genotypes were compared for 2608 common markers across 78 clones (40 tetraploid, 38 diploid). Given the small number of tetraploids, only the 2x CE criterion was computed, and 2341 markers had 2x CE < 0.1 (Figure S6; Supplemental File S13).

3.2 Imputation

A key role for the DArTag genomic markers is to facilitate imputation to higher density platforms for genomic selection. Among the 298 clones genotyped with both the SNP array and V1 DArTag is a five-parent half-diallel population of 85 clones, with F1 family sizes between 1 and 20 (Table 1). Using a leave-one-family-out cross-validation, we compared the accuracy of two imputation methods, Random Forest (RF) vs. Linkage Analysis (LA). Linkage analysis uses a genetic model of recombination and phased parental genotypes to reconstruct progeny in terms of parental haplotypes. The RMSE for imputing 10K SNP array genotypes from DArTag was always lower with LA compared to RF (Table 1), with overall means of 0.15 and 0.95, respectively.

3.3 Trait markers

The V1 DArTag assay had two trait markers, targeting two different resistance genes (Ry_{adg}, Ry_{sto}) for the most economically important viral pest of potato: potato virus Y (PVY). Both variants had previously been targeted with KASP markers, and for 93 samples genotyped with both KASP and V1 DArTag, there were 2 discrepancies for presence/absence of *Ry*$_{adg}$ (Table S1).

Besides the two PVY markers, the V2 DArTag assay had five additional trait markers with reliable results (Table 1). We had good prior knowledge about the distribution of the PVY and golden cyst nematode (*H1*) resistance genes in our germplasm from other marker systems (SCAR and KASP). Four clones tested positive for the *Ry*$_{sto}$ marker: three were expected based on previous testing, and the fourth was plausible based on its pedigree (Table S2). Many samples tested positive for *Ry*$_{adg}$ and *H1*, which was expected given the high frequency of these variants in the US chip processing germplasm, but the allele dosages for *Ry*$_{adg}$ seemed too high—eight samples were even homozygous tetraploids. To investigate further, we analyzed a partial diallel population (N=123) within the V2 DArTag dataset (Figure S7). Treating the *Ry*$_{adg}$ and *H1* markers as dominant traits, joint linkage analysis identified which parental haplotypes carry the R gene (Figure S8), and corrected dosages were determined by reconstructing the progeny in terms of parental haplotypes (Figure 5). Five triplex and two quadriplex calls for *Ry*$_{adg}$ were corrected down to duplex, and the average upward bias was 0.24 dosage. For *H1*, the original calls were more accurate, with an average bias of only 0.05 dosage.

Little is known about resistance to potato wart disease (*S. endobioticum*) in US germplasm, but given the prevalence of the disease in other parts of the world (Obidiegwu et al., 2014), it has become a higher priority for molecular breeding. One trait marker targets the *Sen3* resistance gene (Table 1), which was detected in four individuals with a common parent, AW07791-2rus. Based on pedigree information, we believe the resistance was inherited from its maternal parent, PALB0303-1 (Elison et al., 2021).

Another trait marker targets *Sli*, a non-S locus F-box protein that disrupts the gametophytic incompatibility system and allows for the development of diploid, inbred lines (Ma et al., 2022; Eggers et al., 2022). The GBS marker showed perfect agreement with prior knowledge for 28 diploid samples based on KASP marker screening (Table S3).

A trait marker for the maturity gene *CDF1* targets the location of the 7 bp indel variants that differentiate alleles 2 and 4 from wild-type alleles, collectively designated group 1. Because of the multi-allelic nature of this variant, correct interpretation requires use of the DArT "missing allele discovery count" (MADC) file, which contains read counts for 81 bp haplotypes surrounding each target variant. Five CDF1 haplotypes were detected in the population (Figure 6A): three were full-length variants of CDF1.1 (Ref, Other1, Other2), one was CDF1.4 (Alt), and one was CDF1.2 (Other3). The validity of the assay was confirmed by comparing the read counts with samples of known CDF1 genotype (Figure 6B), with the complication that CDF1.3, which has an 865 bp transposon insertion at the same position, is not detected. As a result, samples with zero (or near zero, due to sequencing error) counts are interpreted as homozygous for allele 3. And since clones selected under long-day conditions are typically not homozygous wild-type, when CDF1.1 alleles are detected but not alleles 2 or 4, the predicted genotype is 1/3.

Several markers were included in the V2 panel to target *OFP20*, an ovate family protein with a major effect on tuber shape (Wu et al. 2018). This is a complex locus with dozens of predicted alleles (van Eck et al. 2022), so the following approach to interpreting the DArTag markers may not work in all germplasm groups. Marker OFP20_M6_CDS_994 was used to estimate the frequency of *OFP20.1*, which is the most common allele in cultivated germplasm and promotes elongated shape (van Eck et al. 2022). *OFP20.1* was present at a higher frequency in the russet (N=21) vs. chip (N=300) samples from the UW breeding program (Fig. 7A), which is consistent with the long vs. round tuber phenotypes required for those market types. Marker OFP20_M6_CDS_24 was used to estimate the frequency of *OFP20.8*, which was present in 13% of the chip samples. Together with OFP20_M6_CDS_171, which provided information about presence/absence of *OFP20* alleles 2, 3, and 7, the DArTag markers were able to correctly predict five different *OFP20* genotypes (Fig. 7B).

4 DISCUSSION

The potato DArTag assay has several applications in potato breeding. For its price point, an ideal stage of deployment is the first clonal evaluation trial (CET), which typically occurs in the second field year of potato breeding and may have several thousand clones. The DArTag genomic markers provide a genetic fingerprint that can be used to correct pedigree errors (Muñoz et al., 2014; Endelman et al., 2017) and provide a reference genotype for quality control. The clonal trial entries are also candidates for genomic selection, both as potential clonal varieties and as parents to begin the next breeding cycle (Slater et al., 2016; Wu et al., 2023). Limited phenotyping for some traits occurs in the CET, and a genomic relationship matrix computed from DArTag markers could enable a multi-location trial to better estimate genetic values for the target population of environments, i.e., "sparse testing" (Endelman et al. 2014; Jarquin et al. 2020).

Based on previous studies, we expect higher selection accuracy if DArTag markers are first imputed to higher density (Cleveland and Hickey, 2013; Gorjanc et al., 2017). The exploitation of pedigree or family structure during marker imputation in diploids is well documented, with a range of methods and software available depending on the structure of the dataset (Meuwissen and Goddard, 2010; Swarts et al., 2014; Hickey et al., 2015; Whalen et al., 2018; Whalen et al., 2020). The present study has confirmed our hypothesis that linkage analysis is also beneficial for imputation in autopolyploids. DArTag panels are available for several autopolyploid crops besides potato, including alfalfa, blueberry, and sweetpotato (Breeding Insight, 2023), so the software developed for this study (Endelman, 2023b) should benefit other breeding communities. Based on the current functionality of the PolyOrigin software (Zheng et al., 2021), only bi-allelic SNPs were used for imputation, but the DArTag missing allele discovery count (MADC) file offers the possibility of using multi-allelic markers, which are generally more informative for linkage analysis (Luo et al., 2001).

DArTag is not the only option for mid-density genotyping. The PlexSeq platform (AgriPlex, Cleveland, USA) is also widely used, for example in soybean and pearl millet (Semalaiyappan et al., 2023). Leyva Pérez et al. (2022) developed their own targeted GBS platform for potato, PotatoMASH, but the number of genomic markers (339) was small compared to the options with DArTag and PlexSeq.

Besides more genomic markers, a major advantage of the V2 DArTag assay is the additional trait markers (Table 2). It is very valuable to select for resistance to three important pests of potato—PVY, wart, and golden cyst nematode—with the same assay used for genomic selection. Notably absent from this list is potato late blight, caused by the pathogen *P. infestans*. Trait marker blb1_chr08_51070621 was designed to target the *RB/Rpi-blb1* gene (Song et al., 2003; van der Vossen et al., 2003) based on a SNP in the 3´UTR that worked well as a KASP marker (Sorensen et al., 2023). However, no haplotypes were detected in the V2 DArTag experiment for three positive samples from the KASP study. The V2 assay also targeted two genes affecting tuber skin color: *f3´5´h* (Jung et al., 2005) and *an2* (Jung et al., 2009). Both loci have complex allelic series (Hoopes et al. 2022), and more information is needed about their functional effects to guide selection. For tuber shape, we confirmed one trait marker estimates the frequency of the most common long allele, *OFP20.1*. This marker can have an immediate impact on parent selection in the russet market type, where round alleles are undesirable due to their partial dominance.

FIGURE CAPTIONS

Figure 1. Total depth per sample, in million (M) read counts, for three submissions of potato V1 DArTag.

Figure 2. (A) Distribution of the mean sample depth (DP) for V1 DArTag markers. (B) Log-log plot of the relationship between the standard deviation and mean for sample DP. Individual marker points are shown only for 2021 to maintain legibility. Combining the data across years, the overall regression line (not shown) has slope 0.79 (SE 0.00) and $R^2 = 0.99$.

Figure 3. Examples of DArTag markers without (A) vs. with (B) allelic bias. Dashed lines correspond to possible tetraploid allele ratios when there is no allelic bias (1:0, 3:1, 1:1, 1:3, 0:1). (A) solcap_snp_c2_36615 with bias = -0.2. (B) PotVar0072076 with bias = 1.8.

Figure 4. Empirical cumulative distribution for the error between the V1 DArTag and SNP array on 1865 common markers. CE = classification error. RMSE = root-mean-squared-error. 2x = pseudo-diploid genotypes. 4x = tetraploid genotypes.

Figure 5. Original vs. corrected genotypes for the trait markers Ryadg_chr11_2499502 and H1_chr05_52349069. The original genotypes were based on R/updog with a "norm" prior and then corrected based on linkage analysis.

Figure 6. (A) Multiple sequence alignment of the DArTag haplotypes discovered for trait marker CDF1.4_chr05_448021. Haplotypes Ref, Other1, Other2 are CDF1.1 alleles, while Alt is CDF1.4 and Other3 is CDF1.2. (B) Haplotype read counts for samples with known CDF1 genotype.

Figure 7. (A) Distribution of sample allele frequencies for OFP20.1 in round chip (N=300) vs. long russet (N=21) germplasm. (B) Comparison of known OFP20 genotypes with V2 DArT markers. Allele frequency (AF) of OFP20.1 was approximated by ALT frequency at marker OFP20_M6_CDS_994. AF of OFP20.8 was approximated by REF frequency at marker

OFP20_M6_CDS_24. Allele depth (AD) at OFP20_M6_CDS_171 was used to distinguish allele 2 (ALT) from alleles 3 and 7 (REF).

TABLES

Table 1. Half-diallel population with five parents. Above diagonal: F1 population sizes; Below diagonal: imputation root-mean-squared-error with linkage analysis (blue, top) vs. random forest (red, bottom).

	P1	P2	P3	P4	P5
P1		3	5	9	20
P2	0.14 0.95		8	1	9
P3	0.15 0.95	0.17 0.96		7	11
P4	0.13 0.95	0.14 0.93	0.13 0.94		12
P5	0.16 0.95	0.16 0.96	0.14 0.94	0.15 0.95	

Table 2. Validated trait markers in the V2 DArTag assay.

Marker	Target Gene (Trait)	Functional Allele	Reference
Rysto_chr12_2352742	Ry_{sto} (PVY)	REF[a]	Nie et al. (2016)
Ryadg_chr11_2499502	Ry_{adg} (PVY)	ALT	Herrera et al. (2018)
H1_chr05_52349069	*H1* (golden cyst nematode)	ALT	Meade et al. (2020)
Sen3_chr11_2563398	*Sen3* (wart)	ALT	Prodhomme et al. (2019)
Sli_chr12_2372490	*Sli* (self-compatibility)	ALT	Clot et al. (2020)
CDF1.4_chr05_4488021	*CDF1* (maturity)	N/A	Gutaker et al. (2019)
OFP20_M6_CDS_24 OFP20_M6_CDS_171 OFP20_M6_CDS_994	*OFP20* (tuber shape)	N/A	van Eck et al. (2022)

[a] The REF/ALT designation for Rysto_chr12_2352742 is reversed compared to the original design file based on the DMv6.1 reference genome. As a result, the functional allele is REF.

SUPPLEMENTAL FILES

During peer review, the supplemental files are available from the Dryad Digital Repository at https://datadryad.org/stash/share/tdvUt18gBCz6bJ568DaE7mLrWtq55kZzJa_C1uLYSfQ. The permanent link for the supplemental files after publication is https://doi.org/10.5061/dryad.8pk0p2nw4.

File S1. Potato DArTag V1 data for 703 samples (VCF).

File S2. Metadata with year submission for the samples in File S1 (CSV).

File S3. Potato DArTag V2 data for 375 samples (VCF).

File S4. DArT Missing Allele Discovery Counts for the samples in File S3 (CSV).

File S5. Potato V4 SNP array data for 298 samples (VCF).

File S6. Pedigree for diallel population in the V1 DArTag dataset (CSV).

File S7. Phased parental genotypes for the diallel population in File S6 (CSV).

File S8. Pedigree for diallel population in the V2 DArTag dataset (CSV).

File S9. Parental genotype probabilities for the diallel population in File S8 (CSV).

File S10. Trait marker phenotypes for the diallel population in File S8 (CSV).

File S11. Sequence alignment and percent identity matrix for *OFP20* (DOCX).

File S12. Marker concordance between V1 DArTag and the SNP array (CSV).

File S13. Marker concordance between V2 DArTag and the SNP array (CSV).

AUTHOR CONTRIBUTIONS

Jeffrey B. Endelman: Conceptualization, Resources, Investigation, Formal analysis, Software, Supervision, Writing – original draft. **Moctar Kante**: Conceptualization, Resources, Investigation, Formal analysis, Writing – original draft. **Hannele Lindqvist-Kreuze**: Conceptualization, Resources, Supervision. **Andrzej Kilian**: Methodology. **Laura M. Shannon:** Conceptualization, Supervision. **Maria V. Caraza-Harter:** Resources. **Brieanne Vaillancourt:** Formal analysis, Data curation. **Kathrine Mailloux:** Investigation, Resources. **John P. Hamilton:** Formal analysis. **C. Robin Buell:** Conceptualization, Supervision. **All authors**: Writing – review & editing.

ACKNOWLEDGMENTS

Development of the V1 DArTag panel was supported by the CGIAR Excellence in Breeding Platform and Crops to End Hunger Initiative. The USDA National Institute of Food & Agriculture (NIFA) Award 2019-51181-30021 supported development of the V2 DArTag panel, with additional support from PepsiCo for the development of genomic resources to validate markers. Genotyping of UW-Madison potato breeding lines was supported by USDA NIFA Awards 2020-51181-32156 and 2021-34141-35447. We thank D. DeKoeyer for suggesting the marker for *Sen3* resistance.

CONFLICT OF INTEREST STATEMENT

J. Endelman is a member of the editorial board for The Plant Genome. A. Kilian is an employee of Diversity Arrays Technology, the company that provides the DArTag genotyping service.

DATA AVAILABILITY STATEMENT

Supplemental Files S1 – S10, which contain the marker and pedigree data needed to reproduce the results of this study, will be available from the Dryad Digital Repository at https://doi.org/10.5061/dryad.8pk0p2nw4 upon publication. Upon manuscript acceptance, PacBio HiFi sequencing data will be available via the NCBI Sequence Read Archive under BioSamples SAMN38982152, SAMN38982165, SAMN38982166, SAMN38982167, and SAMN38982169, and Illumina sequencing data will be available via the NCBI Sequence Read Archive under BioSamples SAMN39419651, SAMN39670896, SAMN39670897, and SAMN39670898.

REFERENCES

Ali, O.A., O'Rourke, S.M., Amish, S.J., Meek, M.H., Luikart, G., Jeffres, C., & Miller, M.R. (2016). Rad capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics*, *202*, 389–400. https://doi.org/10.1534/genetics.115.183665

Alonge, M., Soyk, S., Ramakrishnan, S. (2019). RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, *20*, 224. https://doi.org/10.1186/s13059-019-1829-6

Amadeu, R.R., Muñoz, P.R., Zheng, C., & Endelman, J.B. (2021). QTL mapping in outbred tetraploid (and diploid) diallel populations. *Genetics*, *219*, iyab124. https://doi.org/10.1093/genetics/iyab124

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., & Johnson, E.A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*, e3376. https://doi.org/10.1371/journal.pone.0003376

Bastien, M., Boudhrioua, C., Fortin, G., Belzile, F., & Phillips, D.W. (2018). Exploring the potential and limitations of genotyping-by-sequencing for SNP discovery and genotyping in tetraploid potato. *Genome*, *61*, 449–456. https://doi.org/10.1139/gen-2017-0236

Breeding Insight (2023). Open Source Genetic Marker Panels. https://breedinginsight.org/breeding-solutions/open-source-dartag-marker-panels/ (Accessed 29 December 2023).

de Bem Oliveira, I., Resende, M.F.R., Ferrão, L.F. V., Amadeu, R.R., Endelman, J.B., Kirst, M., Coelho, A.S.G., & Munoz, P.R. (2019). Genomic prediction of autotetraploids; influence of relationship matrices, allele dosage, and continuous genotyping calls in phenotype prediction. *G3: Genes, Genomes, Genetics*, *9*, 1189–1198. https://doi.org/10.1534/g3.119.400059

Campbell, N.R., Harmon, S.A., & Narum, S.R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, *15*, 855–867. https://doi.org/10.1111/1755-0998.12357

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., & Li H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods, 18*, 170-175. https://doi.org/10.1038/s41592-020-01056-5

Cheng, H., Jarvis, E.D., Fedrigo, O., Koepfli, K.P., Urban, L., Gemmell, N.J., & Li, H. (2022). Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnology, 40*, 1332–1335. https://doi.org/10.1038/s41587-022-01261-x

Cleveland, M.A., & Hickey, J.M. (2013). Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. *Journal of Animal Science*, *91*, 3583–3592. https://doi.org/10.2527/jas2013-6270

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., & Davies, R.M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*. https://doi.org/10.1093/gigascience/giab008

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340

Eggers, E.J., van der Burgt, A., van Heusden, S.A.W., de Vries, M.E., Visser, R.G.F., Bachem, C.W.B., & Lindhout, P. (2021). Neofunctionalisation of the Sli gene leads to self-compatibility and facilitates precision breeding in potato. *Nature Communications*, *12*. https://doi.org/10.1038/s41467-021-24267-6

Elison, G.L., Novy, R.G., & Whitworth, J.L. (2021). Russet Potato Breeding Clones with Extreme Resistance to Potato Virus Y Conferred by Rychc as well as Resistance to Late Blight and Cold-Induced Sweetening. *American Journal of Potato Research*, *98*, 411–419. https://doi.org/10.1007/s12230-021-09852-1

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

Endelman, J.B. (2023a). R/MapRtools. https://github.com/jendelman/MapRtools (version 0.32)

Endelman, J.B. (2023b). R/polyBreedR. https://github.com/jendelman/polyBreedR (version 0.36)

Endelman, J.B., Atlin, G.N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M.E., & Jannink, J.L. (2014). Optimal design of preliminary yield trials with genome-wide markers. *Crop Science*, *54*, 48–59. https://doi.org/10.2135/cropsci2013.03.0154

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

Endelman, J.B., Schmitz Carley, C.A., Bethke, P.C., Coombs, J.J., Clough, M.E., da Silva, W.L., De Jong, W.S., Douches, D.S., Frederick, C.M., Haynes, K.G., Holm, D.G., Miller, J.C., Muñoz, P.R., Navarro, F.M., Novy, R.G., Palta, J.P., Porter, G.A., Rak, K.T., Sathuvalli, V.R., Thompson, A.L., & Yencho, G.C. (2018). Genetic variance partitioning and Genome-wide prediction with allele dosage information in autotetraploid potato. *Genetics*, *209*, 77–87. https://doi.org/10.1534/genetics.118.300685

Endelman, J.B., Schmitz Carley, C.A., Douches, D.S., Coombs, J.J., Bizimungu, B., De Jong, W.S., Haynes, K.G., Holm, D.G., Miller, J.C., Novy, R.G., Palta, J.P., Parish, D.L., Porter, G.A., Sathuvalli, V.R., Thompson, A.L., & Yencho, G.C. (2017). Pedigree Reconstruction with Genome-Wide Markers in Potato. *American Journal of Potato Research*, *94*, 184–190. https://doi.org/10.1007/s12230-016-9556-y

Excellence in Breeding (2022). Mid-density genotyping service. https://excellenceinbreeding.org/toolbox/services/mid-density-genotyping-service (accessed 29 December 2023)

Felcher, K.J., Coombs, J.J., Massa, A.N., Hansey, C.N., Hamilton, J.P., Veilleux, R.E., Buell, C.R., & Douches, D.S. (2012). Integration of two diploid potato linkage maps with the potato genome sequence. *PloS ONE*, *7*, e36347. https://doi.org/10.1371/journal.pone.0036347

Fu, A., Narasimhan, B., & Boyd, S. (2020). CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, *94*, 1–34. https://doi.org/10.18637/jss.v094.i14

Gasc, C., Peyretaillade, E., & Peyret, P. (2016). Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Research*, *44*, 4504–4518. https://doi.org/10.1093/nar/gkw309

Gerard, D., Ferrão, L.F.V., Garcia, A.A.F., & Stephens, M. (2018). Genotyping polyploids from messy sequencing data. *Genetics*, *210*, 789–807. https://doi.org/10.1534/genetics.118.301468

Gorjanc, G., Battagin, M., Dumasy, J.F., Antolin, R., Gaynor, R.C., & Hickey, J.M. (2017). Prospects for cost-effective genomic selection via accurate within-family imputation. *Crop Science*, *57*, 216–228. https://doi.org/10.2135/cropsci2016.06.0526

Gutaker, R.M., Weiß, C.L., Ellis, D., Anglin, N.L., Knapp, S., Luis Fernández-Alonso, J., Prat, S., & Burbano, H.A. (2019). The origins and adaptation of European potatoes reconstructed from historical genomes. *Nature Ecology and Evolution*, *3*, 1093–1101. https://doi.org/10.1038/s41559-019-0921-3

Hamilton, J.P., Hansey, C.N., Whitty, B.R., Stoffel, K., Massa, A.N., Deynze, A. Van, Jong, W.S. De, Douches, D.S., & Buell, C.R. (2011). Single nucleotide polymorphism discovery in elite North American potato germplasm. *BMC Genomics*, *302*. https://doi.org/ 10.1186/1471-2164-12-302

Hardenbol, P., Banér, J., Jain, M., Nilsson, M., Namsaraev, E.A., Karlin-Neumann, G.A., Fakhrai-Rad, H., Ronaghi, M., Willis, T.D., Landegren, U., & Davis, R.W. (2003) Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature Biotechnology, 21*, 673–678. https://doi.org/10.1038/nbt821

Hardigan, M.A., Feldmann, M.J., Carling, J., Zhu, A., Kilian, A., Famula, R.A., Cole, G.S., & Knapp, S.J. (2023). A medium-density genotyping platform for cultivated strawberry using DArTag technology. *Plant Genome*, *16*. https://doi.org/10.1002/tpg2.20399

Herrera, M. del R., Vidalon, L.J., Montenegro, J.D., Riccio, C., Guzman, F., Bartolini, I., & Ghislain, M. (2018). Molecular and genetic characterization of the Ryadg locus on chromosome XI from Andigena potatoes conferring extreme resistance to potato virus Y. *Theoretical and Applied Genetics*, *131*, 1925–1938. https://doi.org/10.1007/s00122-018-3123-5

Hickey, J.M., Gorjanc, G., Varshney, R.K., & Nettelblad, C. (2015). Imputation of single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations with a hidden markov model. *Crop Science*, *55*, 1934–1946. https://doi.org/10.2135/cropsci2014.09.0648

Jarquin, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J.W.R., Pazaran, G.C., Burgueño, J., Pacheco, A., Grondona, M., Wimmer, V., & Prasanna, B.M. (2020). Genomic prediction enhanced sparse testing for multi-environment trials. *G3: Genes, Genomes, Genetics*, *10*, 2725–2739. https://doi.org/10.1534/g3.120.401349

Jung, C.S., Griffiths, H.M., De Jong, D.M., Cheng, S., Bodis, M., & De Jong, W.S. (2005). The potato P locus codes for flavonoid 3’,5’-hydroxylase. *TAG. Theoretical and Applied Genetics, 110*, 269–75. https://doi.org/10.1007/s00122-004-1829-z

Jung, C.S., Griffiths, H.M., De Jong, D.M., Cheng, S., Bodis, M., Kim, T.S., & De Jong, W.S. (2009). The potato developer (D) locus encodes an R2R3 MYB transcription factor that regulates expression of multiple anthocyanin structural genes in tuber skin. *Theoretical and Applied Genetics*, *120*, 45–57. https://doi.org/10.1007/s00122-009-1158-3

Kloosterman, B., Abelenda, J. a, Gomez, M.D.M.C., Oortwijn, M., de Boer, J.M., Kowitwanich, K., Horvath, B.M., van Eck, H.J., Smaczniak, C., Prat, S., Visser, R.G.F., & Bachem, C.W.B. (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes.. *Nature*, *495*, 246–50. https://doi.org/10.1038/nature11912

Knaus, B.J., & Grünwald, N.J. (2017). VCFR: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, *17*, 44–53. https://doi.org/10.1111/1755-0998.12549

Leyva-Pérez, M. de la O., Vexler, L., Byrne, S., Clot, C.R., Meade, F., Griffin, D., Ruttink, T., Kang, J., & Milbourne, D. (2022). PotatoMASH—A Low Cost, Genome-Scanning Marker System for Use in Potato Genomics and Genetics Applications. *Agronomy*, *12*. https://doi.org/10.3390/agronomy12102461

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22. https://doi.org/10.1177/154405910408300516

Luo, Z.W., Hackett, C.A., Bradshaw, J.E., McNicol, J.W., & Milbourne, D. Construction of a Genetic Linkage Map in Tetraploid Species Using Molecular Markers. *Genetics, 157,* 1369–1385.

Ma, L., Zhang, C., Zhang, B., Tang, F., Li, F., Liao, Q., Tang, D., Peng, Z., Jia, Y., Gao, M., Guo, H., Zhang, J., Luo, X., Yang, H., Gao, D., Lucas, W.J., Li, C., Huang, S., & Shang, Y. (2021). A nonS-locus F-box gene breaks self-incompatibility in diploid potatoes. *Nature Communications*, *12*. https://doi.org/10.1038/s41467-021-24266-7

Matias, F.I., Xavier Meireles, K.G., Nagamatsu, S.T., Lima Barrios, S.C., Borges do Valle, C., Carazzolle, M.F., Fritsche-Neto, R., & Endelman, J.B. (2019). Expected Genotype Quality and Diploidized Marker Data from Genotyping-by-Sequencing of *Urochloa spp.* Tetraploids. *The Plant Genome*, *12*. https://doi.org/10.3835/plantgenome2019.01.0002

Meuwissen, T., & Goddard, M. (2010). The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. *Genetics*, *185*, 1441–1449. https://doi.org/10.1534/genetics.110.113936

Muñoz, P.R., Resende, M.F.R., Huber, D.A., Quesada, T., Resende, M.D.V., Neale, D.B., Wegrzyn, J.L., Kirst, M., & Peter, G.F. (2014). Genomic relationship matrix for correcting pedigree errors in breeding populations: Impact on genetic parameters and genomic selection accuracy. *Crop Science*, *54*, 1115–1123. https://doi.org/10.2135/cropsci2012.12.0673

Nie, X., Sutherland, D., Dickison, V., Singh, M., Murphy, A.M., & De Koeyer, D. (2016). Development and validation of high-resolution melting markers derived from Rysto STS markers for high-throughput marker-assisted selection of potato carrying Rysto. *Phytopathology*, *106*, 1366–1375. https://doi.org/10.1094/PHYTO-05-16-0204-R

Obidiegwu, J.E., Flath, K., & Gebhardt, C. (2014). Managing potato wart: A review of present research status and future perspective. *Theoretical and Applied Genetics*, *127*, 763–780. https://doi.org/10.1007/s00122-014-2268-0

Pham, G.M., Hamilton, J.P., Wood, J.C., Burke, J.T., Zhao, H., Vaillancourt, B., Ou, S., Jiang, J., & Robin Buell, C. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience*, *9*. https://doi.org/10.1093/gigascience/giaa100

Prodhomme, C., Esselink, D., Borm, T., Visser, R.G.F., Van Eck, H.J., & Vossen, J.H. (2019). Comparative Subsequence Sets Analysis (CoSSA) is a robust approach to identify haplotype specific SNPs; Mapping and pedigree analysis of a potato wart disease resistance gene *Sen3*. *Plant Methods*, *15*. https://doi.org/10.1186/s13007-019-0445-5

Semalaiyappan, J., Selvanayagam, S., Rathore, A., Gupta, S.K., Chakraborty, A., Gujjula, K.R., Haktan, S., Viswanath, A., Malipatil, R., Shah, P., Govindaraj, M., Ignacio, J.C., Reddy, S., Singh, A.K., & Thirunavukkarasu, N. (2023). Development of a new AgriSeq 4K mid-density SNP genotyping panel and its utility in pearl millet breeding. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.1068883

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE, 11,* e0163962. https://doi.org/10.1371/journal.pone.0163962

Sharma, S.K., & Bryan, G.J. (2017). Genome Sequence-Based Marker Development and Genotyping in Potato. In S. Kumar Chakrabarti, C. Xie & J. Kumar Tiwari (Eds.), *The Potato Genome. Compendium of Plant Genomes* (pp. 307–326). Springer.

Slater, A.T., Cogan, N.O.I., Forster, J.W., Hayes, B.J., & Daetwyler, H.D. (2016). Improving Genetic Gain with Genomic Selection in Autotetraploid Potato. *The Plant Genome*, *9*. https://doi.org/10.3835/plantgenome2016.02.0021

Song, J., Bradeen, J.M., Kristine Naess, S., Raasch, J.A., Wielgus, S.M., Haberlach, G.T., Liu, J., Kuang, H., Austin-Phillips, S., Robin Buell, C., Helgeson, J.P., & Jiang, J. (2003). Gene *RB* cloned from *Solanum bulbocastanum* confers broad spectrum resistance to potato late blight. *Proceedings of the National Academy of Sciences, 100,* 9128–9133.

Song, L., & Endelman, J.B. (2023). Using haplotype and QTL analysis to fix favorable alleles in diploid potato breeding. *The Plant Genome, e20339.* https://doi.org/10.1002/tpg2.20339

Sorensen, P.L., Christensen, G., Karki, H.S., & Endelman, J.B. (2023). A KASP Marker for the Potato Late Blight Resistance Gene *RB/Rpi-blb1*. *American Journal of Potato Research*, *100*, 240–246. https://doi.org/10.1007/s12230-023-09914-6

Swarts, K., Li, H., Romero Navarro, J.A., An, D., Romay, M.C., Hearne, S., Acharya, C., Glaubitz, J.C., Mitchell, S., Elshire, R.J., Buckler, E.S., & Bradbury, P.J. (2014). Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. *The Plant Genome*, *7*. https://doi.org/10.3835/plantgenome2014.05.0023

Uitdewilligen, J.G.A.M.L., Wolters, A.A., Bjorn, B.D., Borm, T.J.A., Visser, R.G.F., & Eck, H.J. Van. (2013). A Next-Generation Sequencing Method for Genotyping- by-Sequencing of Highly Heterozygous Autotetraploid Potato, *PLoS ONE, 8*, e0141940. https://doi.org/10.1371/journal.pone.0062355

Vaillancourt, B., & Buell, C.R. (2019). High molecular weight DNA isolation method from diverse plant species for use with Oxford Nanopore sequencing. *bioRxiv*, 783159. https://doi.org/10.1101/783159.

Voorrips, R.E., Gort, G., & Vosman, B. (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models.. *BMC Bioinformatics*, *12*, 172. https://doi.org/10.1186/1471-2105-12-172

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

van der Vossen, E., Sikkema, A., Te Lintel Hekkert, B., Gros, J., Stevens, P., Muskens, M., Wouters, D., Pereira, A., Stiekema, W., & Allefs, S. (2003). An ancient R gene from the wild potato species Solanum bulbocastanum confers broad-spectrum resistance to Phytophthora infestans in cultivated potato and tomato. *Plant Journal*, *36*, 867–882. https://doi.org/10.1046/j.1365-313X.2003.01934.x

van Eck, H.J., Oortwijn, M.E.P., Terpstra, I.R., van Lieshout, N.H.M., van der Knaap, E., Willemsen, J.H., & Bachem, C.W.B. (2022). Engineering of tuber shape in potato (Solanum tuberosum) with marker assisted breeding or genetic modification using StOFP20. *Research Square*, *PREPRINT*. https://doi.org/10.21203/rs.3.rs-1807189/v1

Wang, W., & Yan, J. (2021). Shape-Restricted Regression Splines with R Package splines2. *Journal of Data Science*, 498–517. https://doi.org/10.6339/21-JDS1020

Whalen, A., Gorjanc, G., & Hickey, J.M. (2020). AlphaFamImpute: High-accuracy imputation in full-sib families from genotype-by-sequencing data. *Bioinformatics*, *36*, 4369–4371. https://doi.org/10.1093/bioinformatics/btaa499

Whalen, A., Ros-Freixedes, R., Wilson, D.L., Gorjanc, G., & Hickey, J.M. (2018). Hybrid peeling for fast and accurate calling, phasing, and imputation with sequence data of any coverage in pedigrees. *Genetics Selection Evolution*, *50*. https://doi.org/10.1186/s12711-018-0438-2

Wu, P.Y., Stich, B., Renner, J., Muders, K., Prigge, V., & van Inghelandt, D. (2023). Optimal implementation of genomic selection in clone breeding programs—Exemplified in potato: I. Effect of selection strategy, implementation stage, and selection intensity on short-term genetic gain. *Plant Genome*, *16*. https://doi.org/10.1002/tpg2.20327

Wu, S., Zhang, B., Keyhaninejad, N., Rodríguez, G.R., Kim, H.J., Chakrabarti, M., Illa-Berenguer, E., Taitano, N.K., Gonzalo, M.J., Díaz, A., Pan, Y., Leisner, C.P., Halterman, D., Buell, C.R., Weng, Y., Jansky, S.H., van Eck, H., Willemsen, J., Monforte, A.J., Meulia, T., & van der Knaap, E. (2018). A common genetic mechanism underlies morphological diversity in fruits and other plant organs. *Nature Communications*, *9*. https://doi.org/10.1038/s41467-018-07216-8

Zhao, D., Mejia-Guerra, K.M., Mollinari, M., Samac, D., Irish, B., Heller-Uszynska, K., Beil, C.T., & Sheehan, M.J. (2023). A public mid-density genotyping platform for alfalfa (Medicago sativa L.). *Genetic Resources*, *4*, 55–63. https://doi.org/10.46265/genresj.EMOR6509

Zheng, C., Amadeu, R.R., Munoz, P.R., & Endelman, J.B. (2021). Haplotype reconstruction in connected tetraploid F1 populations. *Genetics*, *219*. https://doi.org/10.1093/genetics/iyab106

Zych, K., Gort, G., Maliepaard, C.A., Jansen, R.C., & Voorrips, R.E. (2019). FitTetra 2.0 - Improved genotype calling for tetraploids with multiple population and parental data support. *BMC Bioinformatics*, *20*. https://doi.org/10.1186/s12859-019-2703-y

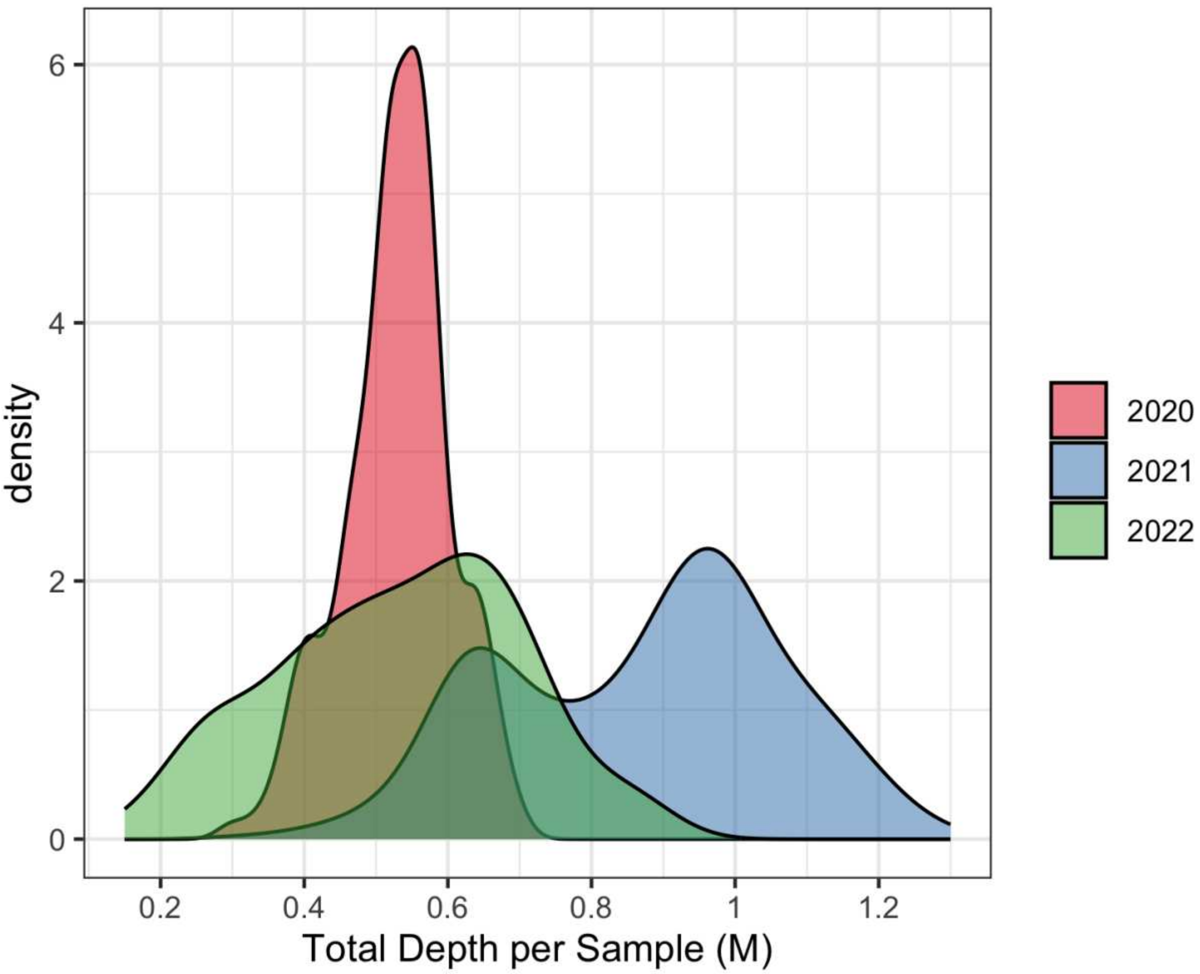


Figure 1. Total depth per sample, in million (M) read counts, for three submissions of potato V1 DArTag.

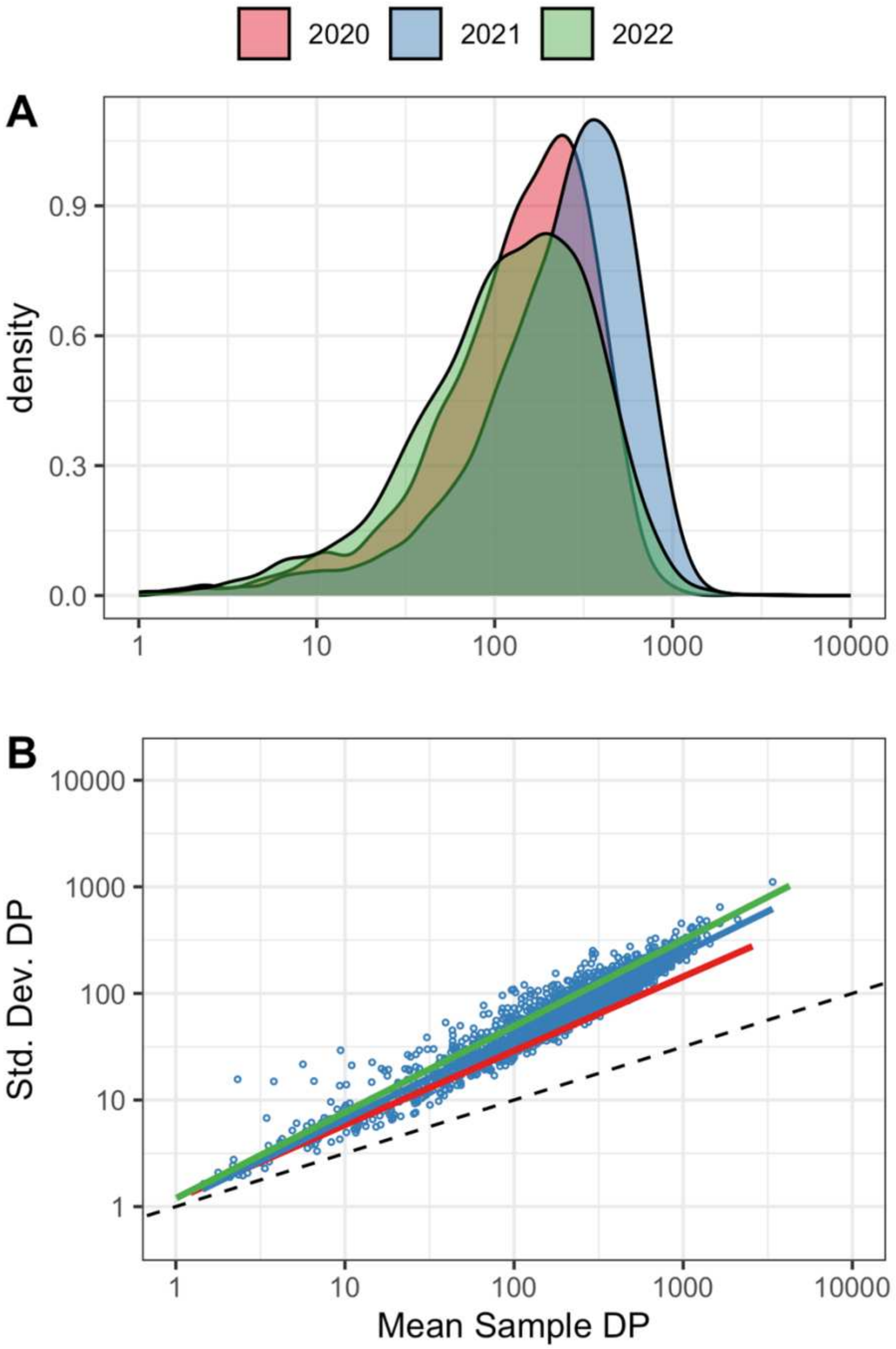


Figure 2. (A) Distribution of the mean sample depth (DP) for V1 DArTag markers. (B) Log-log plot of the relationship between the standard deviation and mean for sample DP. Individual marker points are shown only for 2021 to maintain legibility. Combining the data across years, the overall regression line (not shown) has slope 0.79 (SE 0.00) and $R^2 = 0.99$.

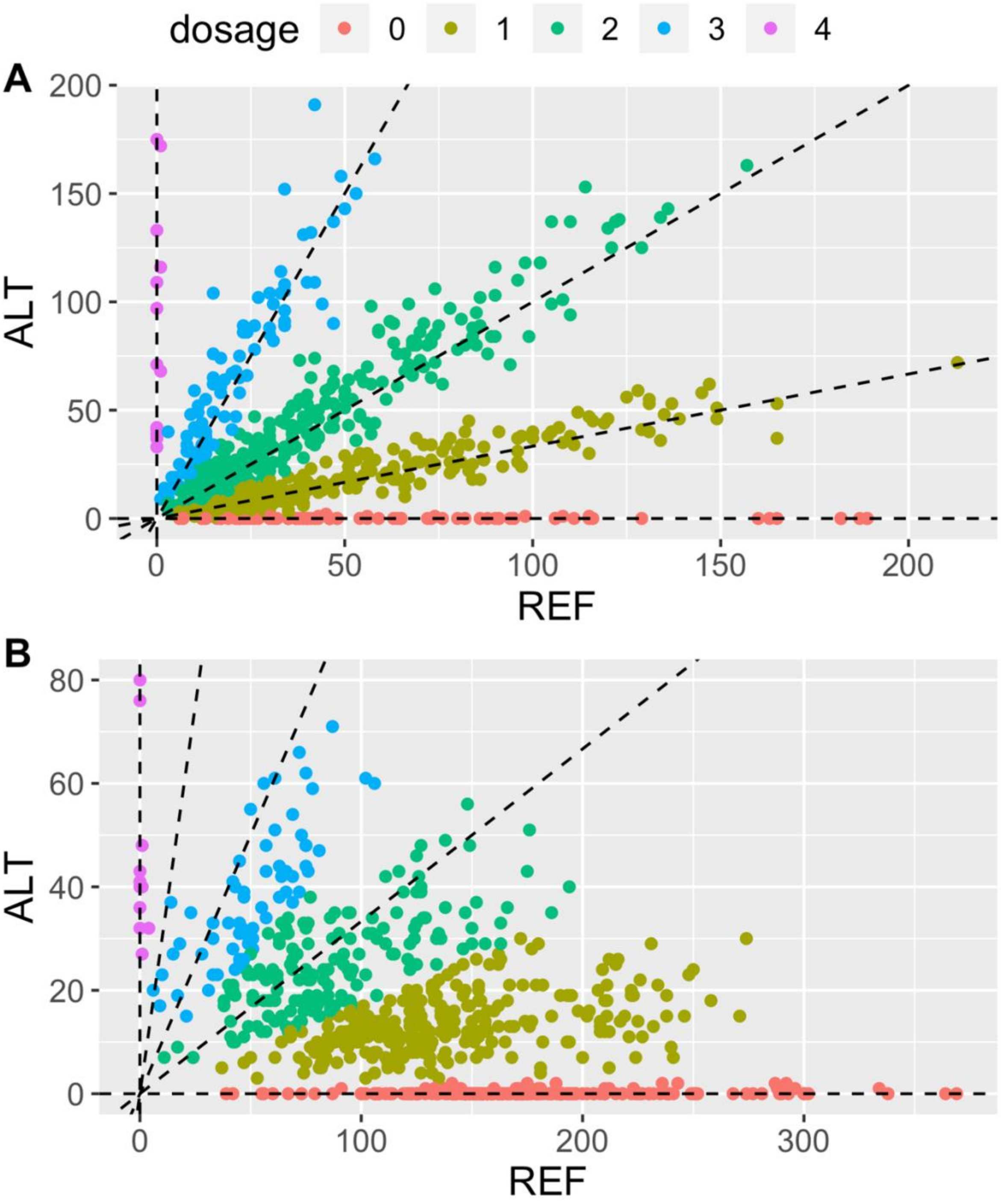


Figure 3. Examples of DArTag markers without (A) vs. with (B) allelic bias. Dashed lines correspond to possible tetraploid allele ratios when there is no allelic bias (1:0, 3:1, 1:1, 1:3, 0:1). (A) solcap_snp_c2_36615 with bias = -0.2. (B) PotVar0072076 with bias = 1.8.

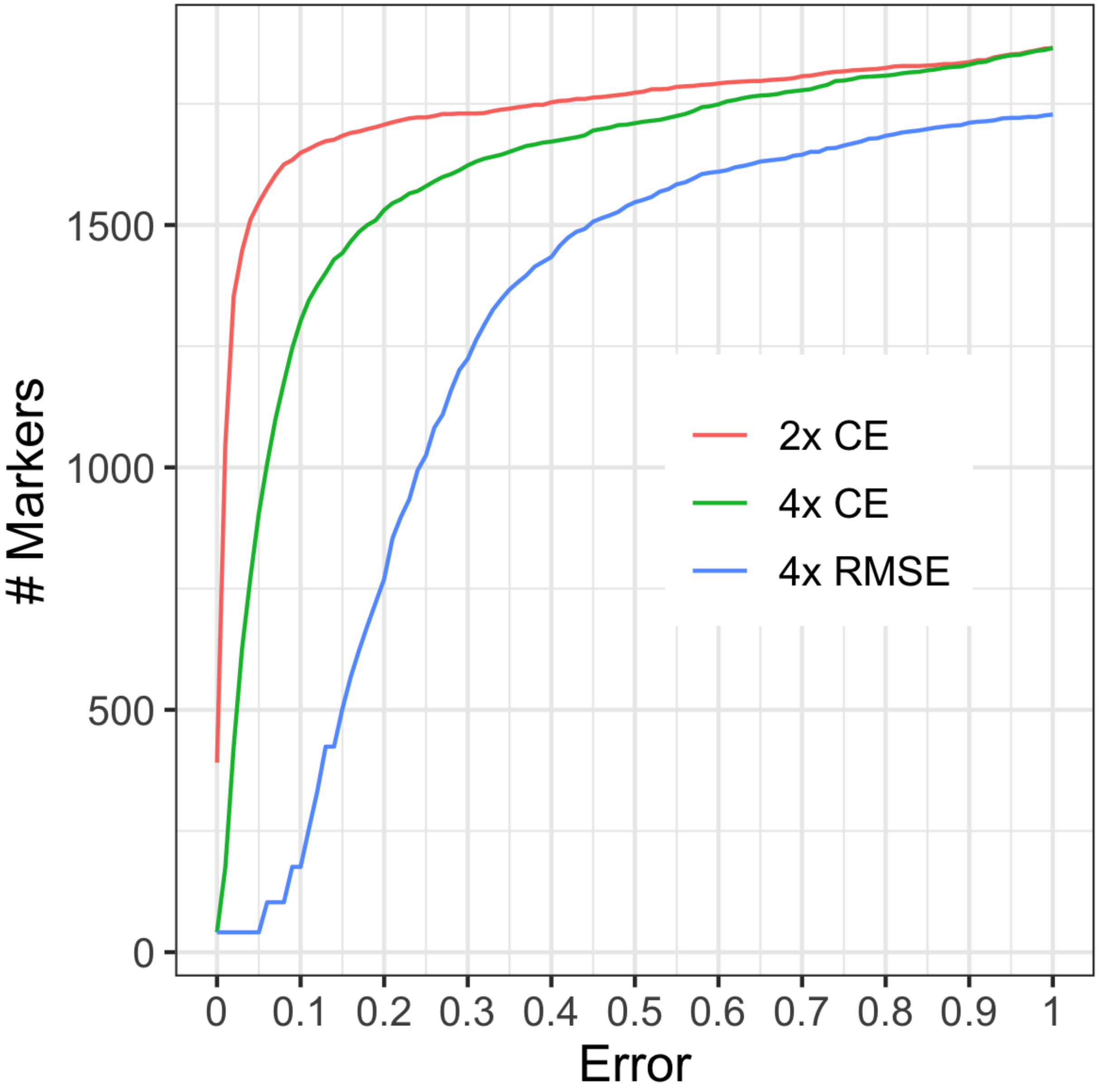
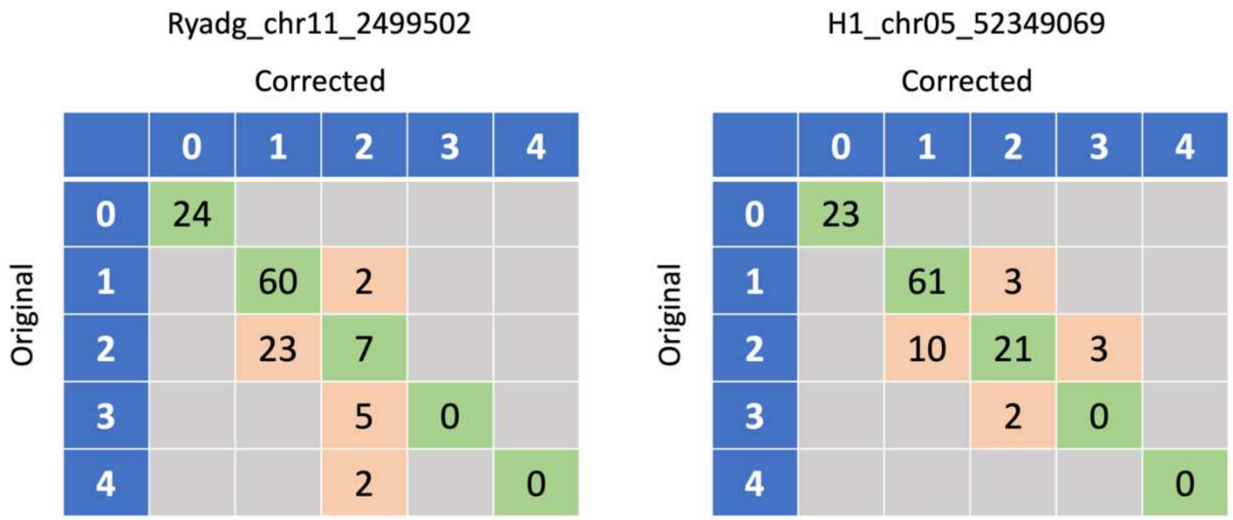


Figure 4. Empirical cumulative distribution for the error between the V1 DArTag and SNP array on 1865 common markers. CE = classification error. RMSE = root-mean-squared-error. 2x = pseudo-diploid genotypes. 4x = tetraploid genotypes.



Ryadg_chr11_2499502

Corrected

Original

	0	1	2	3	4
0	24				
1		60	2		
2		23	7		
3			5	0	
4			2		0

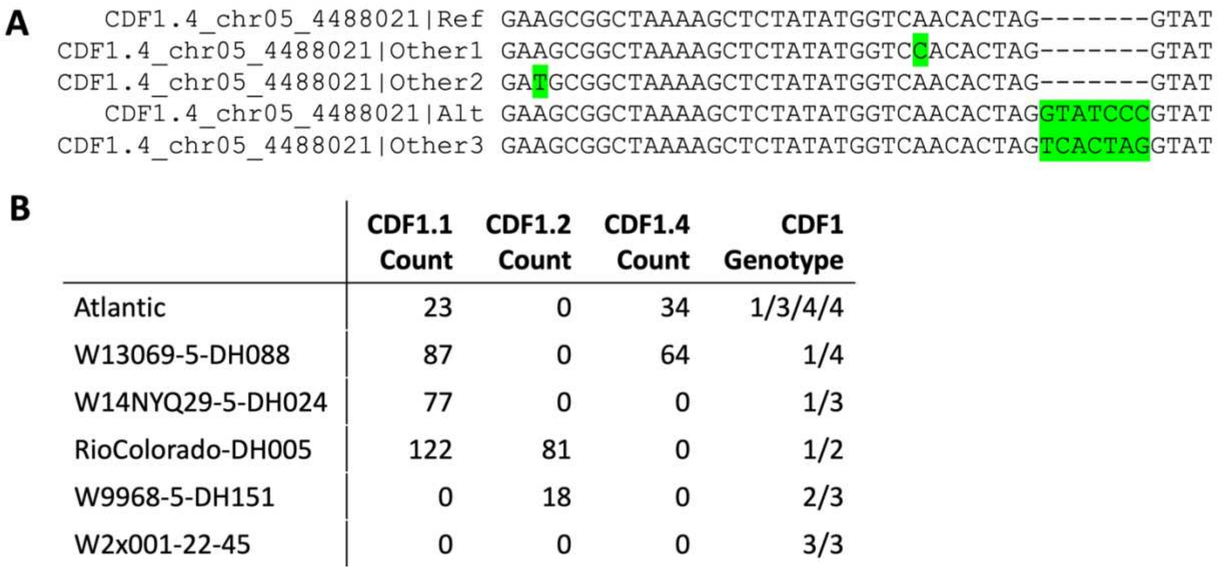
H1_chr05_52349069

Corrected

Original

	0	1	2	3	4
0	23				
1		61	3		
2		10	21	3	
3			2	0	
4					0

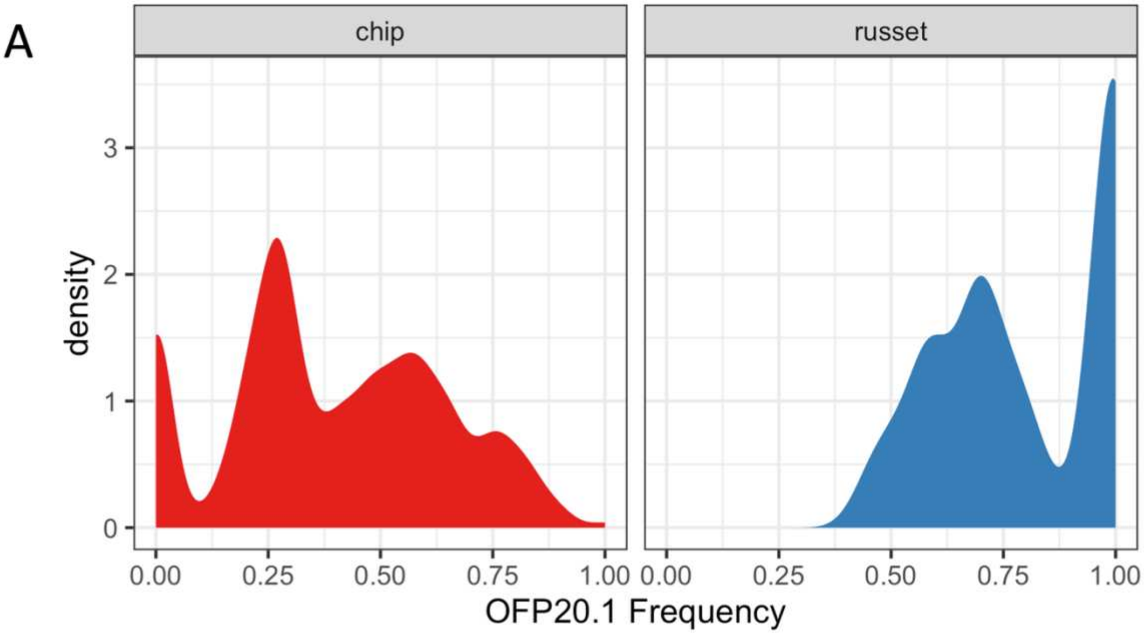
Figure 5. Original vs. corrected genotypes for the trait markers Ryadg_chr11_2499502 and H1_chr05_52349069. The original genotypes were based on R/updog with a "norm" prior and then corrected based on linkage analysis.



```
     CDF1.4_chr05_4488021|Ref GAAGCGGCTAAAAGCTCTATATGGTCAACACTAG-------GTAT
  CDF1.4_chr05_4488021|Other1 GAAGCGGCTAAAAGCTCTATATGGTCCACACTAG-------GTAT
  CDF1.4_chr05_4488021|Other2 GATGCGGCTAAAAGCTCTATATGGTCAACACTAG-------GTAT
     CDF1.4_chr05_4488021|Alt GAAGCGGCTAAAAGCTCTATATGGTCAACACTAGGTATCCCGTAT
  CDF1.4_chr05_4488021|Other3 GAAGCGGCTAAAAGCTCTATATGGTCAACACTAGTCACTAGGTAT
```

	CDF1.1 Count	CDF1.2 Count	CDF1.4 Count	CDF1 Genotype
Atlantic	23	0	34	1/3/4/4
W13069-5-DH088	87	0	64	1/4
W14NYQ29-5-DH024	77	0	0	1/3
RioColorado-DH005	122	81	0	1/2
W9968-5-DH151	0	18	0	2/3
W2x001-22-45	0	0	0	3/3

Figure 6. (A) Multiple sequence alignment of the DArTag haplotypes discovered for trait marker CDF1.4_chr05_448021. Haplotypes Ref, Other1, Other2 are CDF1.1 alleles, while Alt is CDF1.4 and Other3 is CDF1.2. (B) Haplotype read counts for samples with known CDF1 genotype.



B

	OFP20 Genotype	AF Allele 1	AF Allele 8	AD Alleles 2,3+7
CO99076-6R-DH002	1/1	1.00	0.00	0,0
W14NYQ29-5-DH024	3/3	0.00	0.00	0,240
W14NYQ9-2-DH137	1/3	0.49	0.00	0,115
W8890-1R-DH002	1/7	0.64	0.00	0,70
W9968-5-DH084	1/8	0.44	0.60	0,1

Figure 7. (A) Distribution of sample allele frequencies for OFP20.1 in round chip (N=300) vs. long russet (N=21) germplasm. (B) Comparison of known OFP20 genotypes with V2 DArT markers. Allele frequency (AF) of OFP20.1 was approximated by ALT frequency at marker OFP20_M6_CDS_994. AF of OFP20.8 was approximated by REF frequency at marker OFP20_M6_CDS_24. Allele depth (AD) at OFP20_M6_CDS_171 was used to distinguish allele 2 (ALT) from alleles 3 and 7 (REF).

Supplemental Figures and Tables

Endelman *et al.* Targeted genotyping-by-sequencing of potato and software for imputation.

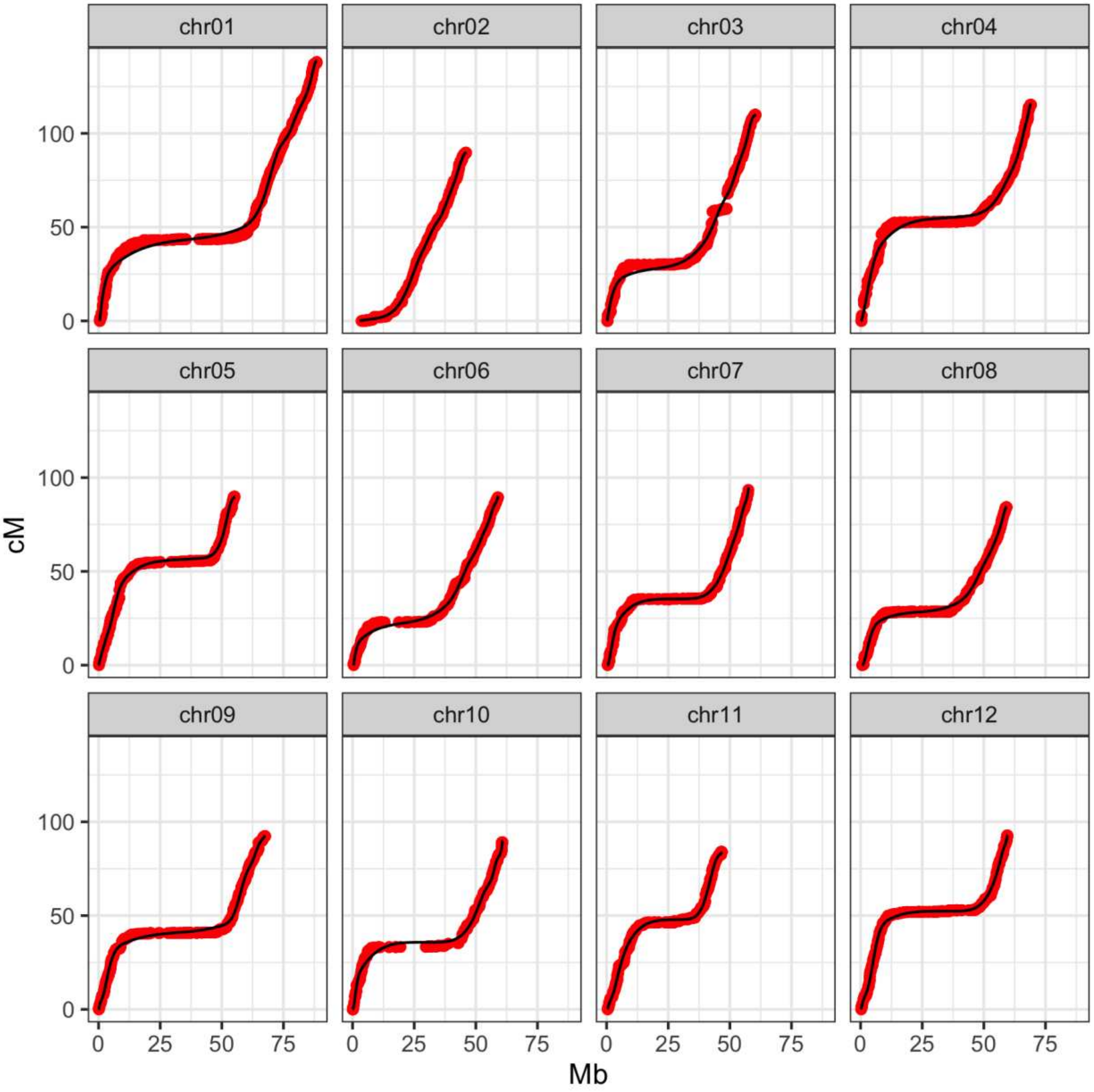


Figure S1. Marey Map of the potato genome. Horizontal axis is the DMv6.1 reference genome position (Pham et al., 2020).

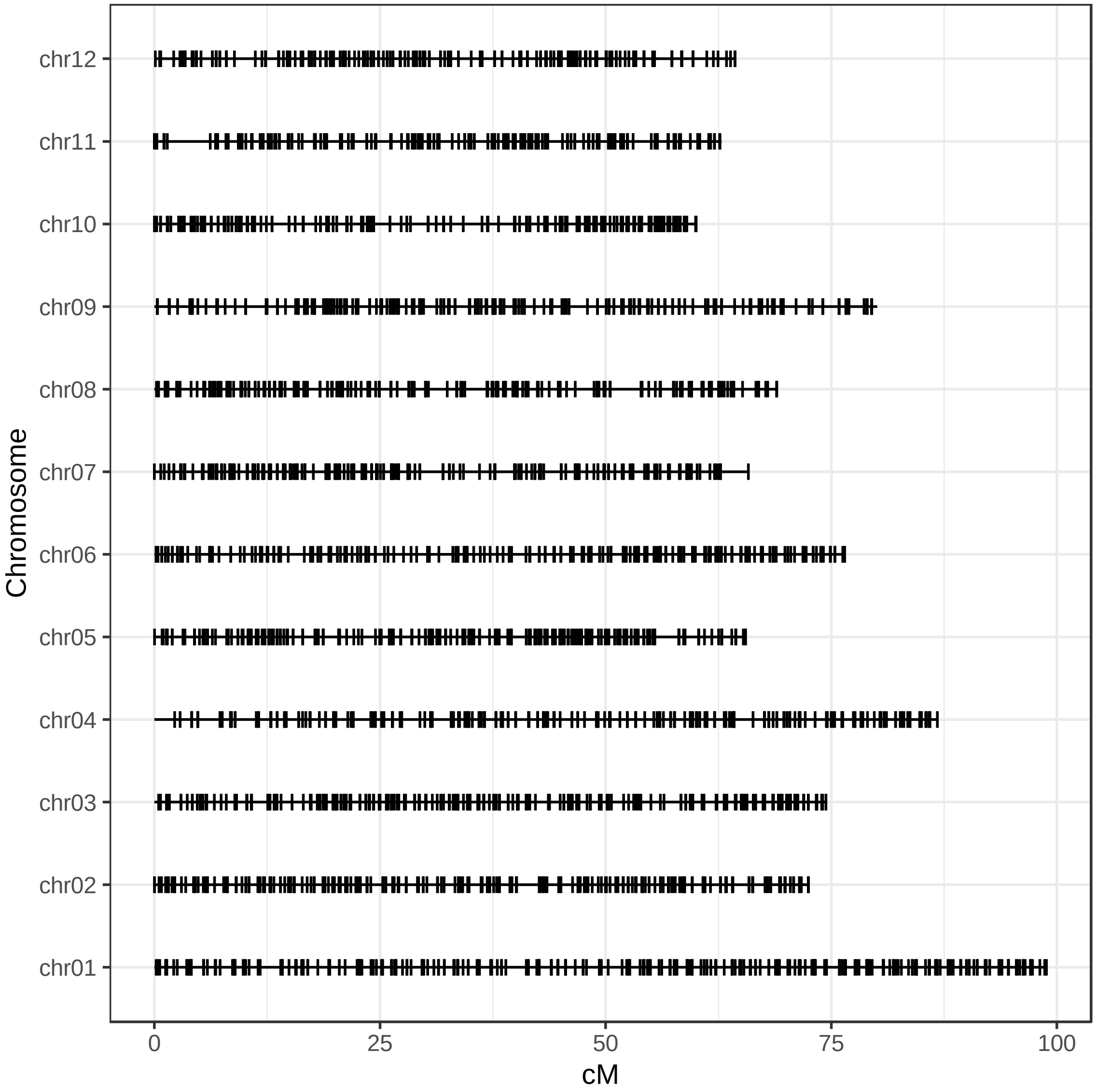


Figure S2. Distribution of the 2501 genomic markers for V1 DArTag.

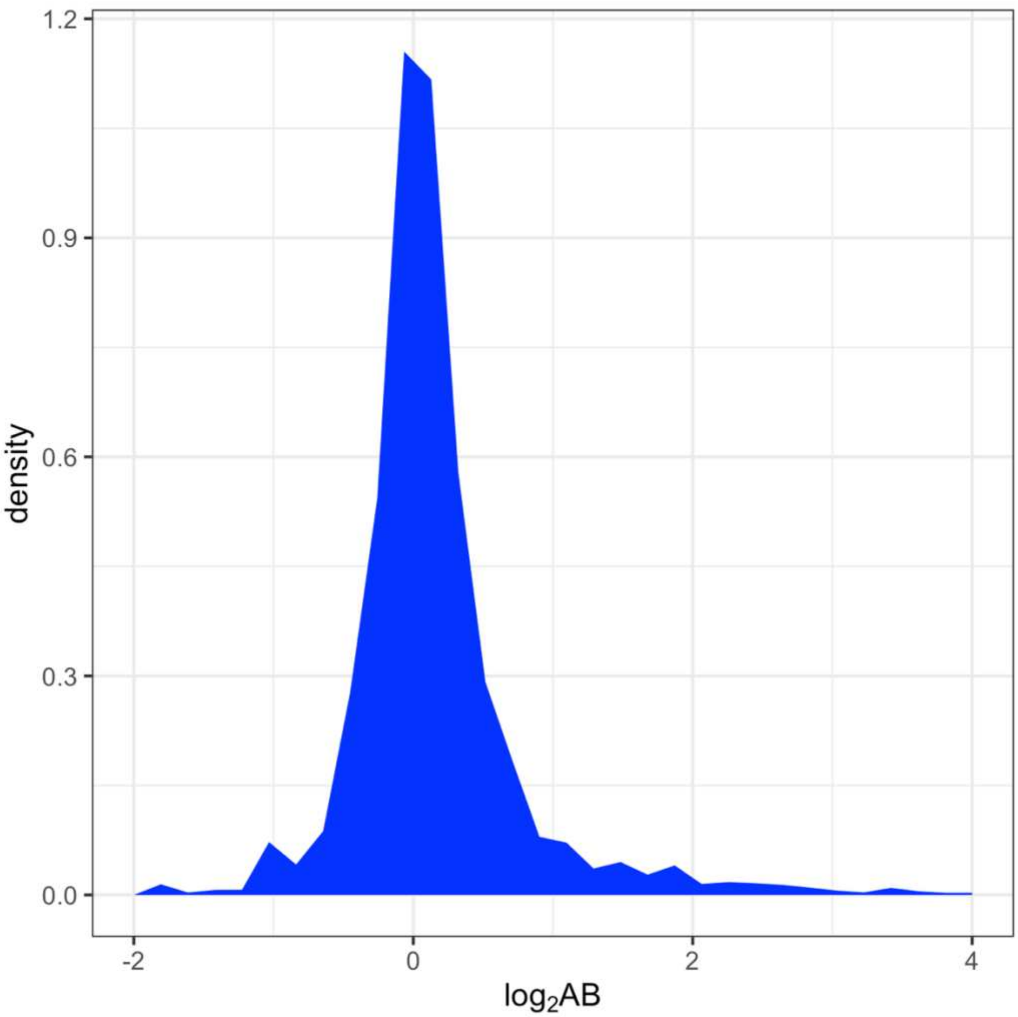
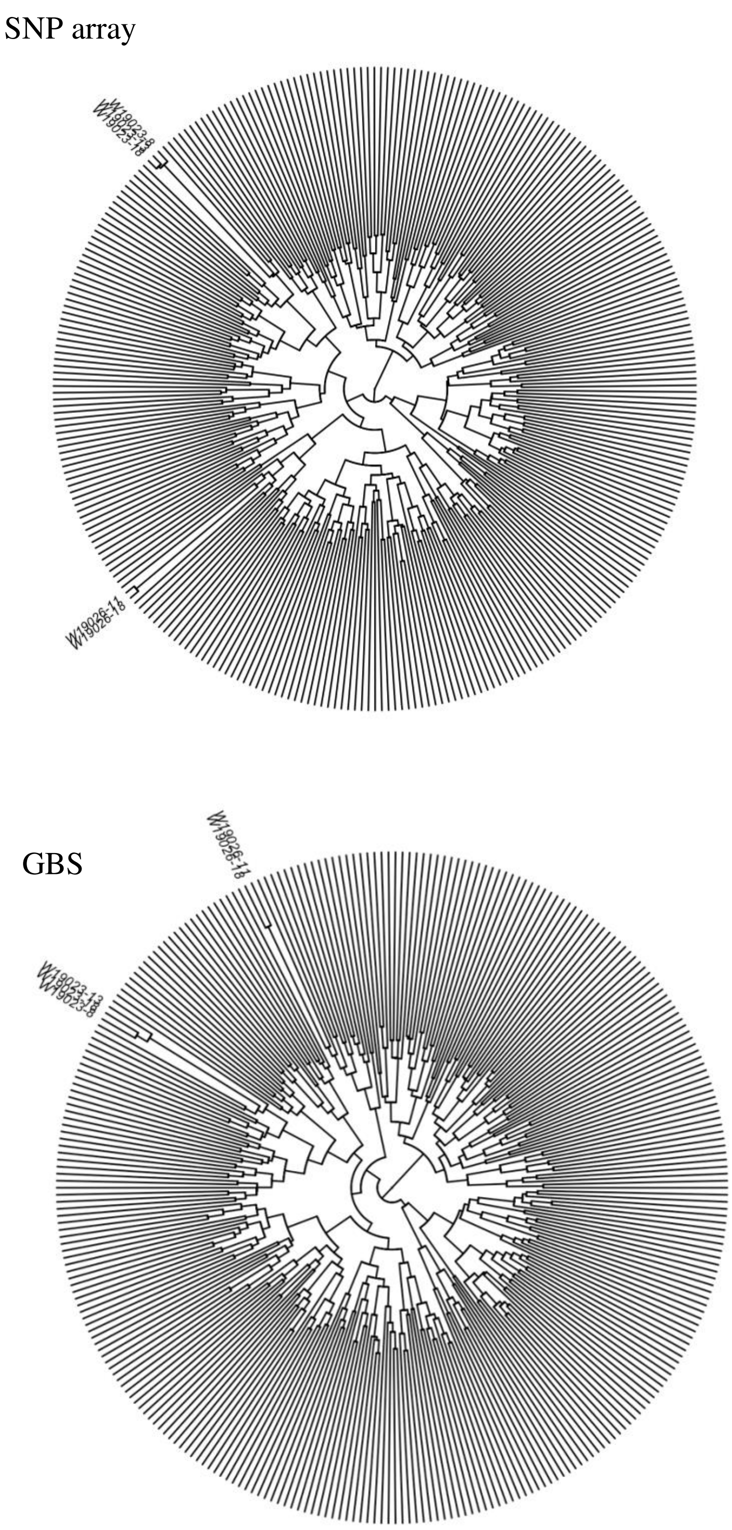


Figure S3. Distribution of allele bias (AB) estimates, where AB=1 indicates no bias, and values greater than 1 indicate bias toward the REF allele



SNP array

GBS

Figure S4. Hierarchical clustering based on SNP array (top) or GBS (bottom) data. Both platforms identified two groups of genetically identical clones, one pair and one threesome, originating from the same F1 populations

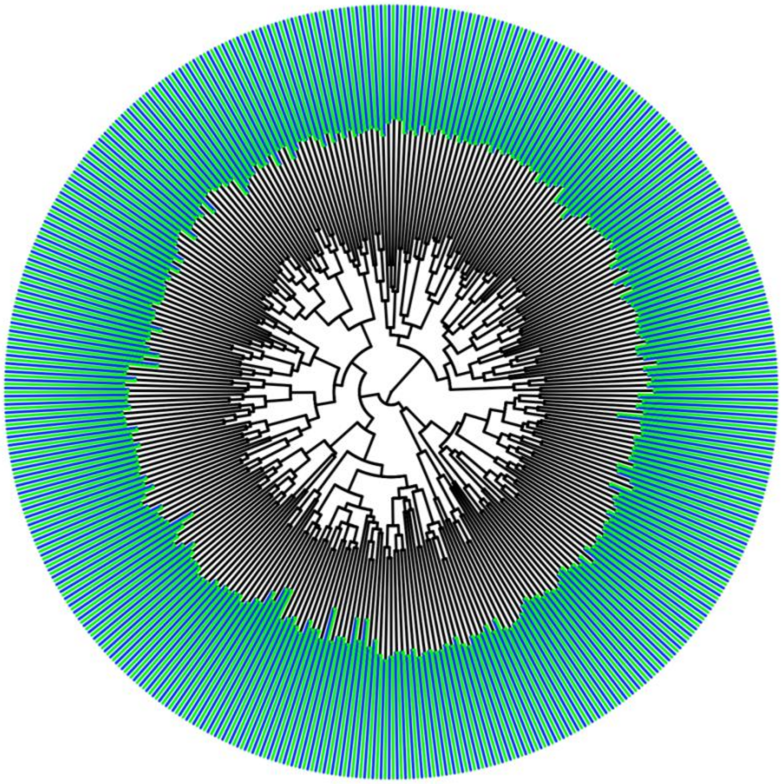


Figure S5. Joint clustering of SNP array (blue) and GBS (green) samples. The two marker profiles for every clone were paired.

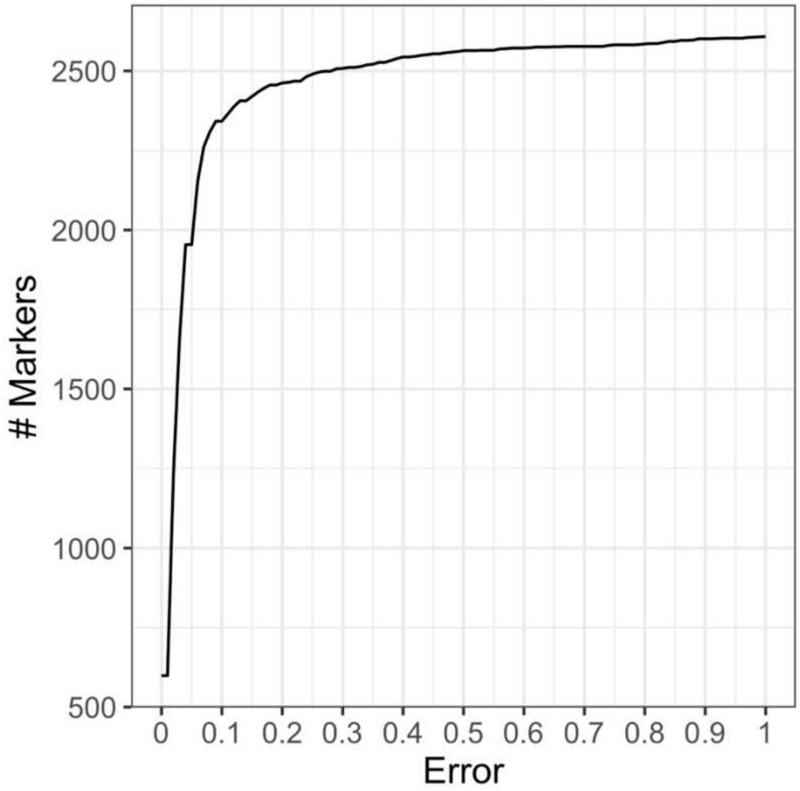


Figure S6. Empirical cumulative distribution for 2x classification error for 2608 common markers between the V4 SNP array and V2 DArTag.

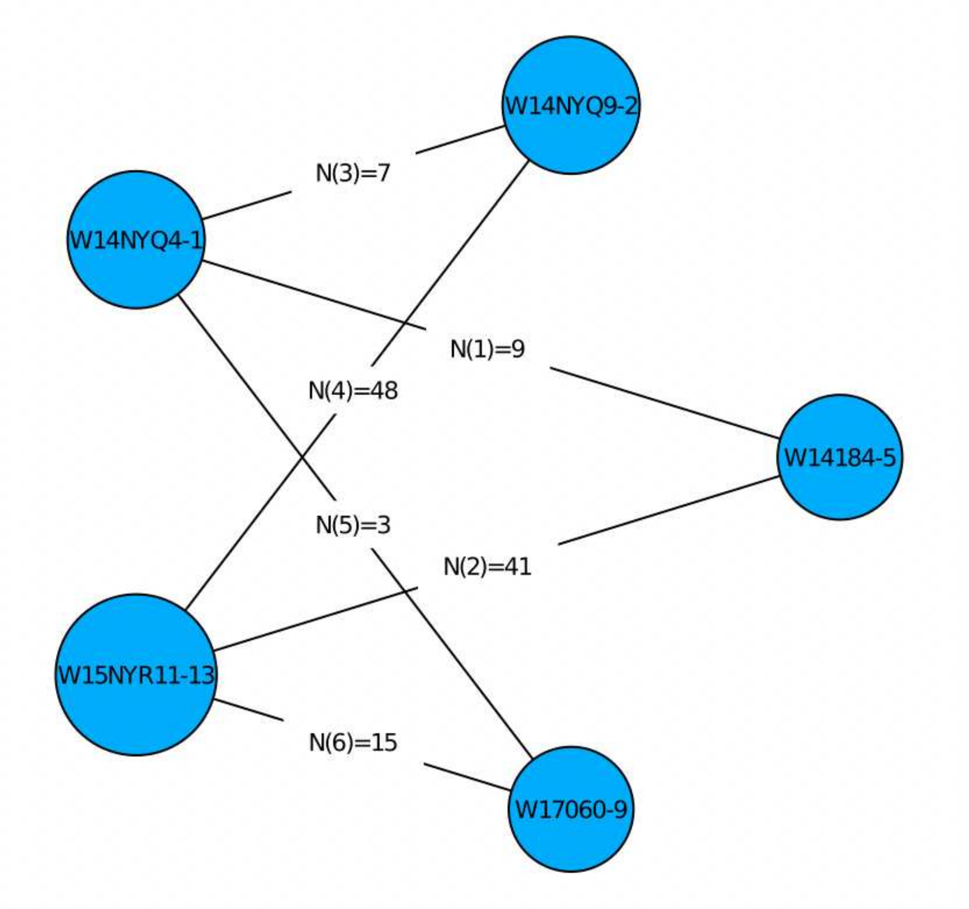


Figure S7. Five-parent partial diallel. Graphical output from PolyOrigin shows the number of progeny per biparental F1 population.

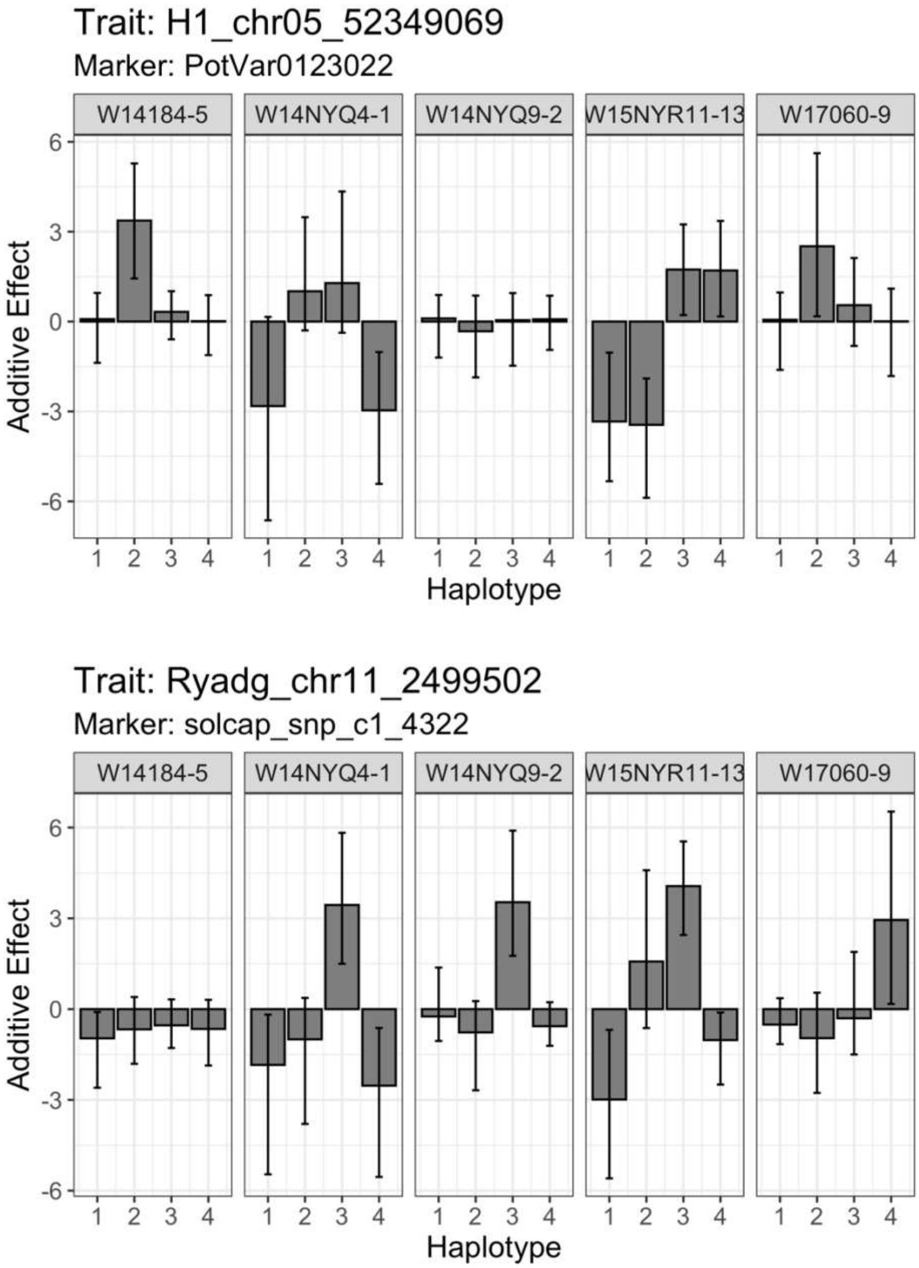


Figure S8. Additive effect estimates for parental haplotypes in the five-parent partial diallel. Positive values indicate presence of the R gene. From left to right, the result indicates the parental dosage of Ryadg is 0, 1, 1. 2, 1, and for H1 the parental dosage is 1, 2, 0, 2, 1. Parents W14NYQ9-2 and W15NYR11-13 were included in the V2 DArTag submission, and their genotype calls agree with these predictions (FileS3).

Table S1. Comparison of KASP and V1 DArTag markers targeting *Ry_{adg}* (snpST00073).

	Absent	**Present**
Absent	7	1
Present	1	84

Table S2. Positive samples for marker Rysto_chr12_2352742 in V2 DArTag.

id	mother	father
W17079-16rus	Payette Russet	AW07791-2rus
W17081-2rus	Payette Russet	W9742-3rus
A12304-1sto	A96953-13sto	Clearwater Russet
W6511-1R	Kankan	W2275-9R

Table S2. Results for marker Sli_chr12_2372490 in V2 DArTag.

id	ALT dosage
W9968-5-DH027	0
CO99076-6R-DH002	0
CO99076-6R-DH033	0
W14NYQ29-5-DH024	0
W14NYQ9-2-DH119	0
W14NYQ9-2-DH132	0
W8890-1R-DH003	0
W9968-5-DH084	0
W9968-5-DH022	0
W13069-5-DH088	0
RioColorado-DH003	0
W9968-5-DH151	0
W12078-76-DH352	0
W12078-76-DH099	0
W13058-4-DH002	0
W14NYQ9-2-DH137	0
W14NYQ9-2-DH146	0
RioColorado-DH005	1
W15263-50R-DH001	1
W15263-50R-DH011	1
W8890-1R-DH002	1
W9426-3R-DH005	1
W9426-3R-DH037	1
W2x150-24	1
W2x113-3	1
W2x001-22-45	2
W2x082-(14/20)-13	2
W2x082-(14/20)-13-2	2