Published March 11, 2016

Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato

Umesh R. Rosyara, Walter S. De Jong, David S. Douches, and Jeffrey B. Endelman*

Abstract

Genome-wide association studies (GWAS) are widely used in diploid species to study complex traits in diversity and breeding populations, but GWAS software tailored to autopolyploids is lacking. The objectives of this research were to (i) develop an R package for autopolyploids based on the **Q** + **K** mixed model, (ii) validate the software with simulated data, and (iii) analyze a diversity panel of tetraploid potatoes. A unique feature of the R package, called GWASpoly, is its ability to model different types of polyploid gene action, including additive, simplex dominant, and duplex dominant. Using a simulated tetraploid population, we confirmed our hypothesis that statistical power is higher when the assumed gene action in the GWAS model matches the gene action at unobserved quantitative trait loci (QTL). Thirteen traits were analyzed in the Solanaceae Coordinated Agricultural Project (SolCAP) potato diversity panel and, consistent with previous studies, significant QTL for tuber shape and eye depth colocalized on chromosome 10. For the other traits, only marginally significant QTL were detected, most likely due to insufficient statistical power: for simulated traits with a heritability (h^2) of 0.3, the median genome-wide power was only 0.01. Our results indicate that both marker density and population size were limiting factors for GWAS with the SolCAP panel.

Published in Plant Genome
Volume 9. doi: 10.3835/plantgenome2015.08.0073

GENOME-WIDE ASSOCIATION STUDIES have become commonplace in diploid plant species as an approach to discovering causal variants. Compared with linkage mapping in biparental crosses, the ability to analyze more diverse germplasm by GWAS promotes the identification of variants with a consistent effect across the discovery population (Myles et al., 2009). Another reason for the intense interest in GWAS is the ability to utilize existing phenotypic and/or genotypic data, which promotes larger population sizes and therefore higher statistical power. Provided sufficient marker density is available, the use of diverse mapping populations with shorter-range linkage disequilibrium also enables finer resolution of QTL positions.

The statistical methods used for GWAS have evolved considerably over the past 15 yr (Balding, 2006; Li et al., 2014). One of the earliest approaches was the transmission disequilibrium test (TDT), which relied on parent-offspring relationships to identify significant associations (Spielman et al., 1993; Allison, 1997). To move beyond familial data, methods were developed to identify subpopulations from the markers and then use these groups as covariates in the analysis. The STRUCTURE (Pritchard et al., 2000) and EIGENSTRAT (Price

U.R. Rosyara and J.B. Endelman, Dep. of Horticulture, Univ. of Wisconsin, Madison, WI 53706; W.S. De Jong, School of Integrative Plant Science, Cornell Univ., Ithaca, NY 14853; D.S. Douches, Dep. of Plant, Soil and Microbial Sciences, Michigan State Univ., East Lansing, MI 48824. Accepted 24 Nov 2015. Received 21 Aug 2015. *Corresponding author (endelman@wisc.edu).

Abbreviations: BIC, Bayesian Information Criterion; DAPC, discriminant analysis of principal components; df, degrees of freedom; GWAS, genome-wide association studies; h^2, broad-sense heritability; LDA, linear discriminant analysis; P3D, population parameters previously determined; PCA, principal component analysis; PEV, prediction error variance; REML, restricted maximum likelihood; SolCAP, Solanaceae Coordinated Agricultural Project; QTL, quantitative trait locus (or loci); SNP, single nucleotide polymorphism.

et al., 2006) programs, which produce matrices typically denoted **Q** and **P**, respectively, are examples of this approach. Yu et al. (2006) demonstrated the value of including a random polygenic effect in the model, with covariance proportional to a marker-estimated kinship (**K**) matrix. As a result of computational innovations (Kang et al., 2008, 2010; Zhang et al., 2010) and the availability of several software packages (Bradbury et al., 2007; Endelman, 2011; Zhou and Stephens, 2012; Lipka et al., 2012), the **Q** + **K** method is now the most widely used single-marker test for GWAS in diploid species.

A few research groups have investigated the use of **Q** + **K** models for GWAS in autopolyploid species, particularly autotetraploid potato (*Solanum tuberosum*, $2n = 4x = 48$). In some cases, diploid models and software have been used because (i) the markers were dominant (Malosetti et al., 2007), or (ii) the markers were codominant but the allele dosage for heterozygotes could not be experimentally determined (Li et al., 2011). Even when tetraploid allele dosage is available, it may be disregarded (i.e., the marker data are "diploidized") to facilitate the use of diploid software (Simko et al., 2006). The first use of tetraploid marker data and association analysis models in potato was in studies of candidate genes (Pajerowska-Mukhtar et al., 2009; Stich and Gebhardt, 2011), and these methods were later extended to true genome-wide analyses (Uitdewilligen et al., 2013). Thus far, however, no one has released polyploid GWAS software targeted to the plant breeding and genetics community.

We report here on the development and validation of R package GWASpoly, which was designed for GWAS with biallelic single nucleotide polymorphisms (SNPs) in autopolyploids using the **Q** (or **P**) + **K** method. One design objective for the software was to incorporate different models of gene action: for ploidy level *N* there are potentially *N* degrees of freedom (df) for the single marker test. Another goal was to investigate different types of kinship (**K**) models for autopolyploids. A number of approaches to modeling kinship have been used in diploids (Bradbury et al., 2007; Zhao et al., 2007; Stich et al., 2008; Kang et al., 2008; Endelman, 2011), but no single method has emerged as clearly superior.

In addition to demonstrating the performance of the software with simulated data, we also present GWAS results for a tetraploid potato diversity panel, which was genotyped and phenotyped as part of the USDA-NIFA SolCAP (Hirsch et al., 2013). The availability of a reasonably priced Infinium array for potato, which originally had 8303 SNPs (Hamilton et al., 2011; Felcher et al., 2012) but has now been extended to more than 12,000, and the ability to reliably call tetraploid dosage for such markers (Voorrips et al., 2011; Hirsch et al., 2013) have created an urgent need for tetraploid GWAS software.

Materials and Methods

Q + K Model for Autotetraploids

The **Q** + **K** linear mixed model for GWAS can be written as (Yu et al., 2006; Kang et al., 2008):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZS}\boldsymbol{\tau} + \mathbf{ZQv} + \mathbf{Zu} + \varepsilon$$

where **y** is a $w \times 1$ vector of observed phenotypes; ε is a $w \times 1$ vector of residuals, with $\mathrm{Var}[\varepsilon] = \mathbf{I}\sigma_e^2$; $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, with incidence matrix **X** used to model environmental effects or other covariate effects; **v** is a $q \times 1$ vector of effects for the subpopulations, with $n \times q$ incidence matrix **Q** for a population of size *n*; and **u** is an $n \times 1$ vector of polygenic effects, with covariance proportional to a kinship (or relationship) matrix, $\mathrm{Var}[\mathbf{u}] = \sigma_g^2\mathbf{K}$. The $w \times n$ incidence matrix **Z** maps genotypes to observations, and the SNP effect is represented by the $d \times 1$ vector $\boldsymbol{\tau}$, where the structure of the $n \times d$ incidence matrix **S** and the dimension *d* depend on the genetic model (see below). The *p*-value for each marker was computed from the *F*-test corresponding to the null hypothesis that the parameters of the SNP effect ($\boldsymbol{\tau}$) are identically zero (for the *F*-statistic formula, see McCulloch and Searle [2001] or Kang et al. [2008]). Two approaches were compared with respect to the estimation of variance components. In the first, σ_g^2 and σ_e^2 were estimated by restricted maximum likelihood (REML) for each marker, using the EMMA algorithm of Kang et al. (2008) implemented in R package rrBLUP (Endelman, 2011). As this approach is computationally demanding, we also estimated the variance components only once per trait, without the SNP effect in the model, which is known as the EMMAX (Kang et al., 2010) or P3D (Zhang et al., 2010) approximation.

For biallelic SNPs in autotetraploids, there are five genotype classes which can be parameterized by the dosage of the minor allele: {0, 1, 2, 3, 4}. The most general type of genetic model allows the fixed effect for each genotype class to be arbitrary. Because it is only the difference between the levels of the fixed effect that matter for the *F*-test , there are 4 df for this model (one less than the number of genotype classes = ploidy level). A number of parameterizations for the general model are possible (Gallais, 2003; Pajerowska-Mukhtar et al., 2009), but for the purpose of the *F*-test, they are statistically equivalent.

In addition to the general model, we present results for four different single-parameter genetic models, which are depicted in Fig. 1: additive, simplex dominant, duplex dominant, and diploidized additive. In the additive model, the SNP effect is proportional to the dosage of the minor allele. In the simplex dominant model, all three heterozygotes are equivalent to one of the homozygotes; as there are two homozygous classes, there are two nonequivalent simplex dominant parameterizations for each marker. There are also two nonequivalent duplex dominant models for each marker, in which the duplex state

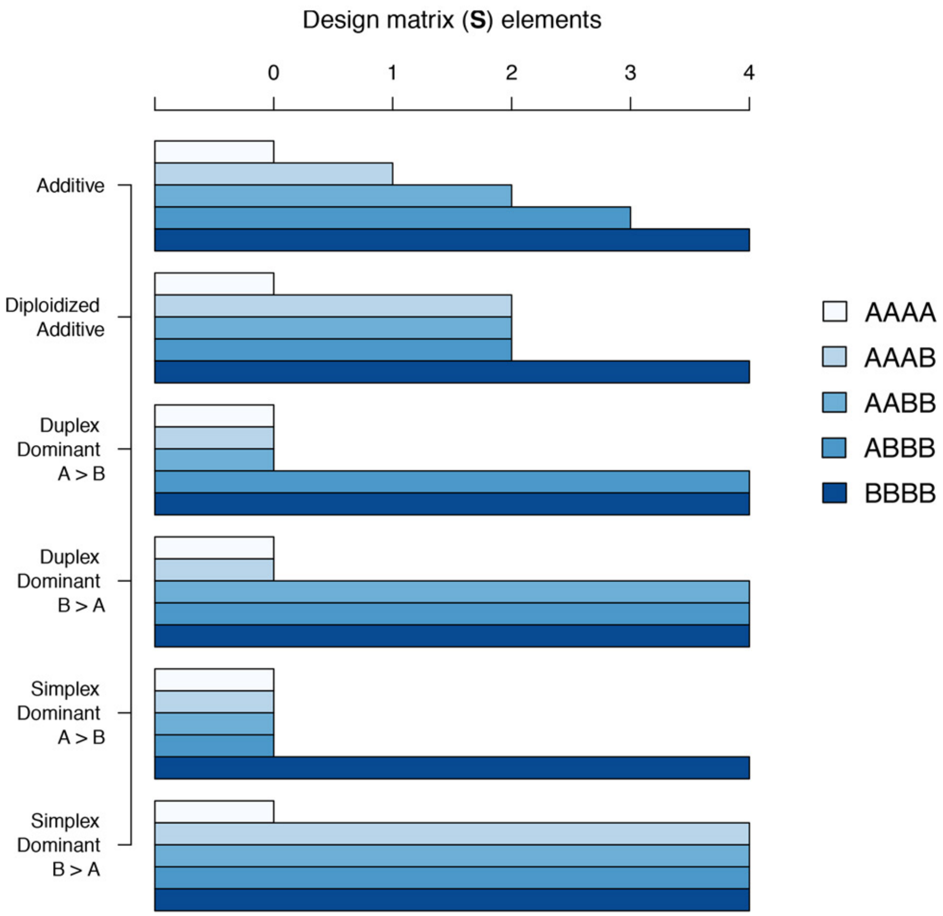

Fig. 1. Graphical depiction of the single nucleotide polymorphim (SNP) effect design matrix elements for tetraploid genetic models. A > B means allele A is dominant over allele B.

(AABB) has a common effect with either the simplex (AAAB) and nulliplex (AAAA) states, or with the triplex (ABBB) and quadriplex (BBBB) states. In the diploidized additive model, all three heterozygous classes are equivalent and exactly halfway between the two homozygotes. We note that, for the simplex dominant and diploidized additive models, the mapping from genotype state to SNP effect is the same regardless of whether the genotypic data are diploidized (called as {0, 1, 2}, failing to differentiate between the heterozygotes) or if tetraploid dosage is called. The additive and duplex dominant models require tetraploid genotype data.

Three tetraploid kinship (or relationship) models were compared. The first is the canonical relationship matrix used in genome-wide prediction studies (VanRaden, 2008), which we call the realized relationship model:

$$\mathbf{K} = \mathbf{M}\mathbf{M}^{\mathrm{T}} \qquad \text{(Realized Relationship)}$$

In the above equation, **M** is the $n \times m$ genotype matrix for a population of size n with m markers, where the genotypes (M_{ij}) have been "centered" by subtracting the population mean for each marker (Endelman and Jannink, 2012). The scaling of the realized relationship matrix is not shown in the above equation, as it is irrelevant for GWAS, but in the software the matrix is scaled such that the mean of the diagonal elements is 1. The second approach is based on the concept of a molecular similarity index (Oliehoek et al., 2006). If a, b, c, d denote the four homologs at locus z in one individual, and e, f, g, h denote the homologs in a second individual, the similarity between the two individuals at that locus is

$$K_z = \frac{1}{16}\sum_{x\in\{a,b,c,d\}}\;\sum_{y\in\{e,f,g,h\}} I_{xy} \qquad \text{(Molecular Similarity)}$$

where the indicator function I_{xy} equals one when homologs x and y have the same allele, and is zero otherwise. The average similarity between the two individuals across m loci is $K = m^{-1}\sum_z K_z$. Whereas the first two models may be considered additive models of relationship, the third model—the Gaussian kernel—involves multigenic interactions (Gianola and van Kaam, 2008; Piepho, 2009). Its formula is

$$K_{ij} = \exp[-(D_{ij}/\theta)^2] \qquad \text{(Gaussian Kernel)}$$

where D_{ij} is the Euclidean distance, normalized to the interval [0,1]:

$$D_{ij}^2 = (16m)^{-1}\sum_k (M_{ik} - M_{jk})^2 \ .$$

The value for the scale parameter θ, which determines how quickly kinship decays with genetic distance, was determined by REML as described in Endelman (2011).

A key diagnostic for GWAS is a quantile-quantile plot of the observed vs. expected –log p values, which should follow a uniform distribution under the null hypothesis. The inflation of p-values above the $y = x$ line in such a plot is an indicator of the failure of the model to control for population structure. Inflation was quantified by the linear regression coefficient of the observed vs. expected $-\log_{10}$ p-values, denoted λ, which has a value of 1 under the null hypothesis (Riedelsheimer et al., 2012).

The average inflation across different GWAS models and traits was compared by analysis of variance, according to

$$\lambda_{ij} = \mu + t_i + \beta_j + \varepsilon_{ij}$$

where t_i is the effect for trait i, and β_j is the effect for model j. The naïve GWAS model was not included in the analysis as it produced residuals with much larger variance (thereby violating an assumption of ANOVA). R package *lsmeans* was used to make means comparisons, with p-values adjusted for multiple testing by Tukey's method.

Three different methods are available in GWASpoly for establishing a p-value detection threshold for statistical significance. The first is the Bonferroni correction, which uses a threshold of α/m to ensure the genome-wide type I error with m markers is no greater than α. The second approach is the random permutation test, in which phenotypes are randomly permuted to explicitly construct the genome-wide null distribution of p-values (Churchill and Doerge, 1994). The third option uses the *qvalue* package (Storey and Tibshirani, 2003) to control the genome-wide false discovery rate (rather than Type I error = probability of false positive). For the simulations, due to their computationally intensive nature, we used the Bonferroni correction with $\alpha = 0.05$. For the analysis of the real potato data, we used the permutation test with 1000 permutations and genome-wide $\alpha = 0.05$.

Simulated Populations

Simulated populations and phenotypes were used to validate the software. Random mating autotetraploid populations were simulated using the software PedigreeSim (Voorrips and Maliepaard, 2012), according to the scheme illustrated in Supplemental Fig. S1. The base population consisted of five individuals, from which 10 mating pairs were randomly selected, and 10 progeny per pair were randomly generated to create a population of 100 individuals in Generation 1. In Generations 2 through 999, 100 mating pairs were randomly selected, each contributing 1 offspring, to keep the population size constant at 100. For the last (1000th) generation, N mating pairs were randomly selected, each contributing one offspring to create a population of size N. Results are shown for $N = 200$, 400, and 600. The simulated genome contained three chromosomes, each 100 cM in length, with 100 loci per cM. Recombination was simulated according to Haldane's mapping function, using the default meiosis parameters governing the formation of quadrivalents. Marker density was varied by subsampling loci ($m = 3$, 10, 50 per centiMorgan).

To estimate power in each simulated population, one marker was randomly designated as the causal QTL and the remaining markers were converted to biallelic SNPs by randomly assigning the 20 founder alleles to biallelic states (A/B), thereby creating markers with an average minor allele frequency of 0.5. Two different schemes were used to simulate genotypic values. In the first, the causal QTL was also converted to a biallelic locus as above, and allelic effects were sampled from the standard normal distribution. This scheme was used to generate Tables 1 and 2. In the second scheme, which was used for Fig. 2, each of the 20 founder alleles was assigned a different effect, drawn from the standard normal distribution. The phenotypic value for each genotype was the sum of its genotypic value and a random deviate, with variance chosen such that the ratio between the genetic and phenotypic variances of the population was $h^2 = 0.3$. Because there were no subpopulations in the simulated population, we used a **K**-only GWAS model with the realized relationship matrix. A QTL was considered detected if a SNP within 5 cM of the unobserved QTL had –log p-value above the significance threshold. Conversely, significant markers greater than 5 cM from the QTL were considered false positives. We report the average power and false positive rate based on 1000 replications, with standard errors computed from the binomial distribution.

Potato Diversity Panel

The genotypic and phenotypic data were collected as part of the SolCAP. The SolCAP potato diversity panel consists of both diploid and tetraploid wild species, genetic stocks, and cultivated potato lines with release dates ranging from 1857 to 2011 (Hirsch et al., 2013). The panel was genotyped with an Infinium SNP array of 8303 markers (Hamilton et al., 2011; Felcher et al., 2012), and tetraploid marker dosage was determined by Hirsch et al. (2013), principally by visual inspection of the cluster boundaries. Our analysis of population structure was conducted using all 221 tetraploid lines in the panel (Supplemental Table S1), while GWAS results are based on the 187 tetraploid lines with both marker and phenotypic data.

Broad-sense heritability and GWAS results are presented for thirteen quantitative traits, which were measured in up to four environments (New York-2010, Wisconsin-2010, New York-2011, Wisconsin-2011). A randomized, complete block design with two replicates was used in each environment, although not all traits were measured in every environment (the number of environments per trait is shown in Table 4). In addition to the four traits analyzed by Hirsch et al. (2013), which were chip color (1–5 scale), tuber shape (1–5 scale), tuber sucrose and glucose (milligrams gram^{-1} fresh wt.), we present GWAS results for total yield (kilograms), tuber size and eye depth (1–9 visual scale), vine maturity 95 and

Table 1. Comparison of the full mixed model (variance components estimated for each marker) vs. the P3D approximation (variance components estimated once) on the statistical power and false positive rate (FPR) in a simulated autotetraploid population with 400 individuals, 600 markers per 100 cM chromosome, and $h^2 = 0.3$. The standard errors for power and FPR were less than 0.015 and 0.008, respectively.

Gene action	Full model	P3D model
Additive	0.93 (FPR 0.03)	0.90 (FPR 0.00)
Simplex dominant	0.37 (FPR 0.06)	0.36 (FPR 0.03)
Duplex dominant	0.89 (FPR 0.04)	0.84 (FPR 0.00)

Table 2. Effect of true vs. assumed (model) gene action on the statistical power and false positive rate (FPR) in a simulated autotetraploid population with 400 individuals, 600 markers per 100 cM chromosome, and $h^2 = 0.3$. The standard errors for power and FPR were less than 0.008.

	True gene action		
Model	Additive	Simplex dominant	Duplex dominant
Additive	0.94 (FPR 0.01)	0.39 (FPR 0.02)	0.78 (FPR 0.02)
Simplex dominant	0.48 (FPR 0.03)	0.67 (FPR 0.03)	0.11 (FPR 0.04)
Duplex dominant	0.68 (FPR 0.03)	0.21 (FPR 0.07)	0.86 (FPR 0.03)
Diploidized additive	0.75 (FPR 0.02)	0.30 (FPR 0.02)	0.06 (FPR 0.03)
General	0.42 (FPR 0.06)	0.11 (FPR 0.04)	0.13 (FPR 0.06)

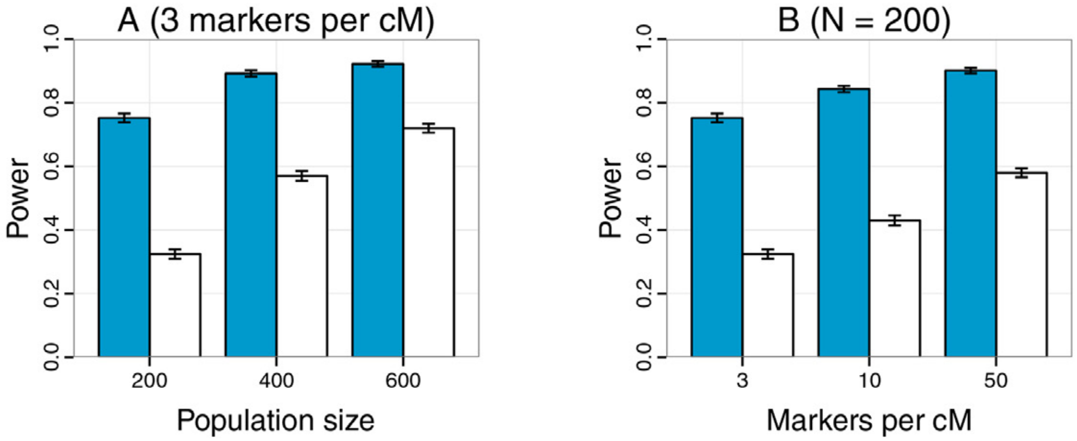

Fig. 2. Determinants of statistical power in simulated autotetraploid populations with biallelic markers. Panel A shows the effect of population size and panel B the effect of marker density. The solid bars correspond to biallelic quantitative trait loci (QTL), while the open bars are for multiallelic QTL. Error bars show ± 1 standard error.

120 d after planting (1–9 visual scale), tuber length (millimeters), tuber width (millimeters), tuber fructose and malic acid content (milligrams gram^{-1} fresh wt.). Phenotypic data were analyzed with the following linear model

$$y_{ijk} = \mu + G_i + E_j + b(E)_{jk} + GE_{ij} + \varepsilon_{ijk}$$

where y_{ijk} is the observation for genotype i in block k of environment j. Variance components were estimated by REML using R package *lme4* (R Development Core Team, 2014). The residuals appeared to be normally distributed for all traits except fructose and glucose, for which a log transformation was used to satisfy model assumptions. Because the experimental design was unbalanced for several traits, the reliability, or heritability, of each genotype was estimated from the prediction error variance (PEV) of the BLUP solution for G_i (Clark et al., 2012): $h^2 = 1 - \text{PEV}/V_G$. For each trait, we report the average heritability for the population. To generate phenotypic values for GWAS, G_i was modeled as a fixed effect (all other effects were random), and the best linear unbiased estimator (BLUE) was computed with *lme4* (Supplemental Table S2).

Three different population structure matrices (**Q**) were compared. The first corresponds to the four subpopulations identified with the program STRUCTURE (Pritchard et al., 2000), as reported by Hirsch et al. (2013). The second matrix was constructed from a principal component analysis (PCA), using centered and scaled marker scores (Price et al., 2006). Since a scree plot of the cumulative percent variation vs. model complexity (Supplemental Fig. S2) showed a gradual increase and no obvious choice for a low-dimensional model, we used four principal components to be consistent with the four covariates used with the other **Q** models. The third matrix was based on the discriminant analysis of principal components (DAPC) method in R package *adegenet* (Jombart et al., 2010). Since DAPC is less widely used than PCA or STRUCTURE, we describe it in more detail. In the first step, *k*-means clustering was used to identify groups. The value $k = 4$ minimized the Bayesian Information Criterion (BIC) and was thus used for GWAS (group membership probabilities in Supplemental Table S2). However, for the purpose of discussing population structure we selected $k = 6$, which was still within the shallow minimum of the BIC curve (Supplemental Fig. S3). In the second step of the DAPC method, linear discriminants were computed based on a reduced-rank representation of the marker matrix (Jombart et al., 2010). Unlike PCA, which maximizes the total variation in the dataset, linear discriminants maximize the ratio of the between-group to within-group sum-of-squares. A cross-validation study revealed that the classification error by linear discriminant analysis (LDA) was minimized over a range of model complexities (Supplemental Fig. S4); we selected 60 principle components for LDA at the upper end of the range.

For each trait, four GWA analyses were conducted, based on the additive, simplex dominant, duplex dominant, and the general SNP models. When multiple significant markers were detected within a 10 Mb region, only the most significant (i.e., lowest *p*-value) was reported, along with the corresponding SNP model.

Statistical power was estimated for the SolCAP panel genotypes using a similar method as for the simulated populations. An additive QTL with $h^2 = 0.3$ was simulated at each marker, which was considered detected if any marker up to 2.5 Mb from the QTL exceeded the detection threshold of $\alpha = 0.05/3242$ (i.e., the Bonferroni correction for a genome-wide scan). Extending the detection interval up to 5 Mb from the QTL did not change the median power for the genome (Supplemental Table S3). The average power for each QTL was based on 1000 simulations.

Results and Discussion

Validation with Simulated Data

The GWAS software was validated using simulated phenotypes and genotypes from a random mating autotetraploid population (details in Methods). Our first objective was to determine the quality of the P3D approximation for the mixed model (Zhang et al., 2010; Kang et al., 2010), which is widely used in diploid GWAS to reduce the computing time. The P3D approximation involves estimating the variance components only once by REML, and then using those values for each single-marker hypothesis test. Table 1 compares the statistical power and false positive rate of the full mixed model vs. the P3D model for three different types of simulated QTL: additive, simplex dominant, and duplex dominant (see Methods for more information on these models). Using the same *p*-value detection threshold for both methods, we observed slightly lower statistical power (0.01–0.05) when using the P3D model but also fewer false positives. If the $-\log_{10} p$ threshold for the P3D model were lowered to achieve the same false positive rate for the two methods, the difference in statistical power would be even smaller. For this relatively small dataset of 400 individuals and 1800 markers (600 for each of three linkage groups), the P3D approximation reduced the computing time by a factor of 20. Thus, given its favorable performance, the P3D approach was used for the remainder of the study.

One of the unique features of the software is its ability to conduct the single marker test for association using different models of gene action. Our hypothesis was that the probability of detecting QTL would be higher if the marker model matched the gene action at unobserved QTL. The results shown in Table 2 confirm this hypothesis: for an additive QTL, analysis with an additive model resulted in a statistical power of 0.94, while the next most powerful model detected the QTL with probability 0.75 (standard errors < 0.01). For a simplex dominant QTL, use of the simplex model in the analysis increased power by 0.28 over the next best model (additive). Table 2 also illustrates the consequences of neglecting dosage information for the heterozygous genotypes, that is, diploidizing the data. If the underlying QTL is simplex dominant, there is no loss of power with diploidized marker data, as the simplex dominant model implies all heterozygous genotypes are equivalent. However, when the QTL was additive or duplex dominant, the best diploid model had significantly less power than the best tetraploid one (losses of 0.19 and 0.67, respectively). Table 2 also shows the potential disadvantage of relying solely on the general tetraploid model, which makes no assumptions about gene action, and thus encompasses the other models. This flexibility comes with a penalty of substantially lower statistical power (more than 0.5 less than the best model) because 4 df are needed for the single marker test. This conclusion still holds when the general model is compared against a combination of multiple single-df models with a higher detection threshold to maintain the same false positive rate (data not shown).

Our third objective was to investigate the effects of marker density and population size on statistical power in autotetraploid GWAS. In diploids it is well established that both factors contribute to higher power (Klein, 2007; Spencer et al., 2009), and this trend was also observed in simulated autotetraploid populations (Fig. 2). The left-most bars in Panels A and B of Fig. 2 correspond to a common scenario of 300 markers per 100 cM chromosome and 200 individuals, which is approximately the size of the real potato dataset analyzed below. Figure 2A shows the effect of increasing population size, while Fig. 2B illustrates higher maker density. For the same proportional increase (e.g., twofold), population size had a bigger effect on power than marker density. The two different series in Fig. 2 (solid vs. open) correspond to different types of QTL models. In both cases the markers are biallelic, but the solid bars correspond to biallelic QTL while the open bars are multiallelic QTL. The loss in power for the latter scenario can be viewed as analogous to the loss in power for the off-diagonal elements in Table 2. In both cases there is a mismatch between the markers in the GWAS model and gene action at the unobserved QTL. This mismatch can potentially be overcome through the use of multimarker haplotypes in GWAS (Lorenz et al., 2010).

GWAS of a Tetraploid Potato Diversity Panel

The SolCAP potato diversity panel included 221 tetraploid lines and 3441 tetraploid SNP markers with minor allele frequency greater than 0.05. Based on version 4.03 of the potato reference genome (Potato Genome

Table 3. Comparison between *k*-means clustering groups (Roman numerals) and market categories.

Market category	I	II	III	IV	V	VI	Total
Fry processing	26	0	1	2	3	2	34
Table Russet	10	0	0	0	3	0	13
Wild species	0	3	0	0	0	0	3
Genetic stock	0	0	0	4	1	0	5
Pigmented	0	1	28	3	0	0	32
Yellow	1	0	3	15	3	5	27
Round White table	2	2	2	8	15	9	38
Chip processing	0	1	0	5	29	34	69
Total	39	7	34	37	54	50	221

Sequencing Consortium, 2011; Sharma et al., 2013), the median distance between markers was 67 kb, with a minimum of 3 bp and maximum of 8.2 Mb.

The diversity panel was comprised of potatoes from eight different market categories, listed in Table 3. Previously, the program STRUCTURE had been used to identify subpopulations in this dataset (Hirsch et al., 2013). A commonly used alternative to STRUCTURE for GWAS is PCA. Figure 3B shows the projection of the population onto the first two principal components, which only account for 8% of the total variation in the marker data (scree plot in Supplemental Fig. S2).

To achieve better separation between subpopulations, we used the DAPC technique (Jombart et al., 2010). In the first step, clusters based on the marker data were compared against the market categories. Table 3 and Fig. 3A show the results for $k = 6$ clusters. As expected, the DAPC technique produced greater separation among groups than PCA, with Groups I–III clearly separated, and Groups IV–VI apparently more closely related. Group I primarily contains the fry processing and table russets, which are closely related and continue to be intermated by breeders. Group II was a small group containing the majority of the wild species in the panel, and Group III contained most of the pigmented (red and purple) types. Group IV contained the majority of the yellow potatoes, along with some round white potatoes for both tablestock and chip processing. Groups V and VI were predominantly round white potatoes, used for both tablestock and chip processing. Hirsch et al. (2013) also observed minor divergence within the round white types based on hierarchical clustering.

One of the hallmarks of using ordinary linear regression (aka, the naïve model) as a test of association in structured populations is the inflation of the –log(p) values relative to the expected value under the null hypothesis (Supplemental Fig. S5; Freedman et al., 2004; Clayton et al., 2005). The use of subpopulation group membership as a covariate in the analysis helps to reduce this inflation. Each boxplot in Fig. 4 shows the distribution of inflation factors across the 13 traits analyzed in this study. All three of the **Q** models tested—DAPC, PCA, and STRUCTURE—were able to reduce inflation relative to the naïve model, with DAPC and PCA slightly better than STRUCTURE. The DAPC approach ($\mathbf{Q}_{\mathrm{DAPC}}$) was selected for subsequent analyses.

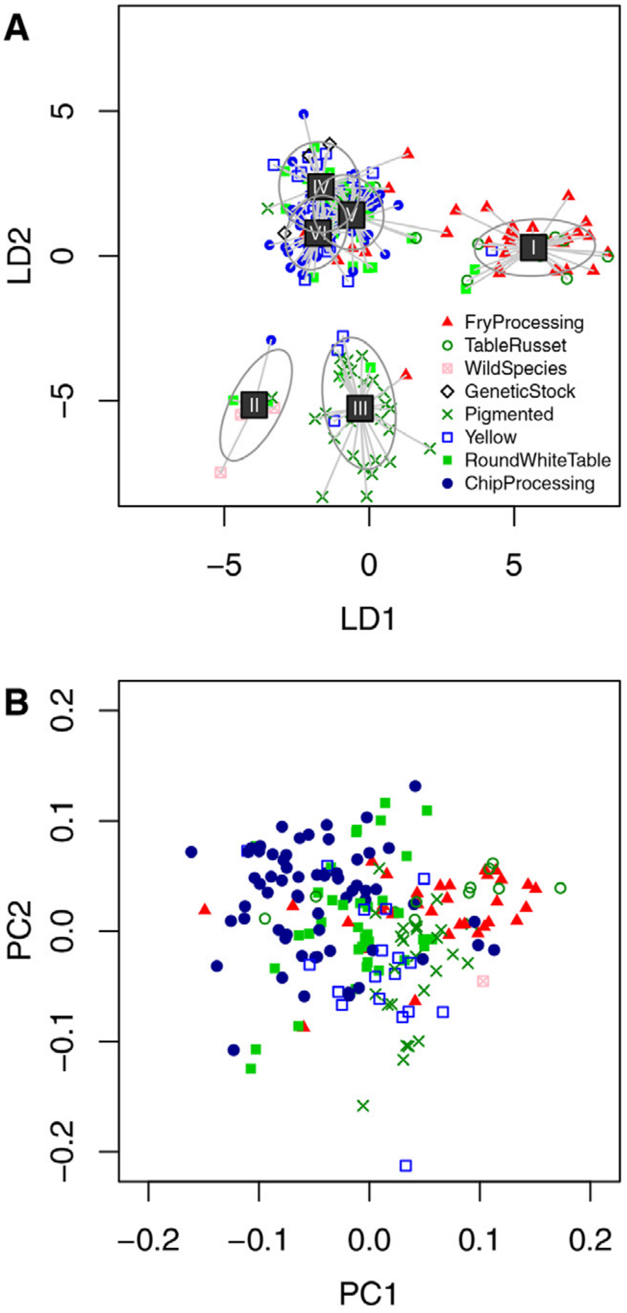

Fig. 3. Projection of the potato diversity panel onto (A) the first two linear discriminants (LD) vs. (B) the first two principal components (PC). See Table 3 for the composition of the six clusters (I–VI) with respect to potato market types.

As first shown by Yu et al. (2006), a random polygenic effect with covariance proportional to a kinship matrix **K** can also reduce inflation. The results in Fig. 4 show that all three kinship matrices we tested were effective, with perhaps a slight advantage to the realized relationship model ($\mathbf{K}_{\mathrm{RR}} = \mathbf{MM}^{\mathrm{T}}$ for genotype matrix **M**), which had significantly less inflation than any of the **Q** models. Subsequent analyses were conducted using the $\mathbf{Q}_{\mathrm{DAPC}} + \mathbf{K}_{\mathrm{RR}}$ model.

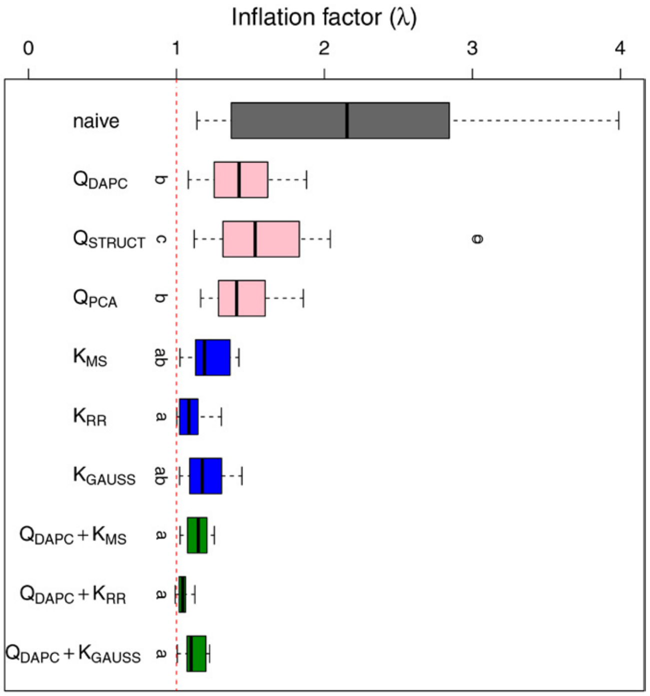

Fig. 4. Influence of the association analysis model on *p*-value inflation for the 13 traits in the potato diversity panel. The inflation factor (λ) is the regression coefficient from a quantile-quantile plot of the $-\log_{10}$p-values, which should equal 1 under the null hypothesis. **Q** refers to the incidence matrix for subpopulation covariates modeled as a fixed effect, while **K** is the kinship matrix for the random polygenic effect. Kinship model abbreviations: DAPC, discriminant analysis of principal components; GAUSS = Gaussian kernel; MS, molecular similarity index; PCA, principal components analysis; STRUCT, STRUCTURE software; RR = realized relationship model.

Table 4. Broad-sense heritability (h^2) and significant quantitative trait loci (QTL) in the potato diversity panel.

Trait	No. environments	h^2	Significant QTL	Model†
Total yield	4	0.73	c2_10614 (chr 4 at 71827521, $-\log_{10}p = 4.8$)	DD
Chip color	4	0.91	none	
Eye depth	4	0.74	c2_11685 (chr 5 at 2288291, $-\log_{10}p = 4.7$)	AD
			c1_8019 (chr10 at 48863165, $-\log_{10}p = 7.2$)	AD
Tuber shape	4	0.94	c1_8019 (chr10 at 48863165, $-\log_{10}p = 9.5$)	AD
Tuber size	2	0.81	none	
Tuber length	2	0.91	c1_8019 (chr10 at 48863165, $-\log_{10}p = 6.5$)	AD
Tuber width	2	0.87	none	
Sucrose	2	0.67	none	
Glucose	2	0.78	none	
Fructose	2	0.85	none	
Malic acid	2	0.60	none	
Vine maturity at 95 d	3	0.69	c2_34548 (chr1 at 84727196, $-\log_{10}p = 4.5$)	DD
			c2_13133 (chr9 at 8245062, $-\log_{10}p = 5.2$)	SD
			c1_9183 (chr11 at 42627957, $-\log_{10}p = 4.7$)	AD
Vine maturity at 120 d	2	0.80	c2_25219 (chr7 at 47348171, $-\log_{10}p = 5.0$)	GEN

† Model with the most significant marker is listed. AD = additive, SD = simplex dominant, DD = duplex dominant, GEN = general.

Several categories of traits were measured on the diversity panel, including agronomic (e.g., total yield, vine maturity), morphological (e.g., tuber shape and eye depth), and biochemical (e.g., tuber sucrose and glucose) properties. Table 4 presents the broad-sense heritability on an entry-mean basis for each trait, which ranged from 0.60 for tuber malic acid content to 0.94 for tuber shape.

Significant QTL were detected for 6 the 13 traits, although many of the QTL were only marginally significant (Table 4; results for all markers in Supplemental Table S4). Significant QTL were not detected for the three tuber sugar traits (sucrose, glucose, and fructose) even though they had heritability comparable to the other traits. This was unexpected as metabolic traits typically have fewer causal loci with larger (and thus more easily detectable) effects compared with a complex trait such as yield (Riedelsheimer et al., 2012). The most significant QTL were for tuber shape and tuber eye depth, both at 48.9 Mb on chromosome 10 (Supplemental Fig. S6). Several biparental linkage mapping studies have mapped major QTL for these traits to the same region (Van Eck et al., 1994; Śliwka et al., 2008; Li et al., 2005; Prashar et al., 2014), although the molecular identities of the QTL have not yet been published. QTL studies in potato frequently detect a major locus affecting plant maturity on chromosome 5 (Bradshaw et al., 2008), which was identified as the *StCDF1* gene by Kloosterman et al. (2013). This locus was not detected in our analysis of the SolCAP plant maturity data, although minor QTL were identified on chromosomes 1, 7, 9, and 11.

To better understand the scarcity of major QTL in the GWAS results for the SolCAP panel, a power simulation was performed using simulated QTL and phenotypes, but with the actual marker data. For a monogenic trait with $h^2 = 0.3$, the genome-wide median for the probability of QTL detection was only 0.01 (results for all loci in Supplemental Table S3). Although low power was expected considering the small size of the population ($N = 187$), this result was even lower than anticipated. To determine if marker density also played a role, the power was plotted against the distance between the QTL and its closest marker (Fig. 5). The red trendline in Fig. 5, which is the 95th percentile, shows that power was lower in regions of lower marker density. We conclude that, in addition to increasing the population size, higher marker density could also improve future GWAS studies in potato.

The GWASpoly software is being distributed under the GNU Public License and can be downloaded from http://potatobreeding.cals.wisc.edu/software (verified 29 Jan. 2016).

Author Contributions

Designed the research: JBE. Contributed phenotypic data: WSD, DSD. Developed the software and analyzed the data: URR, JBE. Wrote the manuscript: URR, JBE.

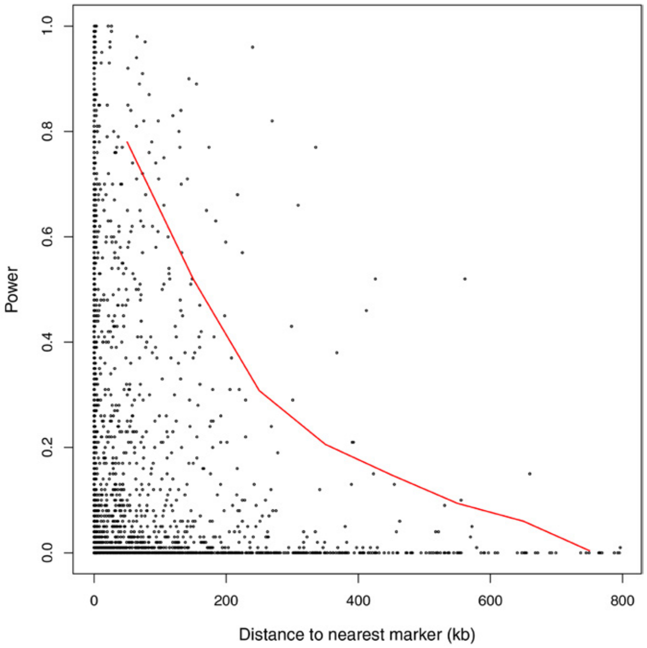

Fig. 5. Influence of marker density on statistical power for the SolCAP panel. The power is the average probability, based on 1000 simulations, of detecting a monogenic trait with $h^2 = 0.3$. The red trendline is the 95th percentile.

Supplemental Information Available

Supplemental information is included with this article.

Supplemental Figures S1–S6.

Supplemental Table S1. Marker data for the potato diversity panel.

Supplemental Table S2. Phenotype data for the potato diversity panel.

Supplemental Table S3. Statistical power for the potato diversity panel.

Supplemental Table S4. GWAS results for the potato diversity panel.

Acknowledgments

Financial support was provided to U.R.R. by USDA-NIFA-SCRI Grant No. 2011-51181-30629 (Improved Breeding and Variety Evaluation Methods to Reduce Acrylamide Content and Increase Quality in Processed Potato Products) and to J.B.E. by USDA-NIFA-Hatch Accession No. 1002731 (Genome-wide Association Analysis and Breeding in Potato). Collection of the phenotype and marker data was supported by USDA-NIFA-AFRI Grant No. 2009-85606-05673 (Translating Solanaceae Sequence Diversity and Trait Variation into Applied Outcomes through Integrative Research, Education, and Extension). We thank Paul Bethke and Shelley Jansky for contributing phenotypic data.

References

Allison, D.B. 1997. Transmission-disequilibrium tests for quantitative traits. Am. J. Hum. Genet. 60:676–690.

Balding, D.J. 2006. A tutorial on statistical methods for population association studies. Nat. Rev. Genet. 7:781–791. doi:10.1038/nrg1916

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinform. 23:2633–2635. doi:10.1093/bioinformatics/btm308

Bradshaw, J.E., C.A. Hackett, B. Pande, R. Waugh, and G.J. Bryan. 2008. QTL mapping of yield, agronomic and quality traits in tetraploid potato (*Solanum tuberosum* subsp. *tuberosum*). Theor. Appl. Genet. 116:193–211. doi:10.1007/s00122-007-0659-1

Churchill, G.A., and R.W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963–971.

Clark, S.A., J.M. Hickey, H.D. Daetwyler, and J.H.J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44:4. doi:10.1186/1297-9686-44-4

Clayton, D.G., N.M. Walker, D.J. Smyth, R. Pask, J.D. Cooper, L.M. Maier, et al. 2005. Population structure, differential bias and genomic control in a large-scale, case–control association study. Nat. Genet. 37:1243–1246. doi:10.1038/ng1653

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Gen. 4:250–255. doi:10.3835/plantgenome2011.08.0024

Endelman, J.B., and J.-L. Jannink. 2012. Shrinkage estimation of the realized relationship matrix. G3: Genes, Genomes, Genet. 2:1405–1413. doi:10.1534/g3.112.004259

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

Freedman, M.L., D. Reich, K.L. Penney, G.J. McDonald, A.A. Mignault, N. Patterson, et al. 2004. Assessing the impact of population stratification on genetic association studies. Nat. Genet. 36:388–393. doi:10.1038/ng1333

Gallais, A. 2003. Quantitative genetics and breeding methods in autopolyploid plants. INRA, Paris.

Gianola, D., and J.B.C.H.M. van Kaam. 2008. Reproducing Kernel Hilbert Spaces Regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303. doi:10.1534/genetics.107.084285

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

Hirsch, C.N., C.D. Hirsch, K. Felcher, J. Coombs, D. Zarka, A. Van Deynze, et al. 2013. Retrospective view of North American potato (*Solanum tuberosum* L.) breeding in the 20th and 21st centuries. G3: Genes, Genomes, Genet. 3:1003–1013. doi:10.1534/g3.113.005595

Jombart, T., S. Devillard, and F. Balloux. 2010. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. BMC Genet. 11:94. doi:10.1186/1471-2156-11-94

Kang, H.M., J.H. Sul, S.K. Service, N.A. Zaitlen, S. Kong, N.B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42:348–354. doi:10.1038/ng.548

Kang, H.M., N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723. doi:10.1534/genetics.107.080101

Klein, R.J. 2007. Power analysis for genome-wide association studies. BMC Genet. 8:58. doi:10.1186/1471-2156-8-58

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

Li, M., X. Liu, P. Bradbury, J. Yu, Y.-M. Zhang, R.J. Todhunter, E.S. Buckler, and Z. Zhang. 2014. Enrichment of statistical power for genome-wide association studies. BMC Biol. 12:73. doi:10.1186/s12915-014-0073-5

Li, X.Q., H. De Jong, D.M. De Jong, and W.S. De Jong. 2005. Inheritance and genetic mapping of tuber eye depth in cultivated diploid potatoes. Theor. Appl. Genet. 110:1068–1073. doi:10.1007/s00122-005-1927-6

Li, X., Y. Wei, K.J. Moore, R. Michaud, D.R. Viands, J.L. Hansen, A. Acharya, and E.C. Brummer. 2011. Association mapping of biomass yield and stem composition in a tetraploid alfalfa breeding population. Plant Gen. 4:24–35. doi:10.3835/plantgenome2010.09.0022

Lipka, A.E., F. Tian, Q. Wang, J. Peiffer, M. Li, P.J. Bradbury, M.A. Gore, E.S. Buckler, and Z. Zhang. 2012. GAPIT: Genome association and prediction integrated tool. Bioinform. 28:2397–2399. doi:10.1093/bioinformatics/bts444

Lorenz, A.J., M.T. Hamblin, and J.-L. Jannink. 2010. Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. PLoS ONE 5(11):E14079. doi:10.1371/journal.pone.0014079

Malosetti, M., C.G. van der Linden, B. Vosman, and F. van Eeuwijk. 2007. A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. Genetics 175:879–889. doi:10.1534/genetics.105.054932

McCulloch, C.E., and S.R. Searle. 2001. Generalized, linear, and mixed models. John Wiley & Sons, New York.

Myles, S., J. Peiffer, P.J. Brown, E.S. Ersoz, Z. Zhang, D.E. Costich, and E.S. Buckler. 2009. Association mapping: Critical considerations shift from genotyping to experimental design. Plant Cell 21:2194–2202. doi:10.1105/tpc.109.068437

Oliehoek, P.A., J.J. Windig, J.A. van Arendonk, and P. Bijma. 2006. Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. Genetics 173:483–496. doi:10.1534/genetics.105.049940

Pajerowska-Mukhtar, K., B. Stich, U. Achenbach, A. Ballvora, J. Lubeck, J. Strahwald, E. Tacke, H.R. Hofferbert, E. Ilarionova, D. Bellin, B. Walkemeier, R. Basekow, B. Kersten, and C. Gebhardt. 2009. Single nucleotide polymorphisms in the *allene oxide synthase 2* gene are associated with field resistance to late blight in populations of tetraploid potato cultivars. Genetics 181:1115–1127. doi:10.1534/genetics.108.094268

Piepho, H.P. 2009. Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49:1165–1176. doi:10.2135/cropsci2008.10.0595

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

Prashar, A., C. Hornyik, V. Young, K. McLean, S.K. Sharma, M. F B. Dale, and G.J. Bryan. 2014. Construction of a dense SNP map of a highly heterozygous diploid potato population and QTL analysis of tuber shape and eye depth. Theor. Appl. Genet. 127:2159–2171. doi:10.1007/s00122-014-2369-9

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909. doi:10.1038/ng1847

Pritchard, J.K., P. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

R Development Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Riedelsheimer, C., J. Lisec, A. Czedik-Eysenberg, R. Sulpice, A. Flis, C. Grieder, T. Altmann, M. Stitt, L. Willmitzer, and A.E. Melchinger. 2012. Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc. Natl. Acad. Sci. USA 109:8872–8877. doi:10.1073/pnas.1120813109

Sharma, S.K., D. Bolser, J. de Boer, M. Sønderkaer, W. Amoros, M.F. Carboni, 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

Simko, I., K.G. Haynes, and R.W. Jones. 2006. Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics 173:2237–2245. doi:10.1534/genetics.106.060905

Śliwka, J., I. Wasilewicz-Flis, H. Jakuczun, and C. Gebhardt. 2008. Tagging quantitative trait loci for dormancy, tuber shape, regularity of tuber shape, eye depth and flesh colour in diploid potato originated from six Solanum species. Plant Breed. 127:49–55. doi:10.1111/j.1439-0523.2008.01420.x

Spencer, C.C., Z. Su, P. Donnelly, and J. Marchini. 2009. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 5(5):E1000477. doi:10.1371/journal.pgen.1000477

Spielman, R.S., R.E. McGinnis, and W.J. Ewens. 1993. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. 52:506–516.

Stich, B., and C. Gebhardt. 2011. Detection of epistatic interactions in association mapping populations: An example from tetraploid potato. Heredity 107:537–547. doi:10.1038/hdy.2011.40

Stich, B., J. Mohring, H.-P. Piepho, M. Heckenberger, E.S. Buckler, and A.E. Melchinger. 2008. Comparison of mixed-model approaches for association mapping. Genetics 178:1745–1754. doi:10.1534/genetics.107.079707

Storey, J.D., and R. Tibshirani. 2003. Statistical significance for genome-wide studies. Proc. Natl. Acad. Sci. USA 100:9440–9445. doi:10.1073/pnas.1530509100

Uitdewilligen, J.G.A.M.L., A.-M.A. Wolters, B.B. D'hoop, T.J.A. Borm, R.G.F. Visser, and H.J. van Eck. 2013. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS ONE 8(5):e62355. doi:10.1371/journal.pone.0062355

Van Eck, H.J., J.M. Jacobs, P. Stam, J. Ton, W.J. Stiekema, and E. Jacobsen. 1994. Multiple alleles for tuber shape in diploid potato detected by qualitative and quantitative genetic analysis using RFLPs. Genetics 137:303–309.

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

Voorrips, R.E., G. Gort, and B. Vosman. 2011. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinform. 12:172. doi:10.1186/1471-2105-12-172

Voorrips, R.E., and C.A. Maliepaard. 2012. The simulation of meiosis in diploid and tetraploid organisms using various genetic models. BMC Bioinform. 13:248. doi:10.1186/1471-2105-13-248

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–208. doi:10.1038/ng1702

Zhang, Z., E. Ersoz, C. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, P.J. Bradbury, J. Yu, D.K. Arnett, J.M. Ordovas, and E.S. Buckler. 2010. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42:355–360. doi:10.1038/ng.546

Zhao, K., M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg. 2007. An Arabidopsis example of association mapping in structured samples. PLoS Genet. 3(1):e4. doi:10.1371/journal.pgen.0030004

Zhou, X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44:821–824. doi:10.1038/ng.2310