QTL Mapping in Outbred Tetraploid (and Diploid) Diallel Populations

Rodrigo R. Amadeu[1], Patricio R. Munoz[1], Chaozhi Zheng[2], Jeffrey B. Endelman[3*]
[1]Horticultural Sciences Department, University of Florida, Gainesville, FL, USA
[2]Biometris, Wageningen University and Research, Wageningen, The Netherlands
[3]Department of Horticulture, University of Wisconsin, Madison, WI, USA
[*]corresponding author: endelman@wisc.edu

Running title: QTL mapping in tetraploid diallel

Key words: F1 population, multiparental, dominance, haplotypes, polyploidy

Abstract

Over the last decade, multiparental populations have become a mainstay of genetics research in diploid species. Our goal was to extend this paradigm to autotetraploids by creating computational tools for the analysis of connected F1 populations derived from a set of shared parents. In a companion paper, software to reconstruct F1 progeny in terms of parental haplotypes was described. For this study, we developed software for quantitative trait locus (QTL) mapping via Bayesian regression of phenotypes on the parental genotype probabilities. Statistical properties of the QTL model were explored by analyzing simulated half-diallel diploid and tetraploid populations with different population sizes, genome sizes, and numbers of parents. As expected, the LOD threshold needed to control the false positive rate increased with genome size, ploidy, and parents. Across the different scenarios, the number of progeny per parental haplotype (pph) largely determined the statistical power for QTL detection and the accuracy of the estimated haplotype effects. A QTL with heritability 0.1 was detected with 90% probability at 60 pph, while only 40 pph were needed to estimate the haplotypes with 90% accuracy. Our methodology includes a comprehensive treatment of dominance for multi-allelic QTL, which was illustrated by analyzing potato tuber shape in a 3×3 half-diallel population. A well-known QTL on chromosome 10 was detected, and the best-fit model included both additive and dominance effects. In terms of practical impacts on breeding, the software is already being used to select offspring based on the effect and dosage of particular haplotypes.

1 Introduction

For over three decades, the genetic mapping of quantitative trait loci (QTL) using DNA markers has been essential to basic and applied research. Early studies in plants focused on experimental diploid populations created from two parents (Lander and Botstein 1989; Knapp *et al.* 1990). For inbred parents, genetic marker alleles are easily coded to uniquely identify the two parental haplotypes at each locus (Young and Tanksley 1989). For outbred parents of ploidy ϕ, there are 2ϕ parental haplotypes but, in the case of bi-allelic SNPs, only two marker alleles. This difficulty was originally circumvented by using single-dose markers to create separate maternal and paternal maps in diploid species (Grattapaglia and Sederoff 1994) and separate maps

for each homologous chromosome in polyploids (Wu *et al.* 1992). Advances in estimation theory (Ritter and Salamini 1996; Wu *et al.* 2002) and software (Van Ooijen and Voorrips 2001; Margarido *et al.* 2007) eventually made it possible to reconstruct F_1 diploid offspring in terms of their parental haplotypes, and QTL mapping soon followed (Van Ooijen 2004). Inference of parental haplotypes in F_1 polyploid offspring based on bi-allelic SNPs is even more challenging. Software to do this in autotetraploids was originally developed by (Hackett *et al.* 2013, 2017), and within the last few years computational tools for even higher ploidy have been published (Bourke *et al.* 2018; Mollinari and Garcia 2019).

Not long after QTL mapping in biparental populations became a reality, several groups explored the use of multiple, connected biparental populations via theory and simulation (Rebai and Goffinet 1993; Muranty 1996; Liu and Zeng 2000). We use the term diallel to represent a general class of mating designs in which offspring are derived from no more than two parents (Figure 1). The transition from theory to practice for diallel mapping began in maize with 25 diverse inbred lines mated to a common parent (Yu *et al.* 2008; Buckler *et al.* 2009), and similar efforts have been published for other crops (Nice *et al.* 2017; Song *et al.* 2017). Examples of QTL mapping in connected F_1 populations derived from outbred diploids have also been published (Rosyara *et al.* 2013; Bink *et al.* 2014). Compared with MAGIC (multi-parent advanced generation intercross) designs, in which offspring contain haplotypes from more than two parents (Huang *et al.* 2015), diallel designs have the advantage of mimicking how plant breeding programs typically work. This is particularly important for breeding programs with limited resources to create specialized populations for discovery research.

The objective of this research was to develop and apply software for QTL mapping in autotetraploid diallel populations (i.e., connected F_1 populations), for which no other software is available. Our software, which is named *diaQTL* and available as a package for the R Computing Environment (R Core Team 2020), regresses phenotypes on the expected dosage of the parental genotypes. For additive effects, the appropriate genotype is the parental haplotype; for digenic dominance effects, it is the parental diplotype, etc. The diaQTL package does not compute the parental genotype probabilities from marker data but instead relies on the user to do this with the companion software PolyOrigin (Zheng *et al.* 2020). diaQTL can also analyze outbred diploid populations based on genotype probabilities computed with the software RABBIT (Zheng *et al.* 2015). diaQTL performs a regression analysis for each marker independently using Bayesian methods implemented in R package BGLR (Pérez and de Los Campos 2014). Because the genotype probabilities from RABBIT or PolyOrigin are based on multi-point linkage analysis, when the number of markers is high relative to the number of recombination breakpoints in the population (which is usually true for modern genotyping platforms), single marker regression is practically equivalent to interval mapping. After exploring the statistical properties of QTL mapping in outbred diallel populations via simulation, the software was used to analyze tuber shape in a 3×3 diallel of tetraploid potato.

2 Materials and Methods

2.1 QTL Model

A Bayesian mixed model is used to relate the phenotype to the effects of ϕp parental haplotypes in a diallel with p parents of ploidy ϕ. For the tth measurement (e.g., plot) of clone i, the linear model is (bold font designates a vector):

$$y_{it} = \mu + \mathbf{x}_{it}\mathbf{c} + \sum_{z=1}^{\delta} \mathbf{w}_i^{(z)} \boldsymbol{\alpha}^{(z)} + g_i + \varepsilon_{it} \tag{1}$$

The response variable y_{it} denotes the phenotype, μ is the intercept, $\mathbf{c}$ is an optional column-vector of fixed effects (e.g., year, location, blocks) with corresponding row-vector of covariates $\mathbf{x}_{it}$, $\boldsymbol{\alpha}^{(z)}$ is the column-vector of QTL effects of order z (1 = additive, 2 = digenic dominant, 3 = trigenic dominant, 4 = quadrigenic dominant) with corresponding row-vector of expected dosages $\mathbf{w}_i^{(z)}$ at a given locus, g_i is a polygenic effect, and ε_{it} is the residuals. The QTL effects of order z are independent and identically distributed (i.i.d.) and sum to zero. Model selection tests (described below) are used to choose the highest order term, which cannot exceed the ploidy ($\delta \leq \phi$). Models are named based on the highest order term but include all lower order terms (e.g., a digenic model includes both additive and digenic effects). The prior density for the QTL effects is the two-component mixture model known as BayesC in the genomic selection literature (Habier *et al.* 2011). The BayesC model combines a point mass and normal distribution, which has the effect of inducing both variable selection (i.e., some effects are exactly zero) and shrinkage. The mixing probability and variance of the normal distribution are also random (hyper)parameters with their own prior densities (beta and inverse chi-squared, respectively), chosen according to the default rules implemented in the BGLR software. The polygenic effect follows the reproducing kernel Hilbert space model in BGLR, which means it has a multivariate normal prior with covariance proportional to the additive relationship matrix (defined below), and the prior density for the variance component (V_{poly}) is inverse chi-squared. For computational efficiency, the polygenic effect is omitted during QTL discovery with *diaQTL::scan1* but included when estimating QTL effects with *diaQTL::fitQTL*, and markers with the same map position in cM are combined in a single marker-bin. The residuals are assumed to be i.i.d. normal with variance V_ε, which has an inverse chi-squared prior.

The coefficients of the genetic effects, $\mathbf{w}_i^{(z)}$, are calculated from an input file containing parental genotype probabilities. For diploid F_1 populations, the number of possible parental genotypes is four. Denoting the haplotypes in parent 1 as $1, 2$ and parent 2 as $3, 4$, these genotypes are 1–3, 1–4, 2–3, 2–4. For diploid S_1 populations, there are three possible genotypes: 1–1, 1–2, and 2–2. In tetraploids, the haplotypes are denoted $1, 2, 3, 4$ and $5, 6, 7, 8$ for parents 1 and 2, respectively. Allowing for double reduction, there are 100 F_1 genotypes and 35 S_1 genotypes (Zheng *et al.* 2020). Focusing on the tetraploid case, let $\pi_{i,a\text{–}b\text{–}c\text{–}d}$ denote the probability of genotype a–b–c–d for clone i, and $(a\text{–}b\text{–}c\text{–}d)_k$ is the set of all k-combinations (i.e., k elements chosen without replacement from the set a, b, c, d and order does not matter). If $\#\{(a\text{–}b\text{–}c\text{–}d)_1 = j\}$ denotes the number of 1-combinations equal to j (e.g., $\#\{(1\text{–}2\text{–}3\text{–}3)_1 = 3\}$ equals 2), then the expected dosage of haplotype j in individual i can be written as

$$w_{i,j}^{(1)} = \sum_{a\text{–}b\text{–}c\text{–}d} \#\{(a\text{–}b\text{–}c\text{–}d)_1 = j\}\pi_{i,a\text{–}b\text{–}c\text{–}d} \tag{2}$$

where the sum in Equation 2 is over all possible parental genotypes. For diploid dominance or digenic dominance in tetraploids, the expected dosage of the corresponding diplotype follows a similar expression:

$$w_{i,j\text{–}j'}^{(2)} = \sum_{a\text{–}b\text{–}c\text{–}d} \#\{(a\text{–}b\text{–}c\text{–}d)_2 = j\text{–}j'\}\pi_{i,a\text{–}b\text{–}c\text{–}d} \tag{3}$$

The coefficients of the trigenic and quadrigenic dominance effects are calculated analogously.

Realized identity-by-descent (IBD) relationship matrices of order z (1 = additive, 2 = digenic dominant, etc.) are calculated from the dosage vectors introduced in Equation 1, using the function *diaQTL::IBDmat*. If $\mathbf{w}_{i_k}^{(z)}$ denotes the row-vector of dosages for clone i at locus k, the IBD relationship coefficient between clones i and j based on m marker-bins is

$$G_{ij}^{(z)} = \left[\binom{\phi}{z} m\right]^{-1} \sum_{k=1}^{m} \mathbf{w}_{i_k}^{(z)} (\mathbf{w}_{j_k}^{(z)})^T \tag{4}$$

where $\binom{\phi}{z}$ is the binomial coefficient and T is the transpose of the vector. When used in *diaQTL::fitQTL*, the additive relationship matrix is calculated while leaving out the chromosome containing the QTL (Yang *et al.* 2014).

Model parameters are estimated using the Gibbs sampling algorithm in the BGLR package, which saves the values of the genetic effects, $\boldsymbol{\alpha}^{(z)}$, at each iteration of the Markov chain following the burn-in phase (accessible using the function *BGLR::readBinMat*). The sample mean of these values is reported as the posterior mean, and user-specified quantiles are reported as a Bayesian credible interval (CI). The function *diaQTL::set_params(q,r)* uses R package *coda* (Plummer et al. 2006) to determine the number of burn-in and total iterations needed to estimate quantile q within the interval $(q - r, q + r)$ with 95% probability (Raftery and Lewis 1992). For each iteration of the Markov chain, the genetic variances for the QTL effects, $V_g^{(z)}$, are estimated as the sample variance of the corresponding genetic values, $u_i^z = \mathbf{w}_i^{(z)}\boldsymbol{\alpha}^{(z)}$, for the clones. For each iteration, the proportion of variance due to each genetic effect is estimated as

$$h_{(z)}^2 = \frac{V_g^{(z)}}{V_\varepsilon + \lambda V_{poly} + \sum_{z'} V_g^{(z')}} \tag{5}$$

with posterior means and CI computed as described above. The parameter λ in Equation 5 equals the mean diagonal elements of the additive relationship matrix (expanded to include replicates, $\lambda = n^{-1}\sum_{it} G_{ii}^{(1)}$ for a dataset with n observations), to account for inbreeding (Endelman and Jannink 2012).

For QTL discovery during a genome scan, the statistical significance of Equation 1 is assessed relative to a null model that includes the general combining ability (GCA) of the parents but no QTL. The GCA model can

be viewed as a restricted additive model with all haplotype effects for a given parent constrained to be equal. The BayesRR (ridge regression) model from the BGLR package is used for the GCA effects, which involves a normal distribution prior and inverse chi-squared distribution prior for the variance. Statistical support for the QTL is quantified using a logarithm of the odds (LOD) score, defined as log_{10} of the ratio between the likelihoods for the QTL vs. null models, evaluated at their respective posterior means. Thresholds for QTL discovery were based on determining the null distribution of the LOD score for the strictly additive model using simulation (see below), but *diaQTL::scan1_permute* can also be used to estimate the null distribution by random permutation of the phenotypes (Churchill and Doerge 1994). For computational efficiency, marker-bins are constructed from the input genetic map, and only one marker per bin is tested during the genome scan.

The function *diaQTL::BayesCI* computes a Bayesian credible interval for the QTL location following the methodology used in *R/qtl* (Broman and Sen 2009). Let $f(x)$ denote the probability density at map position x (in cM), which satisfies the normalization condition $\int f(x)dx = 1$ for the chromosome. When discretized across marker-bins, this normalization can be written as $\sum_k c_k f_k = 1$, where $f_k = 10^{LOD_k}$ is the unnormalized probability for bin k, $c_k = dx_k / \sum_k f_k dx_k$, and dx_k is the bin-width in cM. The cumulative distribution at bin k is $F_k = \sum_{i=1}^{k} c_i f_i$. The lower bin for the credible interval with probability $1 - \alpha$ is one less than the largest k satisfying $F_k \leq \alpha/2$. The upper bin is one more than the smallest k satisfying $F_k \geq 1 - \alpha/2$.

To select the best genetic model (e.g., strictly additive vs. additive + digenic dominance) for a particular QTL, *fitQTL* returns the change in the deviance information criterion relative to the null model (ΔDIC). Each DIC is computed by the BGLR package and defined as the posterior mean deviance plus the effective number of parameters, with lower values representing a better tradeoff between model complexity and fit (Spiegelhalter *et al.* 2002).

2.2 Simulation

Simulations were performed to assess statistical power, positional accuracy, and accuracy of the estimated effects, using a half-diallel mating design with no selfing (Figure 1C). Using the software PedigreeSim V2.0 (Voorrips and Maliepaard 2012) and auxiliary R package PedigreeSimR (`https://www.github.com/rramadeu/PedigreeSimR`), genomes were simulated with 1 to 12 linkage groups, chromosomes with a length of 100 cM, and loci evenly spaced at 0.1 cM. For the tetraploid simulation, only bivalents were allowed with no preferential pairing of the homologous chromosomes. For each simulation, a single additive QTL was randomly assigned to one locus, and all other loci were used as markers for mapping. Phenotypes were simulated by $y_i = \mathbf{w}_i^{(1)} \boldsymbol{\alpha}^{(1)} + \varepsilon_i$, (symbols have the same meaning as Equation 1, but here $\mathbf{w}_i^{(1)}$ is the true parental haplotype dosage), using i.i.d. standard normal QTL effects and i.i.d. normal residuals with variance V_ε. The additive genetic variance for this model equals the ploidy, and the residual variance was chosen to achieve different values of heritability (Equation 5).

A factorial arrangement with $3 \times 3 \times 3 \times 2 = 54$ treatment scenarios was used to investigate the effects of total population size (200, 400, 800), number of parents (3, 4, 6), heritability (0.1, 0.2, 0.4), and ploidy (2 and 4), assuming a genome size of 12 Morgans. Progeny per parental haplotype (pph) was com-

puted as population size divided by the number of parental haplotypes (e.g., population size of 800 with 4 autotetraploid parents has 16 haplotypes, so pph=50). For each scenario, $nrep = 1000$ simulations were subjected to a genome scan. If the realized heritability, computed from the simulated QTL effects and residuals, deviated from the prescribed value by more than 10^{-4} (due to finite sample size), the simulation was discarded before QTL mapping. The number of iterations for the Gibbs sampler was determined based on *diaQTL::set_params(q=0.5,r=0.1)*, using the 90th percentile of the combined results across all haplotypes for the first marker of every chromosome. Statistical power was the proportion of simulations in which both flanking markers of the simulated QTL exceeded the LOD threshold. As the power was 1 for all scenarios with $h^2 = 0.4$, these results are not shown. Accuracy of the inferred QTL position was the distance between the simulated position and marker with the highest LOD score. Accuracy for the posterior means of the additive effects was evaluated based on the estimated heritability and Pearson's correlation with the simulated values. Standard errors were computed as $s/\sqrt{nrep}$, where s is the standard deviation of the statistics. Monotone splines were fit using R package *scam* (Pya and Wood 2015).

The LOD threshold for QTL discovery was set to control the genome-wide false positive rate at $\alpha = 0.05$. This threshold was determined by simulating a trait with $h^2 = 0$ and using the 95th percentile of the maximum LOD score across 1000 simulations. For each ploidy level, a monotone increasing, concave spline for the effects of genome size and number of parents was fit using R package scam and is available through the function *diaQTL::LODthresh*. As expected from theory (Lander and Botstein 1989), population size did not influence the threshold (results not shown).

2.3 Potato dataset

The potato dataset is a half-diallel population (without selfing) for three parents from the University of Wisconsin-Madison (UW) breeding program: Villetta Rose, W6511-1R, and W9914-1R. The population was genotyped using version 3 of the potato SNP array, and allele dosage was called using R package fitPoly (Voorrips *et al.* 2011; Zych *et al.* 2019). There were 5334 markers distributed across 12 chromosome groups, and their physical positions were based on the potato DMv4.03 reference genome (Potato Genome Sequencing Consortium 2011; Sharma *et al.* 2013). Parental genotype probabilities and a genetic map were calculated using the software PolyOrigin (Zheng *et al.* 2020), which identified 19 of the 434 progeny as outliers, and phenotype data were unavailable for 2 more. The remaining 413 clones were distributed as follows: Villetta Rose × W6511-1R (154 individuals), Villetta Rose × W9914-1R (113 individuals), and W6511-1R × W9914-1R (146 individuals). The genetic map produced by PolyOrigin spanned 12.1 Morgans and contained 2781 marker-bins.

The half-diallel population was evaluated as part of a larger field trial in 2018 at the UW Hancock Agricultural Research Station, Hancock, WI. The trial used an augmented design with four incomplete blocks and seven repeated checks per block. Each plot was a single row sown with 15 seeds (tuber pieces) at 30 cm in-row spacing and 90 cm between rows. The trial was harvested mechanically, and tubers were passed through an optical sizer after washing to measure tuber length and width. Tuber shape was calculated as the average length/width ratio of all tubers weighing 170–285g. To improve normality of the residuals, the trait was transformed to $log(L/W–1)$. Initial statistical analysis using ASReml-R v4 (Butler *et al.* 2018) was based on the linear model $y_{ij} = \mu + block_j + clone_i + \varepsilon_{ij}$. Broad-sense heritability on a plot basis was

estimated at 0.91 by treating $clone_i$ and $block_j$ as random effects (i.i.d. normal). BLUEs for each clone were estimated by treating $clone_i$ as fixed and then used as the response variable for analysis with the diaQTL package. The number of burn-in (50) and total (500) iterations for *scan1* was based on the output from *set_params(q=0.5,r=0.1)*. To estimate the 90% CI for the QTL effects with *fitQTL*, the number of burn-in (100) and total (25,000) iterations was based on the output from *set_params(q=0.05,r=0.025)*.

2.4 Data availability

The R/diaQTL package is available under the GPL-3 license at `https://github.com/jendelman/diaQTL`, and the potato dataset is distributed with the package.

3 Results

3.1 Statistical power and accuracy

Simulated half-diallel populations were generated to study the influence of genome size, ploidy, and number of parents on the LOD threshold needed to control the false positive rate (FPR) at $\alpha = 0.05$. The points in Figure 2 are the average for 1000 simulations analyzed with the additive QTL model, and the lines are monotone increasing, concave splines to allow for interpolation. The increase in LOD threshold with genome size was sub-linear, with a slope of 0.14 LOD/M at 5 Morgans (M) but only 0.07 LOD/M at 10 M, and this slope was largely unaffected by ploidy or parent number. The effect of parent number was also sub-linear and independent of genome size, but the slope for tetraploids was approximately twice that of diploids. With 3 parents, the slopes were 0.37 LOD/parent (2x) vs. 0.72 LOD/parent (4x), while for 6 parents the slopes were 0.28 LOD/parent (2x) vs. 0.58 LOD/parent (4x). The LOD threshold for tetraploids was consistently 1.7 times that for diploids, and the threshold for a tetraploid population with 2 parents was comparable to a diploid population with 8 parents (Figure 2). Because the digenic QTL model has more parameters than the additive model, higher LOD thresholds are needed to maintain the FPR at 0.05 (Figure S1).

The influence of population size, number of parents, ploidy, and QTL heritability on statistical power was studied for a genome of 12 Morgans. Power increased with population size and heritability but decreased with the number of parents and ploidy (Figure S2). For a given h^2, the interplay between these factors could be largely summarized by the number of progeny per parental haplotype (pph), although the diploid (tetraploid) results were consistently below (above) the regression spline (Figure 3). At $h^2 = 0.1$, power reached 0.5 and 0.9 with 25 and 60 pph, respectively. For $h^2 = 0.2$, only 10 and 25 pph were needed to achieve powers of 0.5 and 0.9, respectively. As power increased, so did the accuracy of the inferred QTL position, as measured by the distance between the most significant marker and QTL (Figure S3). The positional accuracy was 6 cM when the power was 0.5 and decreased to 3 cM when power was 0.9.

The accuracy of the predicted haplotype effects, as measured by the correlation with the simulated values, was also largely explained by the pph metric (Figure 4). For $h^2 = 0.1$, 40 pph were sufficient to achieve an accuracy of 0.9, while for $h^2 = 0.2$, only 20 pph were needed. Across all of the scenarios tested, the

heritability of the QTL could be estimated with essentially no error (Figure S2), even when the accuracy of the haplotype effects was only 0.7.

3.2 Potato diallel

To illustrate the functionality of the diaQTL software with a real dataset, we analyzed potato tuber shape, measured by the length to width (L/W) ratio, in a 3 × 3 half diallel without selfing. The ideal range of values for tuber shape is determined by the end-product (e.g., long tubers for French fries, round tubers for potato chips) and cultural preferences that vary by country. The parents of this diallel were developed for the red-skin fresh market in the US, for which a round to slightly oblong shape is expected. The distribution of L/W values for the diallel population ranged from 1.0 to 2.0 (Figure S4), while the commercial checks used in the field experiment had values between 1.1 and 1.2.

There are two types of pedigree relationship in a diallel without selfing: full-sibs and half-sibs. The expected values for the additive, digenic dominance, trigenic dominance, and quadrigenic dominance relationship coefficients for full- and half-sib autotetraploids are shown in Table 1 (Gallais 2003). Whereas half-sib diploids have zero covariance between their dominance values, in tetraploids the digenic relationship coefficient is 1/36 because the gametes are diploid. The coefficients of the additive and dominance effects used for QTL mapping can be used to compute realized relationships (Equation 5 and Figure 5), which deviate from the expected values (dashed lines in Figure 5) due to the finite size of the genome (Amadeu *et al.* 2020). For the additive and digenic dominance coefficients, the distribution for the diallel population was fairly symmetric, so the expected value coincided with the mode. For the higher order dominance terms, the distribution was right-skewed, so the expected value was larger than the mode, but the observed averages were consistent with the expected value (Table 1).

A genome scan with the additive model identified a significant QTL for tuber shape on chromosome group 10 (Figure 6A). The most significant marker was solcap_snp_c2_25522 located at 48,617,457 bp, with a 95% Bayesian credible interval of 2.6 Mb (48,203,384 - 50,782,097). This region of chromosome 10 coincides with the location of the classical *Ro* (round) QTL in potato (Van Eck *et al.* 1994), and the causal gene has been identified as *StOVP20*, an *OVATE Family Protein* with orthologs that affect fruit shape in tomato, melon, and cucumber (Wu *et al.* 2018). *StOVP20* is not present in the DM reference genome, but based on synteny analysis (Wu *et al.* 2018), it is flanked by the DM gene models (PGSC0003DMG400006678, PGSC0003DMG400006679). This corresponds to the DM genome interval (48,982,521 - 49,022,687), which lies within the Bayesian CI for the QTL.

Based on DIC, a penalized goodness-of-fit metric analogous to AIC (lower is better), the best genetic model for the chromosome 10 QTL (solcap_snp_c2_25522) included additive and dominance effects (Table 2). The digenic and trigenic models had similar DIC values, with 25% of the variance partitioned as additive, approximately 10% as dominant, and 32% as polygenic. Of the 12 parental haplotypes in the population, the largest additive effect (in magnitude) was estimated for haplotype 2 of parent W6511-1R (Figure 6B), and the negative sign means it reduced the L/W ratio. Haplotypes 2 and 4 of parent Villetta Rose also reduced the L/W ratio. For parent W9914-1R, none of the haplotypes had a strong additive effect but showed large negative digenic effects (Figure 6C), meaning they also reduced L/W ratio.

4 Discussion

Although this is the first publication on statistical power and accuracy in tetraploid diallel populations, some comparisons with previous studies on diploid diallel or biparental tetraploid populations are possible. For biparental populations from inbred lines, Lander and Botstein (1989) derived an approximate expression for the relationship between genome size and LOD threshold. Although the absolute LOD values are different, which is expected due to differences in population design and the statistical model, the trend is similar to our results (Figure S5). Among previous simulation studies, Muranty (1996) is unique for explicitly showing the loss of statistical power as the number of outbred diploid parents increased at a fixed total population size. We observed the same trend (Figure 3), but the exact values for power are not comparable because of differences in the significance level (higher values produce higher power). Muranty (1996) also reported that the specific form of the mating design (e.g., cyclic, factorial, half-diallel) had little impact on power, which was confirmed in our simulation (data not shown). Bourke *et al.* (2019) reported statistical power in a biparental tetraploid population of size 200, which corresponds to 25 pph, was 0.8 for a bi-allelic, additive QTL with $h^2 = 0.1$ and high genotypic information content. This is higher than the value of 0.5 for the regression spline in Figure 3, which may be the result of the simpler genetic architecture (bi-allelic vs. multi-allelic in our simulation).

One of the challenges with multiparental populations is the large number of parameters to estimate, or in other words, model complexity. The number of QTL effects of order z, where $z = 1$ is additive, $z = 2$ is digenic dominant, etc., increases with the number of parents raised to the zth power. For models with higher order dominance, the number of QTL effects can even exceed the number of observations. In the 3×3 tetraploid potato diallel, there were $12 + 78 + 240 + 300 = 630$ QTL effects in the quadrigenic dominance model but only 413 progeny. However, this model complexity challenge is even more severe in genomic prediction, where one attempts to estimate the effect of all markers simultaneously (Meuwissen *et al.* 2001; Habier *et al.* 2011), and the BGLR package used in diaQTL was originally developed for this purpose. The advantages of using random instead of fixed effects (or equivalently, penalized regression methods) for complex models is well established (Robinson 1991; James *et al.* 2013). Random effects have been used in other QTL mapping software, such as FlexQTL (Bink *et al.* 2008, 2014), QTLpoly (da Silva Pereira *et al.* 2020), and several programs designed for MAGIC populations (Verbyla *et al.* 2014; Zhang *et al.* 2014; Wei and Xu 2016). FlexQTL appears limited to bi-allelic QTL, in which case it makes little practical difference whether the effects are random or fixed. Like diaQTL, QTLpoly allows each parental haplotype to have a different effect, but QTLpoly assumes a normal distribution, which is more restrictive than the BayesC model used in diaQTL.

Compared to other software for QTL linkage mapping in tetraploids (Hackett *et al.* 2017; Bourke *et al.* 2018; da Silva Pereira *et al.* 2020), diaQTL is unique in its ability to model dominance for multi-allelic QTL. diaQTL uses the classical decomposition of genetic variance, which is based on orthogonal effects in panmictic populations and was first derived for polyploids by Kempthorne (1955, 1957). A potential downside to this approach is that complete dominance requires the highest order interaction (quadrigenic dominance for a tetraploid) and therefore a large number of parameters when parental haplotypes are treated as distinct QTL alleles (e.g., Figure 6C). Model complexity would be reduced if equivalence relations were imposed on the parental haplotypes, thereby reducing the number of QTL alleles, and examples of this based on model selection tests can be found in the literature (Jannink and Wu 2003; Bourke *et al.* 2018).

Although implementing this approach could improve model interpretability, it would complicate the process of model selection and—we speculate—only lead to marginal improvements in statistical power. We are more interested in pursuing equivalence relations based on sequence data than changes in the likelihood. Toward this end, we are encouraged by recent methodological advances for haplotype sequence assembly in polyploids that make explicit use of family structure (Motazedi *et al.* 2018, 2019).

4.1 Implications for breeding

Large plant breeding programs often maintain distinct research enterprises devoted to genetic discovery vs. variety selection. Genetic discovery projects typically characterize novel populations, such as a panel of distantly related germplasm or larger-than-normal, unselected biparental populations, but this additional effort is not always feasible for small programs. There is also the potential disconnect between the development of a genetic marker in the discovery population and its successful deployment in the breeding populations; such markers often fail because they are not haplotype-specific. With diallel QTL analysis of the breeding population, this additional step would be unnecessary. The diaQTL function *diaQTL::haplo_plot* allows users to visualize which parental haplotypes are present at a particular locus (Figure 6D), and it is straightforward to select progeny based on the dosage of particular haplotypes using *diaQTL::haplo_get*.

To further illustrate the feasibility of using breeding populations for discovery research in polyploids, consider first the potato. An idealized mating design at the University of Wisconsin involves 10 females and 4 males (male fertility is scarce), from which 40 F1 populations with 500 progeny each are created and evaluated as single plants in the first year. Only 5% of the 20,000 = 1000 individuals advance based on visual selection to the second year (first clonal stage), at which point high-throughput phenotyping can be used to efficiently measure plant development and tuber yield. If the parents are equally represented in year 2, there would be 25 progeny per female parental haplotype and 62 per male parental haplotype. Based on our simulation results, this coverage translates into over 90% power for QTL with $h^2 = 0.2$ and over 50% power for QTL with $h^2 = 0.1$. The blueberry breeding program at the University of Florida uses a circular mating design (Figure 1D), with approximately 150 parents and 150 progeny per F1 population. Traditionally, after 1 year, 10% of the seedlings are selected and genotyped. A portion of these parents have been phenotyped for several years, this can guide the choice for contrasting parents for the traits of interest. Using this information, we can increase the number of selected individuals to 30% in the chosen fullsib populations (to guarante 20 pph), and now they can be used simultaneously for breeding and QTL mapping.

The above discussion has focused on the feasibility of phenotyping diallel breeding populations, but the economics of genotyping is equally important. Fortunately, the cost of library preparation has dropped to the point where most public breeding programs can afford to genotype over 500 clones per year. For polyploids, there is the additional challenge of sequencing at sufficient depth to estimate allele dosage. If dosage is determined for each marker independently, the numbers are grim: at least 60 reads are needed to differentiate duplex from simplex dosage in a tetraploid with less than 5% error (Uitdewilligen *et al.* 2013; Gerard *et al.* 2018; Matias *et al.* 2019). However, because PolyOrigin jointly analyzes all markers on the chromosome, the necessary read depth per marker is much lower and inversely related to the number of markers: with 500 markers, a depth of 10 is sufficient to infer parental haplotypes with less than 5% error

(Zheng *et al.* 2020). For simplicity, our simulation study did not consider genotyping error, which has been shown in biparental populations to reduce power and accuracy (Bourke *et al.* 2019). Nonetheless, the results are promising enough that we have begun implementing these ideas for the potato and blueberry breeding programs under our management.

5 Conclusion

The R/diaQTL software is a user-friendly tool for QTL mapping and haplotype analysis of connected, autotetraploid F_1 populations derived from a set of shared parents. Through simulation, we have demonstrated that the number of progeny (or gametes, if selfed populations are used) per parental haplotype is a critical parameter when selecting populations for analysis. As illustrated with the potato dataset, a unique feature of diaQTL is its comprehensive approach to modeling higher order dominance. By enabling selection on the effect and dosage of haplotypes, without having to develop haplotype-specific markers, diaQTL is already having a positive impact on the efficiency of tetraploid breeding programs.

6 Acknowledgments

Financial support provided by USDA NIFA Award No. 2019-67013-29166. We thank Maria Caraza-Harter for testing the diaQTL software.

Author contributions: RRA and JBE developed the software, analyzed the data, and drafted the manuscript. All authors contributed to the methodology and edited the manuscript.

References

Amadeu, R. R., L. A. C. Lara, P. Munoz, and A. A. F. Garcia, 2020 Estimation of molecular pairwise relatedness in autopolyploid crops. G3: Genes, Genomes, Genetics **10**: 4579–4589.

Bink, M. C., M. P. Boer, C. J. Ter Braak, J. Jansen, R. E. Voorrips, *et al.*, 2008 Bayesian analysis of complex traits in pedigreed plant populations. Euphytica **161**: 85–96.

Bink, M. C., J. Jansen, M. Madduri, R. E. Voorrips, C. E. Durel, *et al.*, 2014 Bayesian QTL analyses using pedigreed families of an outcrossing species, with application to fruit firmness in apple. Theoretical and Applied Genetics **127**: 1073–1090.

Bourke, P. M., C. A. Hackett, R. E. Voorrips, R. G. F. Visser, and C. Maliepaard, 2019 Quantifying the power and precision of QTL analysis in autopolyploids under bivalent and multivalent genetic models. G3: Genes, Genomes, Genetics **9**: 2107–2122.

Bourke, P. M., G. van Geest, R. E. Voorrips, J. Jansen, T. Kranenburg, *et al.*, 2018 polymapR—linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. Bioinformatics **34**: 3496–3502.

Broman, K. W. and S. Sen, 2009 *A Guide to QTL Mapping with R/qtl*, volume 46. Springer.

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown, *et al.*, 2009 The genetic architecture of maize flowering time. Science **325**: 714–718.

Butler, D., B. Cullis, A. Gilmour, B. Gogel, and R. Thompson, 2018 *ASReml-R Reference Manual Version 4*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

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

da Silva Pereira, G., D. C. Gemenet, M. Mollinari, B. A. Olukolu, J. C. Wood, *et al.*, 2020 Multiple QTL mapping in autopolyploids: A random-effect model approach with application in a hexaploid sweetpotato full-sib population. Genetics **215**: 579–595.

Endelman, J. B. and J.-L. Jannink, 2012 Shrinkage estimation of the realized relationship matrix. G3: Genes, Genomes, Genetics **2**: 1405–1413.

Gallais, A., 2003 *Quantitative genetics and breeding methods in autopolyploids plants*. INRA, Paris, first edition.

Gerard, D., L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens, 2018 Genotyping polyploids from messy sequencing data. Genetics **210**: 789–807.

Grattapaglia, D. and R. Sederoff, 1994 Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. Genetics **137**: 1121–1137.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the bayesian alphabet for genomic selection. BMC bioinformatics **12**: 186.

Hackett, C. A., B. Boskamp, A. Vogogias, K. F. Preedy, and I. Milne, 2017 TetraploidSNPMap: software for linkage analysis and QTL mapping in autotetraploid populations using SNP dosage data. Journal of Heredity **108**: 438–442.

Hackett, C. A., K. McLean, and G. J. Bryan, 2013 Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. PloS one **8**: e63939.

Huang, B. E., K. L. Verbyla, A. P. Verbyla, C. Raghavan, V. K. Singh, *et al.*, 2015 MAGIC populations in crops: current status and future prospects. Theoretical and Applied Genetics **128**: 999–1017.

James, G., D. Witten, T. Hastie, and R. Tibshirani, 2013 *An introduction to statistical learning*, volume 112. Springer.

Jannink, J.-L. and X.-L. Wu, 2003 Estimating allelic number and identity in state of QTLs in interconnected families. Genetical Research **81**: 133–144.

Kempthorne, O., 1955 The correlation between relatives in a simple autotetraploid population. Genetics **40**: 168–74.

Kempthorne, O., 1957 One-Locus Polyploid Population. In *An introduction to genetic statistics*, chapter 18, pp. 400–409, John Wiley & Sons, New York.

Knapp, S., W. Bridges, and D. Birkes, 1990 Mapping quantitative trait loci using molecular marker linkage maps. Theoretical and applied genetics **79**: 583–592.

Lander, E. S. and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185–199.

Liu, Y. and Z.-B. Zeng, 2000 A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. Genetics Research **75**: 345–355.

Margarido, G. R., A. P. Souza, and A. A. Garcia, 2007 OneMap: software for genetic mapping in outcrossing species. Hereditas **144**: 78–79.

Matias, F. I., K. G. Xavier Meireles, S. T. Nagamatsu, S. C. Lima Barrios, C. Borges do Valle, *et al.*, 2019

Expected genotype quality and diploidized marker data from genotyping-by-sequencing of *Urochloa* spp. tetraploids . The Plant Genome **12**: 190002.

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

Mollinari, M. and A. A. F. Garcia, 2019 Linkage analysis and haplotype phasing in experimental autopolyploid populations with high ploidy level using hidden Markov models. G3: Genes, Genomes, Genetics **9**: 3297–3314.

Motazedi, E., D. de Ridder, R. Finkers, S. Baldwin, S. Thomson, *et al.*, 2018 TriPoly: haplotype estimation for polyploids using sequencing data of related individuals. Bioinformatics **34**: 3864–3872.

Motazedi, E., C. Maliepaard, R. Finkers, R. Visser, and D. de Ridder, 2019 Family-based haplotype estimation and allele dosage correction for polyploids using short sequence reads. Frontiers in Genetics **10**: 335.

Muranty, H., 1996 Power of tests for quantitative trait loci detection using full-sib families in different schemes. Heredity **76**: 156–165.

Nice, L. M., B. J. Steffenson, T. K. Blake, R. D. Horsley, K. P. Smith, *et al.*, 2017 Mapping agronomic traits in a wild barley advanced backcross–nested association mapping population. Crop Science **57**: 1199–1210.

Pérez, P. and G. de Los Campos, 2014 Genome-wide regression and prediction with the BGLR statistical package. Genetics **198**: 483–495.

Potato Genome Sequencing Consortium, 2011 Genome sequence and analysis of the tuber crop potato. Nature **475**: 189–195.

Pya, N. and S. N. Wood, 2015 Shape constrained additive models. Statistics and Computing **25**: 543–559.

R Core Team, 2020 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rebai, A. and B. Goffinet, 1993 Power of tests for QTL detection using replicated progenies derived from a diallel cross. Theoretical and Applied Genetics **86**: 1014–1022.

Ritter, E. and F. Salamini, 1996 The calculation of recombination frequencies in crosses of allogamous plant species with applications to linkage mapping. Genetics Research **67**: 55–65.

Robinson, G. K., 1991 That BLUP is a good thing: The estimation of random effects. Statistical Science **6**: 15–32.

Rosyara, U. R., M. C. Bink, E. van de Weg, G. Zhang, D. Wang, *et al.*, 2013 Fruit size QTL identification and the prediction of parental QTL genotypes and breeding values in multiple pedigreed populations of sweet cherry. Molecular Breeding **32**: 875–887.

Sharma, S. K., D. Bolser, J. de Boer, M. Sønderkær, W. Amoros, *et al.*, 2013 Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. G3: Genes, Genomes, Genetics **3**: 2031–2047.

Song, Q., L. Yan, C. Quigley, B. D. Jordan, E. Fickus, *et al.*, 2017 Genetic characterization of the soybean nested association mapping population. The Plant Genome **10**: 1–14.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde, 2002 Bayesian measures of model complexity and fit. Journal of the royal statistical society: Series b (statistical methodology) **64**: 583–639.

Uitdewilligen, J. G. A. M. L., A.-M. A. Wolters, B. B. D'hoop, T. J. A. Borm, R. G. F. Visser, *et al.*, 2013 A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS ONE **8**: e62355.

Van Eck, H. J., J. M. Jacobs, P. Stam, J. Ton, W. J. Stiekema, *et al.*, 1994 Multiple alleles for tuber shape

in diploid potato detected by qualitative and quantitative genetic analysis using RFLPs. Genetics **137**: 303–309.

Van Ooijen, J., 2004 *MapQTL® 5, software for the mapping of quantitative trait loci in experimental populations of diploid species*. Kyazma B.V., Wageningen, Netherlands.

Van Ooijen, J. and R. Voorrips, 2001 *Joinmap® 3, software for the calculation of genetic linkage maps*. Kyazma B.V., Wageningen, Netherlands.

Verbyla, A. P., A. W. George, C. R. Cavanagh, and K. L. Verbyla, 2014 Whole-genome QTL analysis for MAGIC. Theoretical and Applied Genetics **127**: 1753–1770.

Verhoeven, K. J., J. L. Jannink, and L. M. McIntyre, 2006 Using mating designs to uncover QTL and the genetic architecture of complex traits. Heredity **96**: 139–149.

Voorrips, R. E., G. Gort, and B. Vosman, 2011 Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinformatics **12**: 172.

Voorrips, R. E. and C. A. Maliepaard, 2012 The simulation of meiosis in diploid and tetraploid organisms using various genetic models. BMC Bioinformatics **13**: 248.

Wei, J. and S. Xu, 2016 A random-model approach to QTL mapping in multiparent advanced generation intercross (MAGIC) populations. Genetics **202**: 471–486.

Wu, K., W. Burnquist, M. Sorrells, T. Tew, P. Moore, *et al.*, 1992 The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theoretical and Applied Genetics **83**: 294–300.

Wu, R., C.-X. Ma, I. Painter, and Z.-B. Zeng, 2002 Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. Theoretical population biology **61**: 349–363.

Wu, S., B. Zhang, N. Keyhaninejad, G. Rodríguez, H. Kim, *et al.*, 2018 Estimation of molecular pairwise relatedness in autopolyploid crops. Nature communications **9**: 1–12.

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price, 2014 Advantages and pitfalls in the application of mixed-model association methods. Nature genetics **46**: 100–106.

Young, N. and S. Tanksley, 1989 Restriction fragment length polymorphism maps and the concept of graphical genotypes. Theoretical and Applied Genetics **77**: 95–101.

Yu, J., J. B. Holland, M. D. McMullen, and E. S. Buckler, 2008 Genetic design and statistical power of nested association mapping in maize. Genetics **178**: 539–551.

Zhang, Z., W. Wang, and W. Valdar, 2014 Bayesian modeling of haplotype effects in multiparent populations. Genetics **198**: 139–156.

Zheng, C., R. R. Amadeu, P. Munoz, and J. B. Endelman, 2020 Haplotype reconstruction in connected tetraploid F1 populations. bioRxiv https://www.biorxiv.org/content/10.1101/2020.12.18.423519v1.

Zheng, C., M. P. Boer, and F. A. van Eeuwijk, 2015 Reconstruction of genome ancestry blocks in multiparental populations. Genetics **200**: 1073–1087.

Zych, K., G. Gort, C. A. Maliepaard, R. C. Jansen, and R. E. Voorrips, 2019 FitTetra 2.0–improved genotype calling for tetraploids with multiple population and parental data support. BMC Bioinformatics **20**: 148.

Figures and Tables

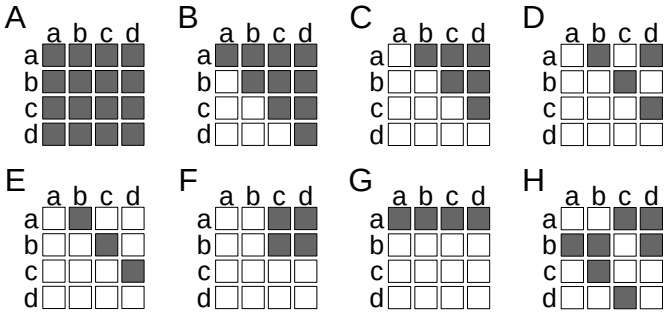

Figure 1: Examples of diallel populations with four parents (labeled a,b,c,d); each square represents a full-sib family. (A) full diallel; (B) half diallel with selfing; (C) half diallel without selfing; (D) circular; (E) linear; (F) factorial; (G) testcross or nested design; (H) arbitrary partial diallel. Figure adapted from Verhoeven *et al.* (2006).

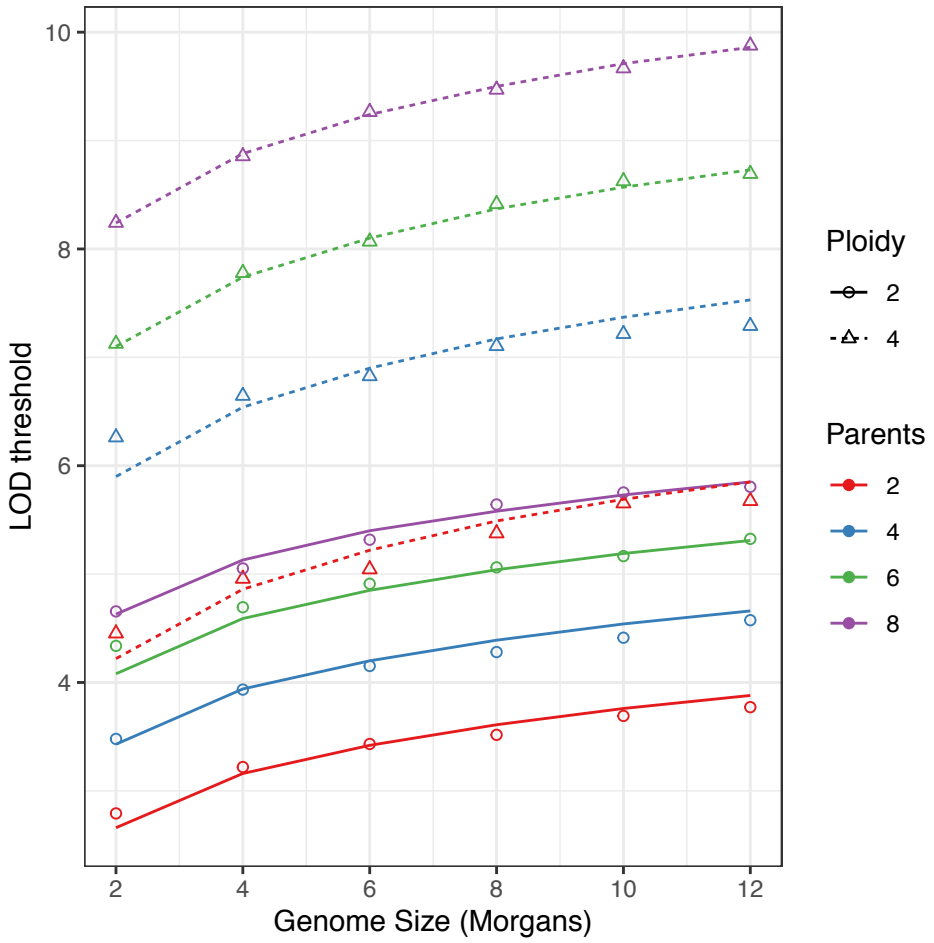

Figure 2: LOD threshold needed to control the false positive rate at 0.05, for a half-diallel design without selfing and the additive QTL model. Each point is based on 1,000 simulations, and the lines are monotone increasing, concave splines.

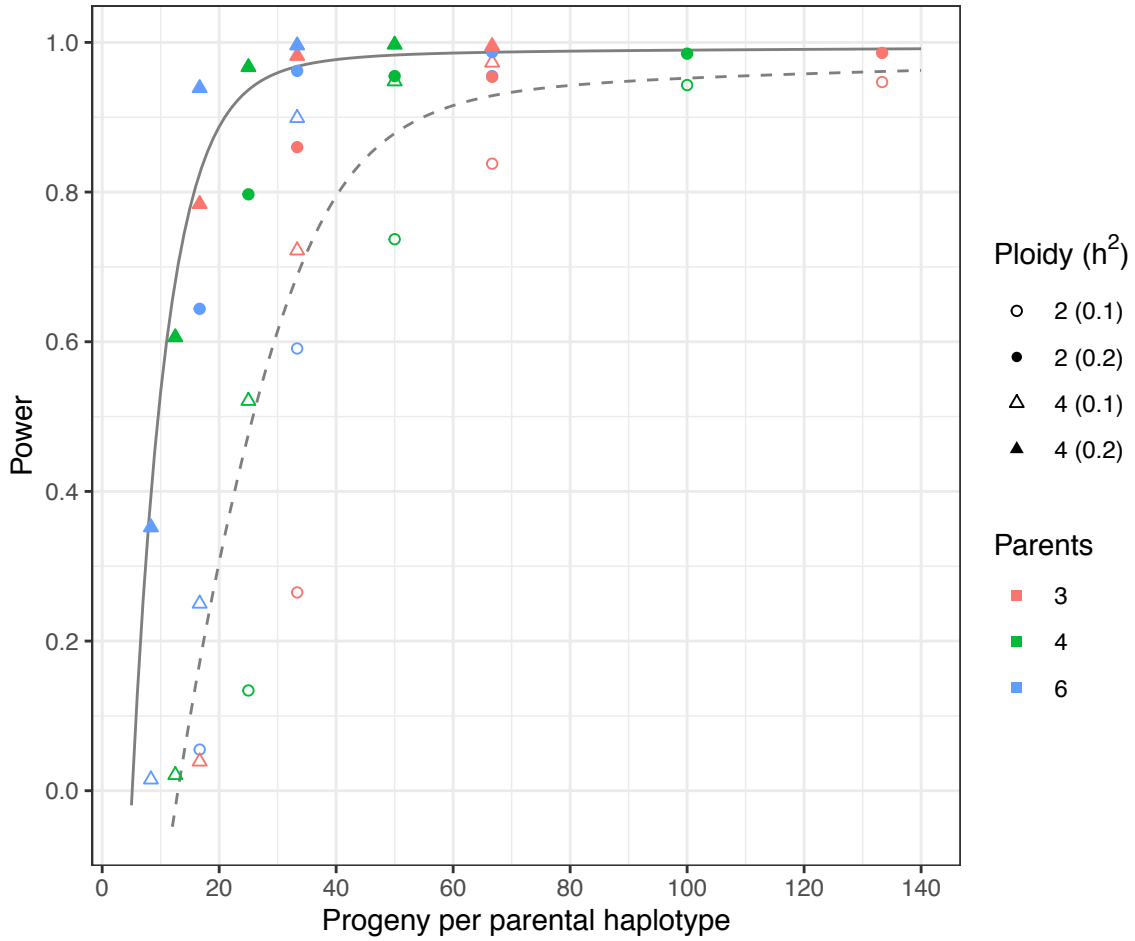

Figure 3: Statistical power to detect QTL when the false positive rate is $\alpha = 0.05$, as a function of the number of progeny per parental haplotype, number of parents, ploidy, and QTL heritability (h^2). Each point is the average of 1,000 simulations with an additive model. The dashed line is a monotone increasing, concave spline for $h^2 = 0.1$, and the solid line is the spline for $h^2 = 0.2$.

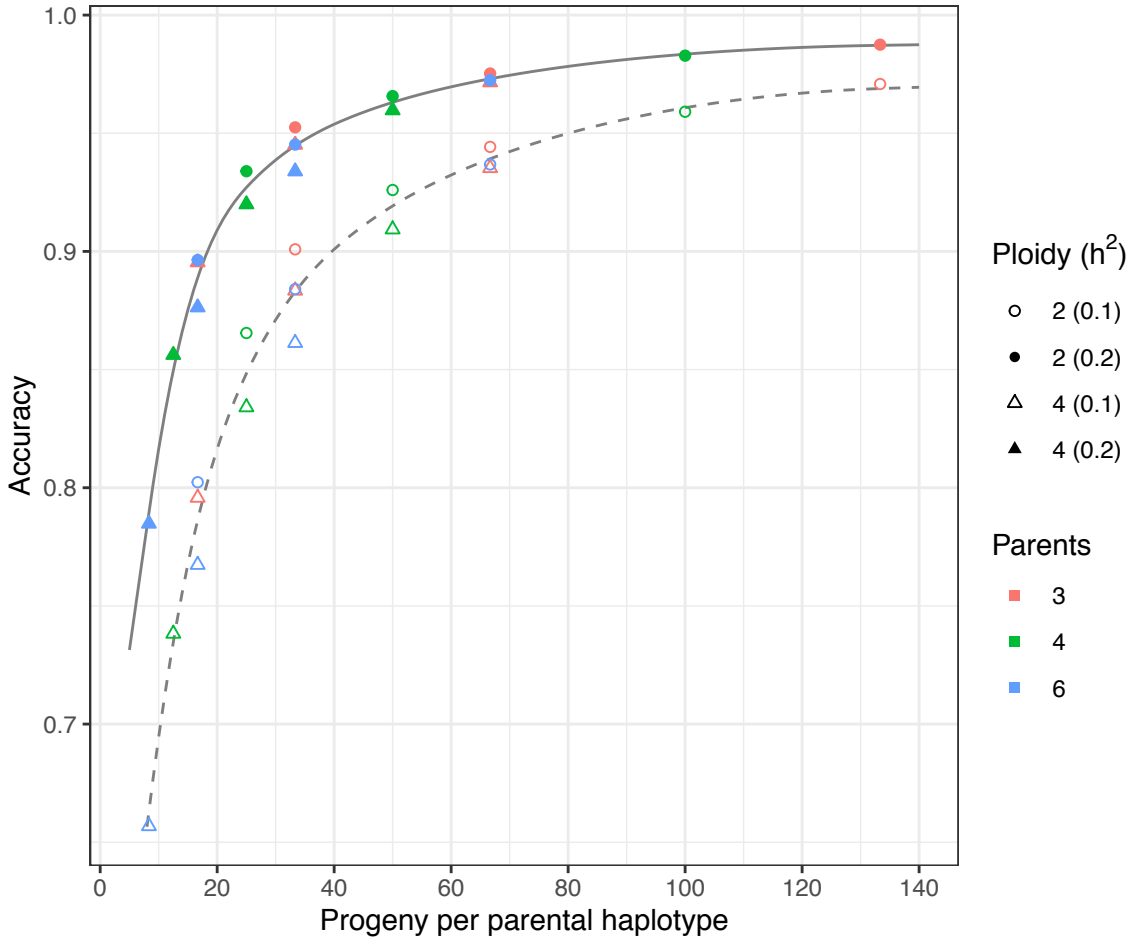

Figure 4: Accuracy of the estimated haplotype effects, as measured by the Pearson's correlation between the posterior mean and simulated values. Each point is the average of 1,000 simulations with an additive model. The dashed line is a monotone increasing, concave spline for $h^2 = 0.1$, and the solid line is the spline for $h^2 = 0.2$.

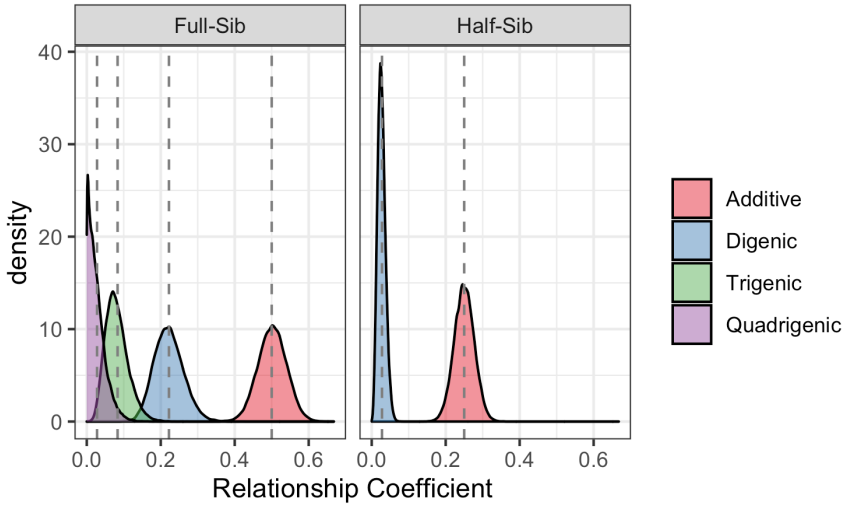

Figure 5: Distribution of realized relationship coefficients (Equation 4) between potato full-sibs and half-sibs in a 3×3 half-diallel population. The vertical dashed lines are the expected values based on pedigree (Table 1).

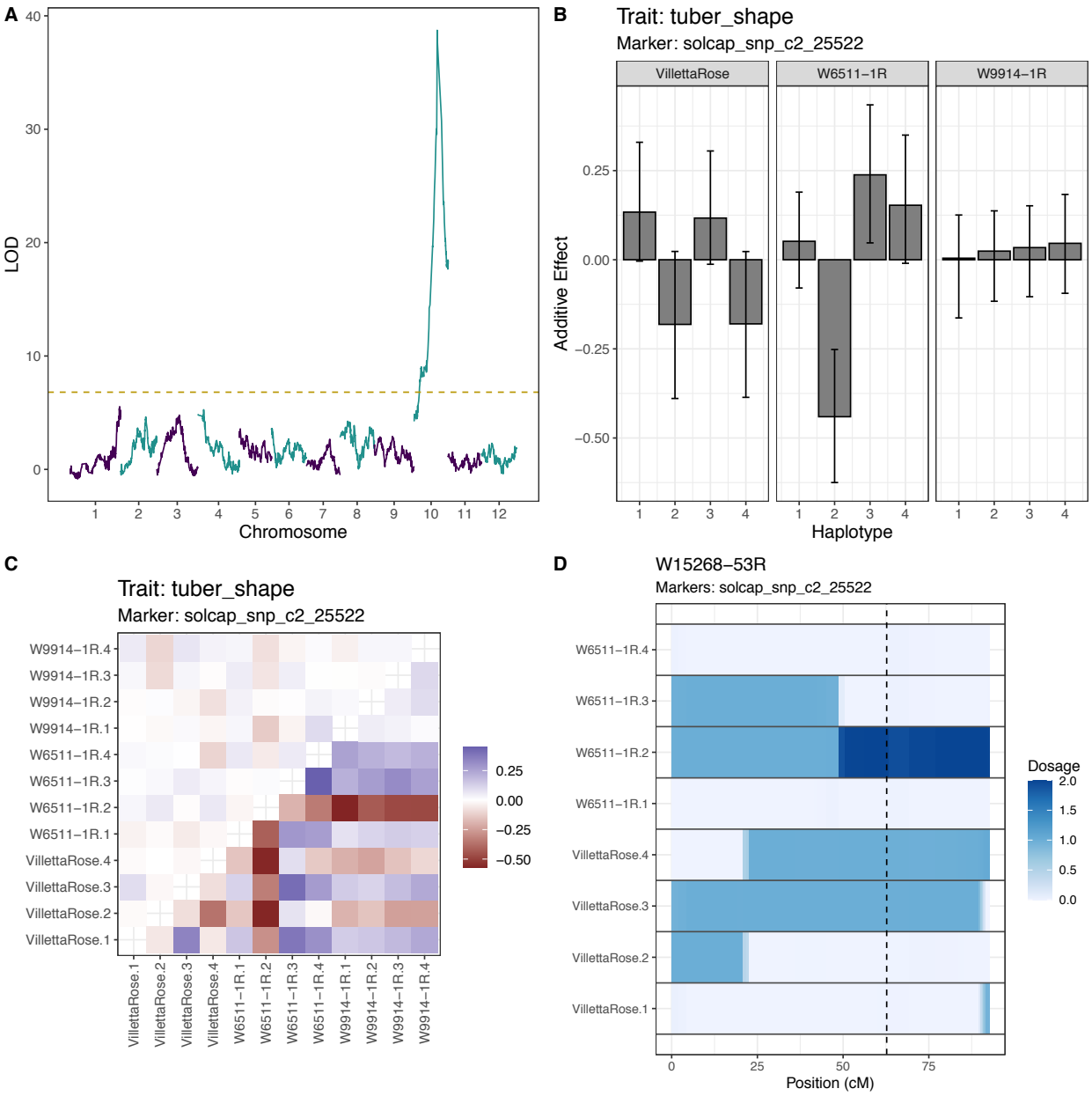

Figure 6: Figure 6: R/diaQTL results for potato tuber shape using a half-diallel population with three parents: VillettaRose, W6511-1R, and W9914-1R. (A) Genome scan with *diaQTL::scan1*. The most significant marker was solcap_snp_c2_25522 on potato chromosome 10. (B) Additive effect estimates for the 12 parental haplotypes using *diaQTL::fitQTL*. Error bars are the 90% CI. (C) Diplotype effect estimates using *diaQTL::fitQTL*, where the haplotypes for each parent are labeled .1 to .4. Values above the diagonal are the digenic dominance effect, and values below the diagonal are the sum of the additive and digenic effects. D) Parental haplotype visualization of chromosome 10 for one F_1 offspring (W15268-53R) from the parents VilletaRose and W6511-1R, using *diaQTL::haplo_plot*. The color shows the expected haplotype dosage, and the dashed vertical line is the position of marker solcap_snp_c2_25522. The figure shows there are two copies of the W6511-1R.2 haplotype at this locus due to double reduction.

Table 1: Expected vs. observed values for the realized relationship (Equation 4) between potato full-sibs and half-sibs in a 3 × 3 half-diallel population.

Order	Relationship	Expected	Observed
Additive	Full-sib	$1/2 = 0.500$	0.501
Additive	Half-sib	$1/4 = 0.250$	0.250
Digenic	Full-sib	$2/9 = 0.222$	0.222
Digenic	Half-sib	$1/36 = 0.028$	0.027
Trigenic	Full-sib	$1/12 = 0.083$	0.081
Trigenic	Half-sib	0	0
Quadrigenic	Full-sib	$1/36 = 0.028$	0.026
Quadrigenic	Half-sib	0	0

Table 2: Model fit and variance decomposition for potato tuber shape in a 3 × 3 half-diallel population, using models of increasing complexity.

Model[a]	# random effects for highest order term	ΔDIC^b	h^2_{poly}	h^2_1	h^2_2	h^2_3	h^2_4
Additive	12	-248.0	0.37	0.31	-	-	-
Digenic	78	-259.4	0.33	0.25	0.09	-	-
Trigenic	240	-260.2	0.32	0.25	0.07	0.05	-
Quadrigenic	300	-256.5	0.32	0.25	0.06	0.04	0.03

[a]Models are named based on the highest order effect but include all lower order effects.
[b]Change in the Deviance Information Criterion relative to the null model; lower values are better. Posterior means are reported for the proportion of variance for each QTL effect, denoted $h^2_{(z)}$ (Equation 5), and the polygenic effect h^2_{poly}. The additive relationship corresponds to $z = 1$, digenic dominance is $z = 2$, trigenic dominance is $z = 3$, and quadrigenic dominance is $z = 4$.

Supplemental Material for

QTL Mapping in Outbred Tetraploid (and Diploid) Diallel Populations

Rodrigo R. Amadeu[1], Patricio R. Munoz[1], Chaozhi Zheng[2], Jeffrey B. Endelman[3*]
[1]Horticultural Sciences Department, University of Florida, Gainesville, FL, USA
[2]Biometris, Wageningen University and Research, Wageningen, The Netherlands
[3]Department of Horticulture, University of Wisconsin, Madison, WI, USA
*corresponding author: endelman@wisc.edu

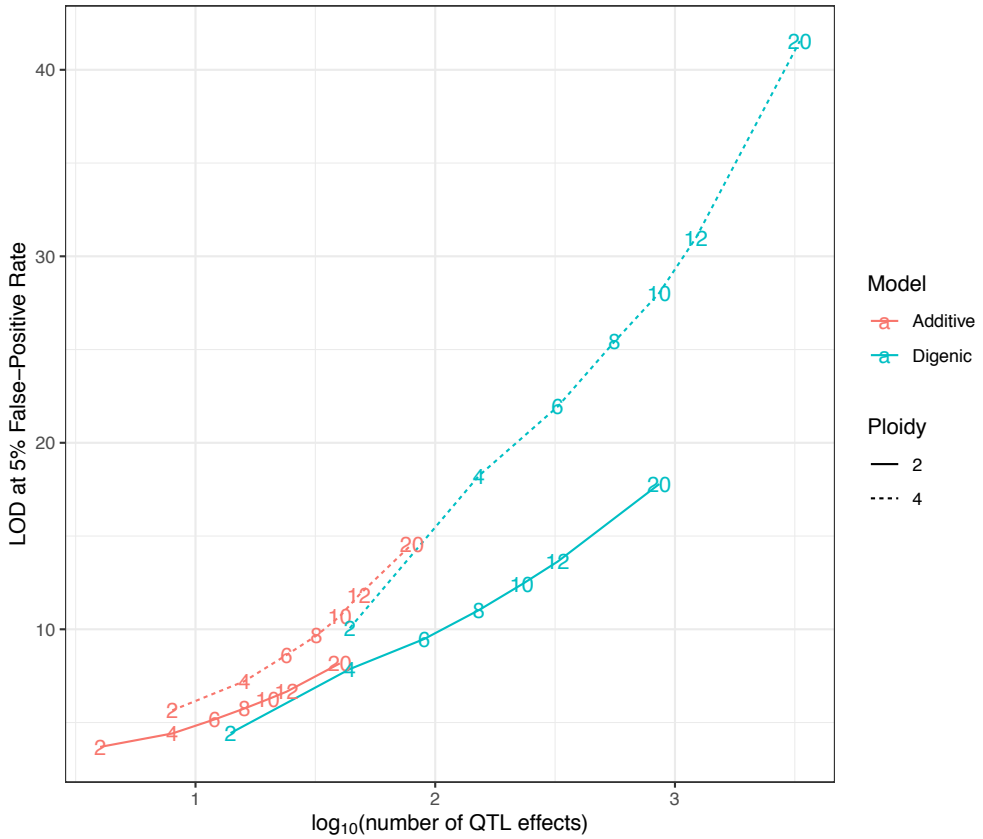

Figure S1: Comparing the LOD threshold needed to control the false positive rate at 0.05, for a genome of 10 Morgans. Numbers in the graphic represent the total number of parents used in the half-diallel simulation. Digenic model also has additive effects.

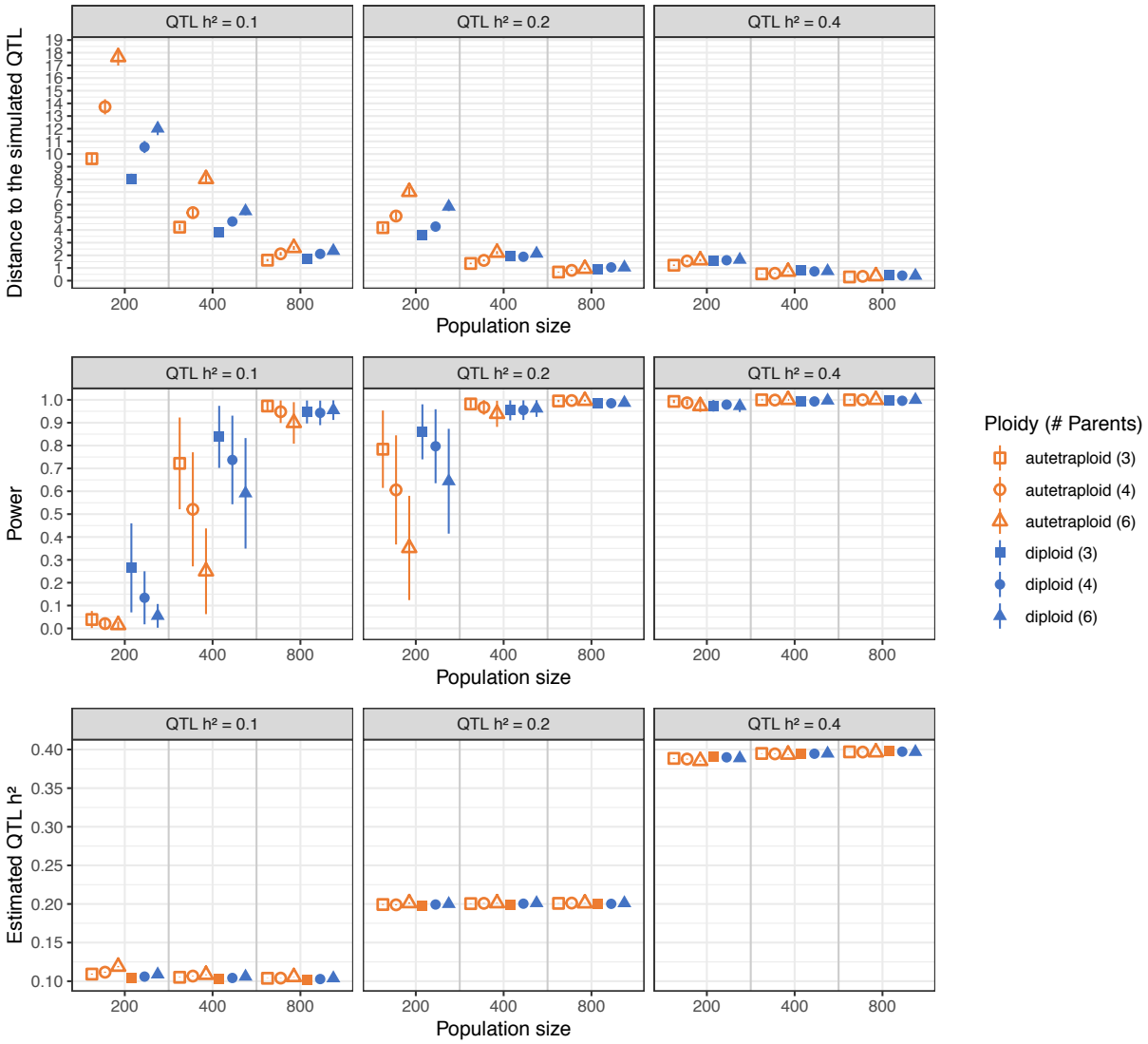

Figure S2: QTL mapping statistics for simulated data considering population sizes, QTL heritability, number of parents (3, 4, and 6) in a half-diallel, and diploid or autotetraploid species. The genome-wide false positive rate was controlled at 5%. A) Average distance in cM between the peak LOD score position and simulated QTL position. B) Average power to detect the QTL. C) Average estimated QTL heritability based on the diaQTL linear model. Error bars are standard errors. Each data point represents the average of 1,000 simulations.

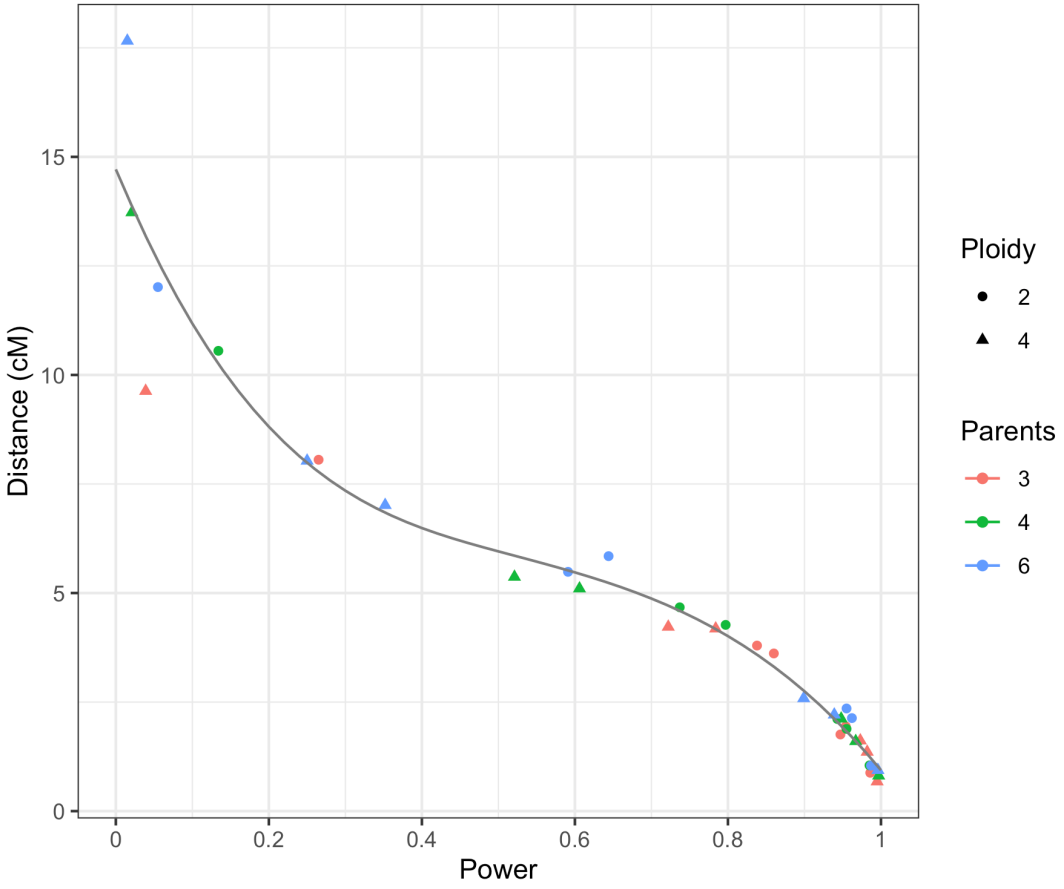

Figure S3: Average distance between the peak LOD score position and simulated QTL position, as function of power, ploidy and number of parents. Each point represents an average over the multiple population sizes and heritability values shown in Figure S1.

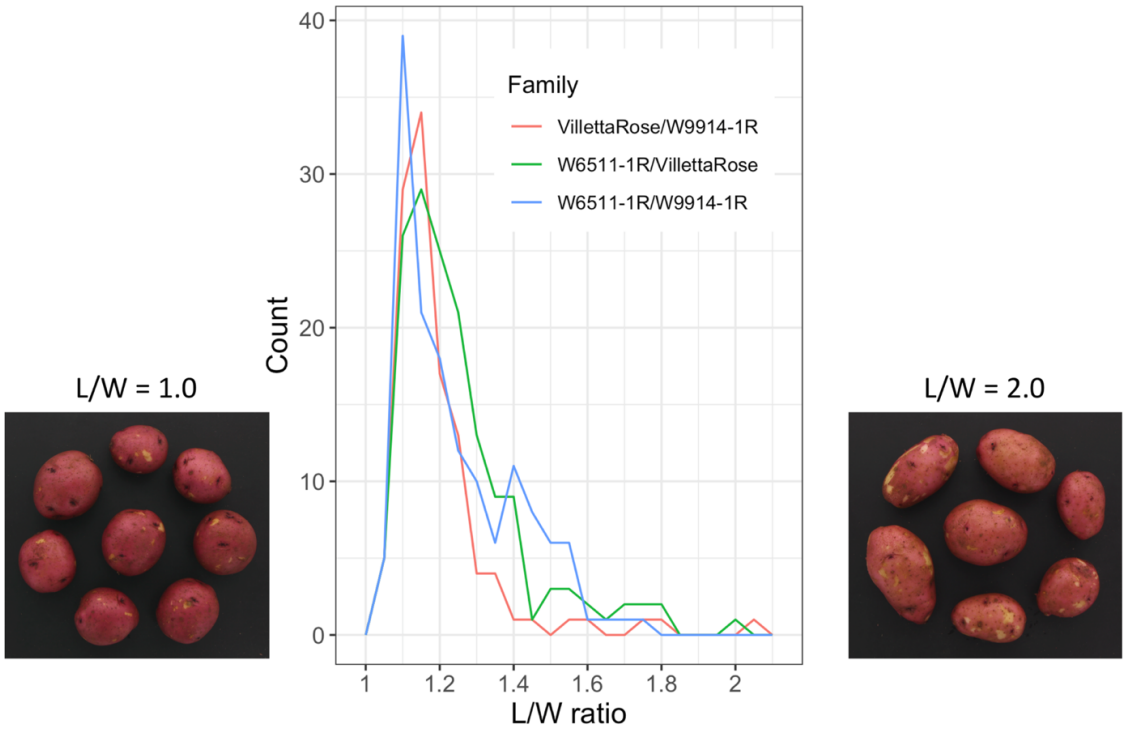

Figure S4: Distribution of the length to width (L/W) ratio for tubers of three full-sib families of autotetraploid potato. The images illustrate the extreme phenotypes of this population with L/W = 1.0 as a round tuber shape, L/W = 2.0 as an elongated tuber shape.

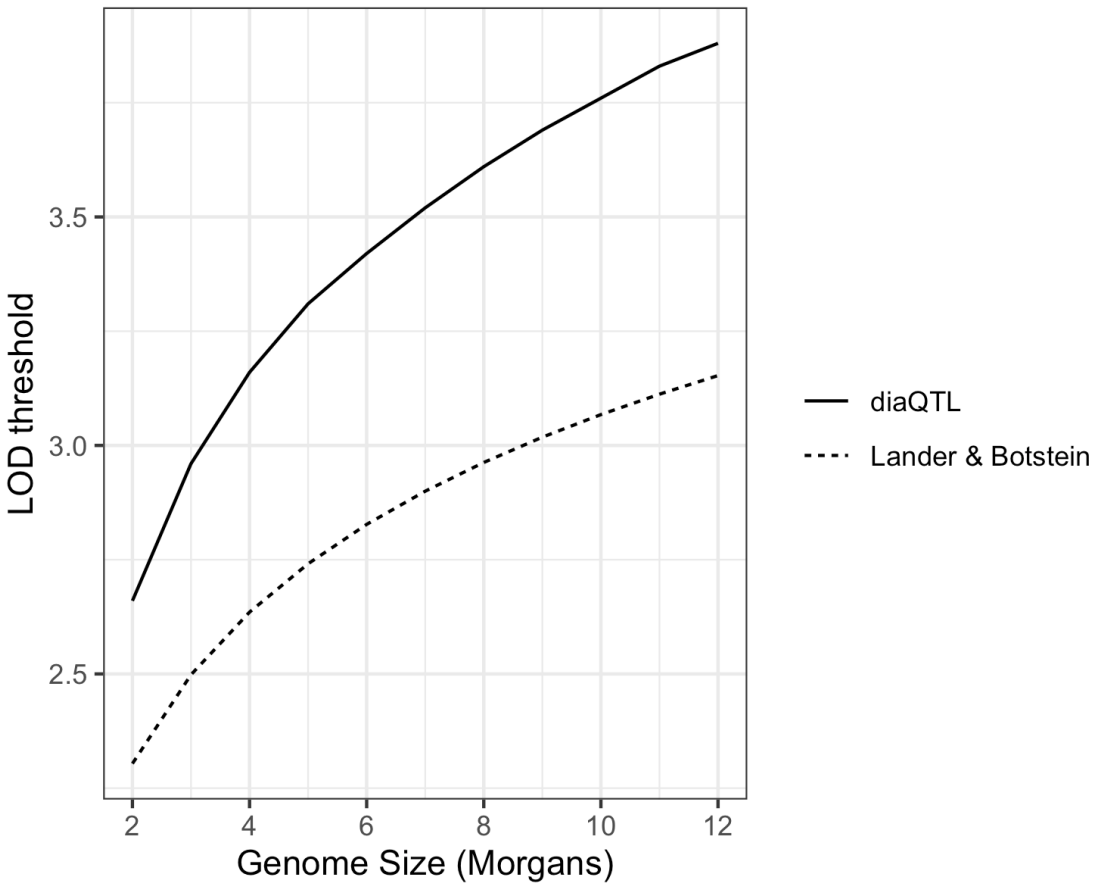

Figure S5: LOD threshold to control the genome-wide false positive at 0.05, comparing the result with diaQTL for an F_1 population with two outbred parents vs. Proposition 2 of Lander and Botstein (1989) for a backcross population from two inbred parents and chromosomes of length 1 Morgan.

References

Lander, E. S. and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185–199.