

RESEARCH

Open Access



Multiallelic models for QTL mapping in diverse polyploid populations

Alejandro Thérèse Navarro, Giorgio Tumino, Roeland E. Voorrips, Paul Arens, Marinus J. M. Smulders, Eric van de Weg and Chris Maliepaard*

*Correspondence:
chris.maliepaard@wur.nl
Plant Sciences Group,
Department of Plant
Sciences, Wageningen
University and Research,
Droevendaalsesteeg
1, P.O. Box 386, 6700
AJ Wageningen, The
Netherlands

Abstract: Quantitative trait locus (QTL) analysis allows to identify regions responsible for a trait and to associate alleles with their effect on phenotypes. When using biallelic markers to find these QTL regions, two alleles per QTL are modelled. This assumption might be close to reality in specific biparental crosses but is unrealistic in situations where broader genetic diversity is studied. Diversity panels used in genome-wide association studies or multi-parental populations can easily harbour multiple QTL alleles at each locus, more so in the case of polyploids that carry more than two alleles per individual. In such situations a multiallelic model would be closer to reality, allowing for different genetic effects for each potential allele in the population. To obtain such multiallelic markers we propose the usage of haplotypes, concatenations of nearby SNPs. We developed “mpQTL” an R package that can perform a QTL analysis at any ploidy level under biallelic and multiallelic models, depending on the marker type given. We tested the effect of genetic diversity on the power and accuracy difference between bi-allelic and multiallelic models using a set of simulated multiparental auto-tetraploid, outbreeding populations. Multiallelic models had higher detection power and were more precise than biallelic, SNP-based models, particularly when genetic diversity was higher. This confirms that moving to multi-allelic QTL models can lead to improved detection and characterization of QTLs.

Key message: QTL detection in populations with more than two functional QTL alleles (which is likely in multiparental and/or polyploid populations) is more powerful when using multiallelic models, rather than biallelic models.

Keywords: Polyploidy, Multiparental, QTL, Multiallelic, Genetic diversity

Introduction

Quantitative trait locus (QTL) analyses are those experiments in which a population is genotyped with many markers that cover the whole genome, and phenotyped for traits of interest. Once that is done, positions along the genome are tested for association, either defined by the markers or by some clever estimate such as those used in interval mapping [1, 2]. QTL studies have been extremely useful in unravelling genomic regions that control or contribute to important plant traits such as disease resistance, yield, crop quality or tolerance to abiotic stresses. The precision of these studies has been improved by the advent of high-throughput technologies, that facilitated genotyping of thousands



to millions of Single Nucleotide Polymorphisms (SNPs) in a single analysis. This is nowadays also possible in polyploid organisms, thanks to statistical and computational developments in the areas of genotyping, linkage map construction and QTL analysis [3, 4].

When trying to find QTLs two aspects will define the outcome obtained: the type of population studied, and the QTL modelling approach chosen.

Population types

A classical population type is the biparental cross, a population of siblings obtained by crossing two parents, usually contrasting in the trait of interest. If both parents are homozygous, as is the case in many self-fertilizing species, QTLs found in this type of population will reflect the allelic differences between the two parents. If the parents are diploids, there will likely be only two alleles per QTL segregating in that population. Since the cross contains only a small fraction of the genetic diversity of the species, QTL results from these populations may not be applicable to other populations and markers linked to QTLs cannot easily be used in other crosses.

Another possibility is to use a genome-wide association study (GWAS), in which a large set of diverse individuals are studied, and thus a large number of QTL alleles is expected to segregate. Unlike in biparental crosses, an association between markers and QTLs is expected due to Linkage Disequilibrium (LD) rather than direct family linkage. These studies produce more widely applicable QTL results, but introduce some drawbacks: (1) rare allele variants, which will be present at low frequency in a GWAS panel, will easily be missed even if they affect the phenotype, and (2) linkage disequilibrium (LD) is not spread homogeneously across the population or the genome, an effect known as “genetic structure”, and this may generate false positives if not taken into account [5, 6].

Nevertheless, as described in [7], mapping in biparental populations or GWAS panels represent two extremes of a genetic diversity gradient. An intermediate form can be found in multi-parental populations (MPP). An MPP is formed by individuals that share a limited number of known ancestors, for instance, a set of connected biparental crosses, or multiple lines originating from a small set of founders. As such, the number of QTL alleles will be at most of $ploidy \times founders$. Additionally, as the genetic structure in an MPP originates from mostly known pedigree relationships, it will be less complex than that of GWAS populations, and the allele frequencies will often be more balanced.

The MPP concept fits well the type of populations usually available in breeding programmes, where multiple crosses are made with some interesting parents. Breeding populations become then ad-hoc MPPs and instead of analysing each cross separately, the whole breeding program could be analysed at once, increasing statistical power. The idea that utilizing breeding populations for QTL analysis might be a better option than creating specific experimental populations has been studied previously [7–10], although in diploid species under biallelic models.

Modelling approaches

The type of mathematical model used for QTL analysis will heavily depend on the population under study. In a classical biparental population an analysis of variance (ANOVA)

will easily provide accurate QTL estimates. In contrast, in a GWAS panel, genetic structure must be taken into account, usually in the form of a mixed model [6]. In the case of a MPP, a similar mixed model could be used, although if the genetic structure is simple enough, a fixed factor accounting for subpopulations may perform well also [6].

The number of modelled QTL alleles is also relevant. Typically, since biallelic markers are used, two alleles per QTL are modelled. Assuming the presence of only two alleles, however, is sensible under very few scenarios. As ploidy, heterozygosity or the number of founders of a population increase, the number of expected QTL alleles rises. The larger the number of alleles, the less realistic the biallelic model becomes for describing the observed variance. Nevertheless, as SNP markers have become the standard polymorphism in modern genotyping, using them directly implicitly tests a biallelic scenario. However, SNP information can be used differently. By combining adjacent SNPs, biallelic SNPs can be turned into multiallelic haplotype markers [11].

Due to the increased genetic diversity present in GWAS and MPP populations, it is foreseeable that moving to multiallelic QTL models will provide a gain in statistical power. Nevertheless, biallelic models are simpler and thus more powerful, and they have a long trajectory of success. There is currently no software available that can perform multiallelic QTL analyses in polyploid populations in the presence of genetic structure, but such software is being developed. Under which circumstances, if any, will a genetically diverse population benefit from a multiallelic QTL modelling approach?

To answer this question, we have simulated a series of autotetraploid MPPs with different levels of genetic diversity. Populations were designed following the Nested Association Mapping (NAM) structure, where one central parent is crossed with many peripheral parents [12]. We adapted the QTL modelling approach presented in [13] for diploid MPPs with inbred founders, expanding it to a polyploid and heterozygous case. We present this approach as an R package [14] named “mpQTL” to perform QTL analysis. This package together with the simulated MPPs allowed us to assess the effect of biallelic or multiallelic markers on QTL detection and QTL precision under different genetic diversity scenarios.

Materials and methods

Statistical models

Mixed models allow to correct for dependence between observations due to genetic structure. Yu et al. (2006) defined a “unified mixed model”, also known as the $Q + K$ model [4], that can accommodate both a population structure matrix (Q) and a kinship matrix (K):

$$y = X\beta + Qv + Zu + \epsilon \quad \text{Var}(\mathbf{u}) = K\sigma_G^2 \quad \text{Var}(\epsilon) = R\sigma_\epsilon^2 \quad (1)$$

where y is the vector of phenotypic trait values, $X\beta$ represents the incidence matrix and marker effects (SNP effect in [6]); Qv are the population structure matrix and vector, respectively; Zu are design matrix and vector of genetic background effects (polygene component in [6]); and ϵ is the residuals vector. The variances of the random effects, \mathbf{u} and ϵ are also defined: K is the kinship matrix and σ_G^2 , the genetic variance; R is a matrix

with off-diagonal numbers being 0 and the diagonal is the reciprocal of the number of observations underlying each genotype estimation, and σ_e^2 is the residual variance.

Fixed term: allele parametrization

Definition of X requires a genetic model, that is, a method to transform genetic data into an incidence matrix X . Polyploid genetic models have existed for a long time [15] and have inspired more recent versions applied to SNP data [16, 17]. The simplest of them is the *biallelic model* (model B in [7], association mapping in [18]), which considers SNP alleles as equal to QTL alleles. In a biallelic model, the SNP dosages are used to predict genetic effects, giving the $X\beta$ term the following form:

$$X_b\beta = \begin{bmatrix} 1 & \delta_1 \\ 1 & \delta_2 \\ \vdots & \vdots \\ 1 & \delta_n \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} \quad (2)$$

where δ_i are the dosages (a value from 0 to *ploidy*) of one of the SNP alleles, μ is the intercept and β the genetic effect of that SNP allele. We denote the incidence matrix as X_b for this modelling strategy. Note that this represents an additive model without intra or inter-locus interaction, i.e. no dominance or epistasis between alleles is modelled.

Alternatively, Identity-By-Descent (IBD) information can be used to generate an ancestral model [13], also known as a PBA model [10] or an LDLA model [9, 19]. Under the ancestral model, the dosage of each ancestral allele or haplotype in the NAM population is used to estimate genetic effects. The shape of the $X\beta$ term then takes the form:

$$X_a\beta = \begin{bmatrix} 1 & \delta_{11} & \delta_{12} & \dots & \delta_{1k} \\ 1 & \delta_{21} & \delta_{22} & \dots & \delta_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & \delta_{n1} & \delta_{n2} & \dots & \delta_{nk} \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad (3)$$

In this case, the dosages of all alleles *except one* (the reference allele) are specified. Therefore, k is the number of alleles -1 . Each β represents the additive genetic effect of each ancestral allele, relative to the effect of the reference ancestral.

Random term: kinship matrix calculation

In this model, a kinship matrix K is calculated using the *realized relationship* [4]:

$$K = \frac{DD^t}{\Delta} \quad \Delta = \overline{\text{diag}(DD^t)}$$

where D is a dosage matrix with markers on columns and individuals on rows, and the mean of each column is zero (column means have been subtracted for each column); and Δ is the mean of the diagonal of the DD^t matrix. If haplotypes are used instead of biallelic SNPs, D can consist of concatenated matrices similar to X_a (without the intercept column), so that the number of columns is equal to the total number of alleles present across all markers used. To mitigate the bias due to differences in marker density across

the genome, kinship estimates are calculated on a subset of evenly distributed SNPs (one marker per cM).

Haplotyping

Haploblocks were arbitrarily defined using a sliding window of 6 consecutive SNPs with an overlap of 4 SNPs (first haplotype is SNP1-SNP2-...-SNP6, second is SNP3-SNP4-...-SNP8). A haploblock of length 6 can tag a maximum of $2^6 = 64$ alleles if all combinations are present, although in our simulations the number of observed alleles was much lower, with the average number of observed alleles ranging from 11.23 in NAM1 to 21.8 in NAM10. To obtain a haploblock position, the average position of the 6 SNP markers was taken. Haplotypes were obtained from the simulated phased SNP genotypes generated by PedigreeSim.

Power study

Definition of QTL interval

Single marker QTL methods do not provide an estimate for the QTL interval, yet with a defined threshold and a genetic map one can interpret the p value distribution to obtain them. Since adjacent markers are not independent, and the closer to a true QTL position, the more significant the p -value becomes, one expects a chain of increasingly significant markers, pointing towards a true QTL position. Based on this, we define a QTL interval as a set of ordered markers above the significance threshold such that:

$$QTL = \{m_1, \dots, m_n\} \quad \text{where } d_{ij} < l$$

where d_{ij} is the distance between adjacent significant markers i and j , and l represents a *linking distance*. As a result, a QTL interval is defined by a chain of significant markers, where adjacent significant markers are at a distance smaller than l . Therefore, for each value of l we can define a set of detected QTL intervals. Since the choice of l is arbitrary, we performed power calculations with l from 0 to 10 cM in steps of 0.5 cM.

Significance threshold

To adjust for multiple testing, an empirical permutation threshold was calculated for each QTL analysis [20]. Thresholds were obtained with 100 permutations on a single population for each model, as threshold values did not change substantially between populations.

Power estimates

To evaluate the QTL models here presented we will use (1) QTL detection power, the probability of detecting a QTL position when present; (2) false positive rate, the probability of having a significant marker outside a QTL region; (3) QTL accuracy, the closeness of a QTL peak (position of maximum probability within an interval) to the true position and (4) QTL and marker precision, the probability that a significant QTL interval or marker is a true positive.

QTL detection power can be calculated as the proportion of true QTLs that are found by the model. While this is informative, one can easily increase detection power by increasing the number of false positives. To estimate the false positive rate, we must define the true negative markers (N). We considered as true negatives all markers

outside a 10 cM interval around our true QTL positions (5 cM above and 5 cM below). We then define as false positives (FP) those markers that are above the significance threshold (they have lower p values, higher significance) and are outside the 10 cM true interval. Lastly the false positive rate is calculated as FP/N .

The range of a QTL interval is defined by the positions of its leftmost and rightmost markers. QTL intervals will be considered *true positives* if the QTL range includes the simulated QTL position. All markers belonging to a true positive QTL interval are considered true positive markers, whereas the rest of significant markers present in other QTL intervals will be considered *false positives*. Isolated significant markers will be ignored.

Under this framework we can define detected QTLs, true QTLs, significant markers and true positive markers. We will use these values to calculate the precision (proportion of true positives over all positives) for both QTLs and markers.

$$QTL_{precision} = \frac{\text{true positive QTLs}}{\text{detected QTLs}}$$

$$\text{marker}_{precision} = \frac{\text{true positive markers}}{\text{significant markers}}$$

Finally, we considered the ability of a model to predict the position of QTL within an interval. We can define a QTL peak as the most significant marker within a QTL interval, as is done when applying logarithm of odds (LOD) scores. QTL accuracy can then be calculated as the average distance of a QTL peak in a true QTL to the true QTL position.

Power measures were calculated for each of the three models in 11 populations for each level of genetic diversity (total of 44 populations).

Implementation

All computations in this study were done in R [14].

Ridge regression using a restricted maximum likelihood procedure was used to obtain the mixed model estimates, which in this context are equivalent to the Best Linear Unbiased Predictions (BLUP) [21, 22]. Such calculations can be performed using the mpQTL package, where the solution algorithm, F-test approximation and p value calculation were based on the mixed.solve() function of the rrBLUP package [23].

To improve computational efficiency, the EMMAX/P3D approach was applied [24, 25], which approximates variance components once, and recycles these components at each marker position, reducing the amount of large matrix multiplications that must be performed.

Simulation

Multiparental population design and genotype simulation

Nested Association Mapping (NAM) populations were generated using PedigreeSim V2.0, a simulation software that can simulate not only diploid but also polyploid meiosis [26]. PedigreeSim generates genotypes given a genetic map, a pedigree and the

genotypes of the first generation (founders) of that pedigree. Simulations were performed using Haldane's mapping function, allowing only bivalents with random pairing and the parameter "NATURALPAIRING" set to 1.

To speed up the calculations, an adapted tetraploid potato genetic map was used [27] containing only the first five chromosomes (3509 markers representing 485 cM). The individuals used in this study were simulated in a two-stage process: firstly, ancestor individuals were generated and used to obtain ten separate populations (ancestral groups); secondly, from each ancestral group a set of NAM founders were chosen to obtain parallel NAM populations.

For each ancestral group (AG), 10 ancestor individuals were generated with random SNP scores at each marker. Each SNP position is also given an "IBD allele", unique to each homologue of each ancestor (even if the SNP state is the same). Each ancestral group has 10 founders, and thus a total of 40 IBD alleles will segregate in each AG. These alleles we will name *ancestral alleles*. Each ancestor is randomly crossed (without selfing, as potato is an outbreeder) to produce a first generation of 100 individuals, which will serve as parents of the second generation. This process was repeated for 50 generations, maintaining a constant generation size of 100 individuals. Finally, 100 individuals per AG were obtained as potential parents for the creation of NAM populations.

A NAM population consisted of one central parent crossed with nine peripheral parents, without any of the subsequent inbreeding that was originally proposed for NAM crossing scheme for selfing crops [12]. Each cross produced 50 offspring, thus totalling at 460 individuals per NAM. To simulate NAMs with different degrees of genetic diversity, parents were sampled from the same or from different AGs. A NAM1 contains parents from only one AG, while a NAM5 contains parents from 5 different AGs, with the same number of parents per group when possible. When the numbers of parents per AG was not equal the central parent always originated from the AG providing the most parents. For each level of genetic diversity, 11 populations were simulated. At the end of the process, the genotypes of each individual were obtained in terms of ancestral alleles (IBD alleles) and in terms of SNP dosages.

Phenotype simulation

Phenotypes were simulated based on the simulated genotypes: genotypic values were obtained by assigning genetic effects to the ancestral alleles at pre-defined QTL positions. Each individual will then harbour four QTL alleles at each QTL position and the final phenotype is equal to the added effects of all QTL alleles plus a normally distributed noise. No interactions between alleles in one QTL or among QTL loci were simulated, and thus additive phenotypes were obtained.

We considered a situation where three unique QTL positions (at chromosome 1, 67.88 cM; chromosome 2, 61.2 cM and chromosome 4, 100.49 cM) were segregating. Each AG has a random allelic mean, and allele effects are drawn from a normal distribution around that mean. Additionally, 50 small-effect QTLs were added randomly across the genome to simulate a polygenic effect.

For further information see Additional file 1.

Results

Population simulation

Ten Ancestral Groups (AGs) were simulated, each of them being founded with 40 different founder alleles. After 50 generations of random mating with a generation size of 100 individuals, each locus contained 8 to 20 founder alleles, with an average between 12.5 and 13.5 depending on the AG.

Parents from the last generation of AGs were used to obtain NAM populations. Different degrees of genetic diversity were simulated by sampling parents from the same or different AGs, thus producing genetic structure. This is visualized for one example in Fig. 1, which shows a heatmap of the relatedness matrix K and a Principal Coordinate Analysis (PCoA) plot of the same matrix. On the left, we see how cross 3, 4 and 5, derived from crosses between AG1 and AG2 ($A1 \times A2$ in Fig. 1), have a higher relatedness between them than with any other cross. Similarly, in the PCoA plot we observe how the individuals from these crosses (light blue dot cloud) cluster together in the midpoint between X (from AG1) and the three parents B (from AG2). These indications confirm that our two-step approach was successful in generating NAM populations with genetic structure. A similar outcome can be observed in the NAM1 to NAM10 simulations.

Population comparison

For each level of diversity, 11 populations were tested with the three proposed models. In almost all cases, all models were able to detect all QTL regions. Regardless of the linking distance used for QTL estimation, lower diversity resulted in higher detection power (Table 1). This can be observed at $l = 3$ using haplotype markers: NAM1 has a detection power of 1 (all QTLs were found in the 11 populations), but this power decreases to 0.818 in NAM10. Similarly, the false positive rate decreases as diversity increases and is lowest in the SNP model than in IBD or haplotype models. In Fig. 2 the 99th percentile profiles also highlight the increased power in lower diversity populations, where the dark blue line representing NAM1 populations had higher significance values for all

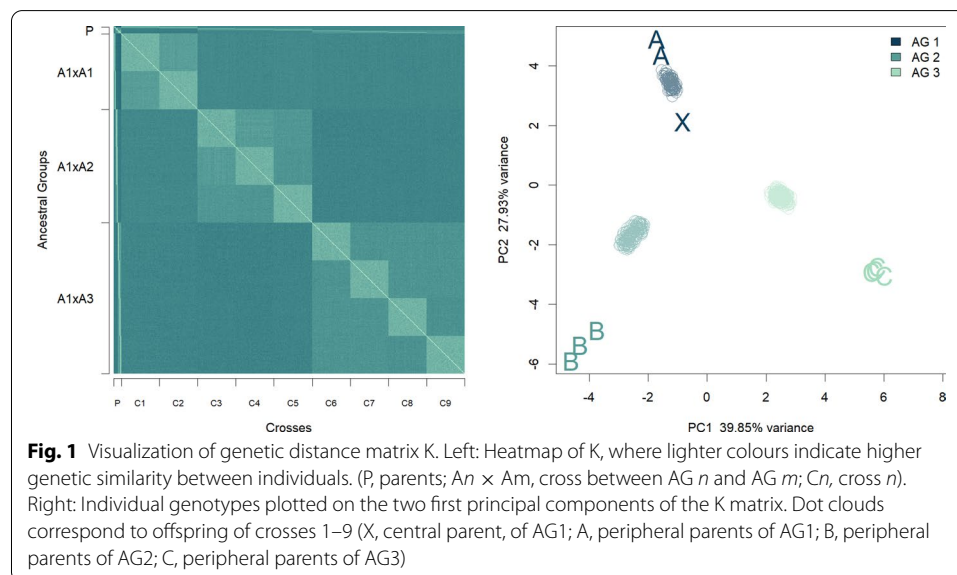
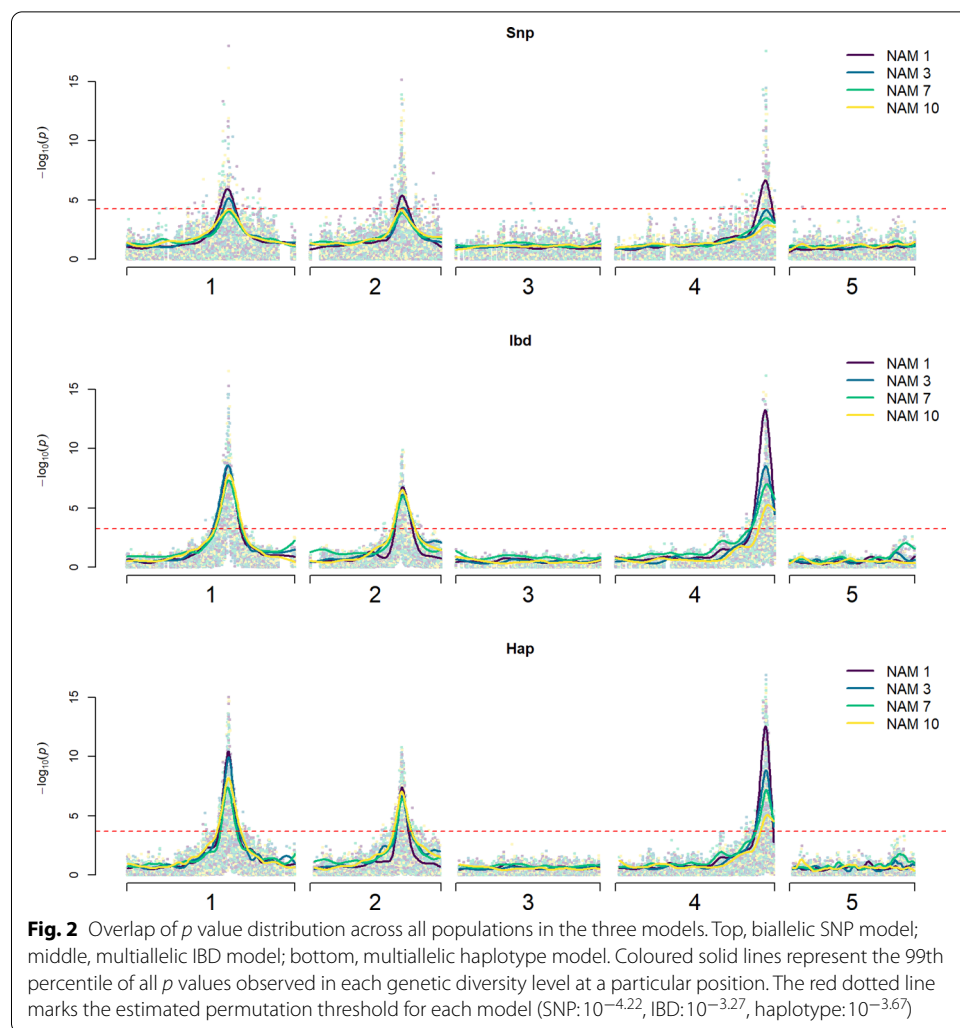


Table 1 Power comparison across genetic diversity and marker types. Each estimate is an average of 11 populations for each diversity level, with $l = 3$ cM. SNP refers to the biallelic model, IBD refers to the ancestral, multiallelic model and Hap refers to the haplotype-based approach. For detection power and QTL precision, higher numbers indicate a better model, while for false positive rate and accuracy, lower numbers indicate a better model

	Detection power			False positive rate			QTL precision			Accuracy (cM from true position)		
	SNP	IBD	Hap	SNP	IBD	Hap	SNP	IBD	Hap	SNP	IBD	Hap
NAM1	0.939	1	1	0.012	0.066	0.055	0.917	0.850	0.941	0.593	0.161	0.121
NAM3	0.909	0.970	0.970	0.008	0.065	0.054	0.865	0.814	0.886	0.550	0.192	0.130
NAM7	0.545	0.939	0.909	0.005	0.064	0.040	0.697	0.842	0.879	0.687	0.325	0.331
NAM10	0.606	0.848	0.818	0.005	0.055	0.038	0.773	0.932	0.850	0.665	0.312	0.621



QTL peaks and for all models. As diversity increases, a similar decrease can be observed for QTL precision. Finally, the mean peak distance from the QTL peak to the true QTL position was also larger (lower accuracy) at a higher level of diversity in the populations (Table 1).

Marker comparison

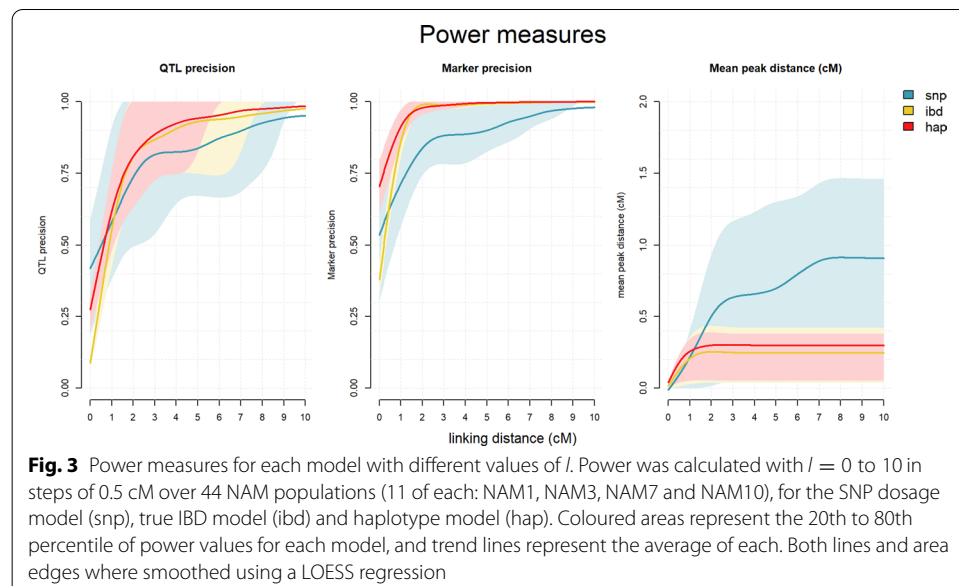
Across NAM populations and at a linking distance (l) of 3 cM, detection power averaged at 0.74 for SNPs, 0.93 for IBD and 0.92 for haplotypes and was stable for $l > 1$ cM. The decrease in detection power as genetic diversity increased was markedly larger in the SNP models than in the multiallelic models (Table 1). This can be clearly observed in the 99th percentile lines in Fig. 2: when diversity increases, the trend line is below the significance threshold in the SNP models, while for both multiallelic models all trend lines stay well above their respective thresholds. In Fig. 3 left and centre panels, we can see how the proportion of true positives increases as the value of l increases. For $l > 1$ cM, QTL precision is on average higher for multiallelic models (0.91 IBD, 0.92 haplotype) than for the SNP model (0.86). Marker precision is also higher for the multiallelic models (0.99 IBD, 0.99 haplotype, 0.92 SNP). The choice of l has an impact on this difference, as for lower values of l (but above 1) precision is much lower for the SNP model. This is due to the presence of significant markers further away from the true QTL position in the SNP model than in the multiallelic models (Fig. 4).

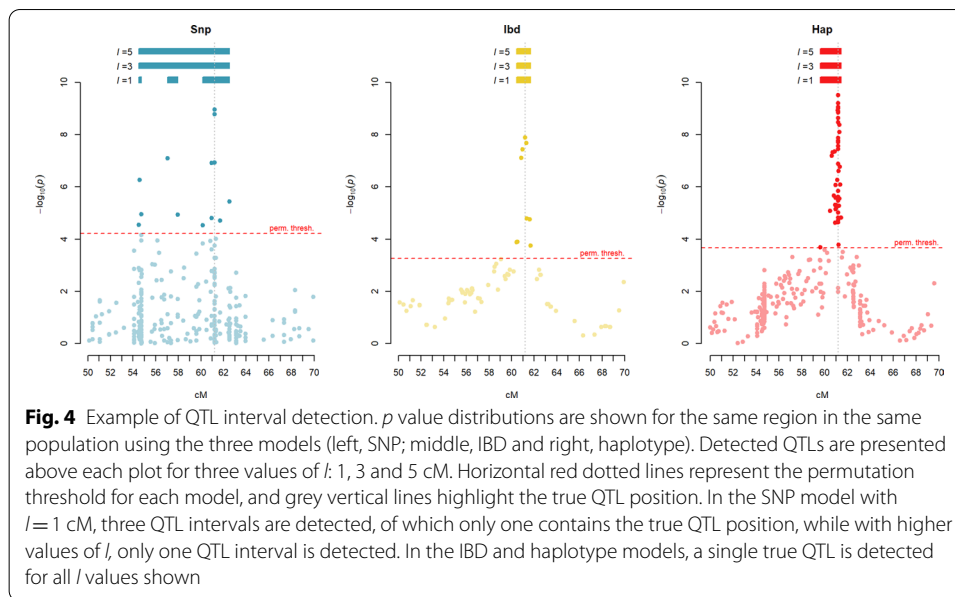
Peak accuracy (Fig. 3, right panel) is stable from $l > 1$ at 0.25 cM for IBD and 0.30 cM for haplotype models. In the SNP model, peak accuracy is lower and shows more variation. At $l = 1$ peak accuracy is similar to the IBD and haplotype models, yet many false positives are present in the QTL analysis (see Fig. 4). At higher l , average peak distance increases from 0.33 cM at $l = 2$ to 0.83 cM at $l = 7$.

Discussion

Model comparison

The essence of a QTL study is the genetic linkage between observed markers and unobserved QTL alleles. When dense genetic maps are used, the purpose of a QTL model should be to obtain an increasing marker significance as the analysis approaches a true QTL position. The definition of QTL interval used in this study stems from such





reasoning: we expect a chain of contiguous significant markers that form a peak structure, pointing towards the true QTL position.

Classical QTL experiments were carried out on inbred diploid experimental crosses. In this setup one can expect only two alleles per QTL to segregate, and thus biallelic SNP markers are able to uniquely tag each allele. In this context, a SNP regression is equivalent to testing the difference in phenotype due to having 0, 1 or 2 copies of each marker allele [1, 28]. However, when we move to scenarios where more than two alleles per QTL are expected to segregate at a single locus, for instance when heterozygosity is expected to be high or in multiparental populations, single SNPs no longer tag QTL alleles uniquely. Thus, each SNP allele might tag more than one functional QTL allele, creating a situation where the regression test is being performed between groups that do not represent a unique effect. Only if, by chance, those groups happen to divide functional alleles between those with large effects and those with small effects, will SNP markers be significant. Since two factors are affecting the significance of biallelic markers (i.e. distance to the true QTL position and the grouping of multiple effects), they become worse at estimating the true QTL position.

Figure 4 illustrates this situation. The three panels represent the same population being analysed with the three models presented in this study. It can be seen how in the SNP model there are three significant markers at approximately 54.5 cM, while the true QTL position is at 61.2 cM. Meanwhile there are quite some markers near the true position that are not significant. Such behaviour is not seen in the multiallelic models where markers near the true QTL position form a clear peak and more distant markers show no significance.

The consequences of this can be seen in Figs. 2 and 3. First, SNP models have overall lower significance at the QTL regions (Fig. 2), an effect that is increased when genetic diversity increases and biallelic markers become increasingly worse at tracking the multiple effects present in the population. This explains the lower detection power of

biallelic models when genetic diversity is increased (Table 1). Secondly, we see how at low linking distances, SNP models have a high number of significant markers in false-positive QTL intervals (Fig. 3 middle). As l is increased, marker precision increases (there are less false-positive QTLs), but at the cost of accuracy (Fig. 3 left): the QTL intervals become larger (Fig. 4), including markers at some distance of the QTL position with higher significance than those at the simulated QTL position.

Thus, in a context of high genetic diversity, the usefulness of SNP models will depend on marker density, as higher density gives higher chances of having at least one marker at the QTL position that divides functional QTL alleles in two groups with statistically different means. Even if such a marker is found and the location of the QTL is detected, the effect estimated by a regression model does not realistically represent the true functional alleles present in the population.

Considering the lower detection power, lower accuracy and inability of biallelic QTL models to estimate effects for multiple alleles, it is clear that SNP-based biallelic models are a limited and limiting tool when applied to multiallelic populations.

Multiallelic markers

In order to apply multiallelic models, one must be able to obtain multiallelic genotypes. One possibility is to utilize markers that are multiallelic per se, such as SSR markers, but these markers are less common along the genome, their detection cannot be automated, and they are therefore hard to apply within high-throughput pipelines.

Alternatively, several studies have proposed the use of multiallelic haplotypes: groups of phased adjacent SNPs. This type of markers has the advantage of being predictive of two parts of IBD: family IBD, regions of chromosomes from two individuals that originate from the same *parental chromosome*; and ancestral IBD, chromosomal regions originating from the same *ancestral chromosome* that could occur in more than a single founder and that are broken down by recombination events [29].

While in our simulations haplotyping was trivial because the genotype of each individual was known, haplotyping of real SNP data requires *phasing*. For instance, if two adjacent marker genotypes of an individual are *AAAB* and *AAAB*, the underlying four haplotypes could be both *AA-AA-AB-BA* or *AA-AA-AA-BB*. Some approaches have been developed for haplotyping in polyploids [30–32] but regardless of the method, haplotype estimation from SNP data carries a certain degree of uncertainty due to the high number of possible solutions with similar probabilities. This uncertainty is not present in the haplotypes used in this study, meaning that the haplotype model here presented is performing better than what would be expected with real data, depending on the accuracy of haplotype estimation.

Nevertheless, sequencing technologies are becoming a mainstream approach for genotyping, and haplotypes can be directly observed in longer sequencing reads. Identifying haplotypes for different individuals given a set of reads is a complex mathematical problem that has spurred the development of a variety of tools [31, 33–35]. The haplotypes obtained from these methods could also be used with the multiallelic polyploid model introduced in this paper, allowing to perform QTL analysis in genetically diverse polyploid populations based on sequence data.

Lastly, in this simulated population each founder allele had a different QTL effect. In nature this might not be the case, as it is well known that many mutations are in fact neutral and thus do not change the QTL effect of that mutated allele. This could imply that the number of haplotypes would be higher than the number of QTL effects in a population, thus decreasing the usefulness of haplotype-based multiallelic markers.

Preparing multiparental populations

When organizing an MPP, the power to be able to detect the effects of an allele at a QTL depends on its frequency. The more individuals harbour one QTL allele, the more information the MPP provides about it. The expected frequency of founder alleles is directly affected by two factors: founder genetic diversity and offspring per founder.

The number of alleles segregating in a population is a direct reflection of the genetic diversity of its founders. When relatedness between founders is high, the chances of two founder chromosomes harbouring the same allele is also high. In MPPs where founders are very related, ultimately not many alleles can be expected to segregate. In contrast, when relatedness between founders is low, they have high chances to contribute unique alleles. The approach here presented estimates one parameter per each allele in the population, and thus, if population size is maintained constant, the power of the model decreases as the number of alleles increases. This hypothesis was confirmed by our simulation study where systematically, higher diversity populations, which require more allele effect parameters, presented lower QTL detection power, lower precision and lower QTL accuracy (Fig. 2, Table 1).

A second aspect to be considered is the number of offspring per founder. The larger the contribution of a founder to the individuals of the MPP, the higher the power to detect and estimate the effects of its alleles [36]. For instance, using our NAM design, the alleles present in the central parent were present in all crosses. Alleles from peripheral parents not shared with the central parent had fewer individuals contributing to their effect estimation, meaning these estimations will be less powerful.

Considering the previous points, we suggest that MPPs should be developed with an intermediate diversity and ensuring that those alleles to be studied are kept at a relatively high frequency. Following this logic, a few parents from the same ancestral group (AG) can be selected (which likely share some alleles) and crossed with several other AGs. If all AGs are equally interesting for the QTL study, then all AGs should have a similar contribution to the offspring [36]. If an MPP is designed from an already-existing set of connected F1 crosses, then each cross should be of similar size and the number of crosses per AG should be similar. When more complex pedigrees are used, ancestry coefficients can help guide the design of MPP.

Conclusion

Genetic diversity is the basis of breeding, and thus, characterizing it becomes essential in the development of new varieties. The methods developed within the “mpQTL” package add to the growing toolset for polyploid organisms. It is now possible to apply multiallelic models in polyploid organisms in the presence of genetic structure, which we have shown are more powerful, especially in the presence of high genetic diversity.

Additionally, this study supports an alternative approach to the study of genetic diversity. Instead of using a diversity panel to perform a GWAS, a selection of these diverse accessions can be used as founders of an MPP. Each biparental cross within the MPP will add information to the QTL study, and future crosses can be added to the overall MPP analysis. This approach shows much promise in the context of breeding, particularly for its ability to connect and share information between crosses that in traditional approaches would remain separate.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-022-04607-z>.

Additional file 1. Allelic effect sampling and phenotype simulations.

Acknowledgements

The contribution of all breeding companies involved in the project, and the Top sector Horticulture and Starting Materials (<https://topsectortu.nl/nl>) is highly appreciated.

Authors' contributions

Conceptualization of the research was made by A.T.N., G.T., R.E.V., and C.M. Data simulation and statistical analysis was performed by A.T.N. with substantial help from G.T. Supervision during the research was granted by G.T., R.E.V. and C.M. The first draft of this paper was written by A.T.N. Extensive review was granted by G.T., R.E.V., P.A., M.J.M.S., E.vdW. and C.M. Funding was acquired by P.A. and C.M. All authors have reviewed and approved this manuscript.

Funding

This research was funded under the TKI project: "Novel genetic and genomic tools for polyploid crops" (KV 1605-020; BO 26.03 009 004, BO 50 002 022). The funding committee evaluated the research proposal of which this study was part of.

Availability of data and materials

The genotypes and phenotypes generated and analysed during the current study are available in the Figshare repository: <https://doi.org/10.6084/m9.figshare.14315867>.

The QTL results generated and analysed during the current study are available in the Figshare repository: QTL results, <https://doi.org/10.6084/m9.figshare.14316068>.

Code availability

PedigreeSim can be found on GitHub: <https://github.com/PBR/pedigreeSim>.

The package mpQTL together with vignette and example datasets can be found on Github (<https://github.com/Alethere/mpQTL>).

The code for producing the datasets used in this article can be found on Figshare: <https://doi.org/10.6084/m9.figshare.14316107>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have no conflicts of interest regarding this article.

Received: 18 June 2021 Accepted: 12 January 2022

Published online: 14 February 2022

References

1. Lander ES, Botstein D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*. 1989;121:185–99.
2. Akond Z, Alam MJ, Hasan MN, Uddin MS, Alam M, Mollah MNH. A comparison on some interval mapping approaches for QTL detection. *Bioinformatics*. 2019;15:90. <https://doi.org/10.6026/97320630015090>.
3. Bourke PM, Voorrips RE, Visser RGF, Maliepaard C. Tools for genetic studies in experimental populations of polyploids. *Front Plant Sci*. 2018;9:513. <https://doi.org/10.3389/fpls.2018.00513>.

4. Rosyara UR, De Jong WS, Douches DS, Endelman JB. Software for genome-wide association studies in autopolyploids and its application to potato. *Plant Genome*. 2016. <https://doi.org/10.3835/plantgenome2015.08.0073>.
5. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*. 2013;9:29. <https://doi.org/10.1186/1746-4811-9-29>.
6. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*. 2006;38:203–8. <https://doi.org/10.1038/ng1702>.
7. Würschum T. Mapping QTL for agronomic traits in breeding populations. *Theor Appl Genet*. 2012;125:201–10. <https://doi.org/10.1007/s00122-012-1887-6>.
8. Jansen RC, Jannink J-L, Beavis WD. Mapping quantitative trait loci in plant breeding populations: use of parental haplotype sharing. *Crop Sci*. 2003;43:829. <https://doi.org/10.2135/cropsci2003.0829>.
9. Bardol N, Ventelon M, Mangin B, Jasson S, Loywick V, Couton F, et al. Combined linkage and linkage disequilibrium QTL mapping in multiple families of maize (*Zea mays* L.) line crosses highlights complementarities between models based on parental haplotype and single locus polymorphism. *Theor Appl Genet*. 2013;126:2717–36. <https://doi.org/10.1007/s00122-013-2167-9>.
10. Bink MCAM, Jansen J, Madduri M, Voorrips RE, Durel C-E, Kouassi AB, et al. Bayesian QTL analyses using pedigreed families of an outcrossing species, with application to fruit firmness in apple. *Theor Appl Genet*. 2014;127:1073–90. <https://doi.org/10.1007/s00122-014-2281-3>.
11. Leroux D, Rahmani A, Jasson S, Ventelon M, Louis F, Moreau L, et al. Clusthaplo: a plug-in for MCQTL to enhance QTL detection using ancestral alleles in multi-cross design. *Theor Appl Genet*. 2014;127:921–33. <https://doi.org/10.1007/s00122-014-2267-1>.
12. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, et al. Genetic properties of the maize nested association mapping population. *Science* (80-). 2009;325:737–40. <https://doi.org/10.1126/science.1174320>.
13. Garin V, Wimmer V, Mezouk S, Malosetti M, van Eeuwijk F. How do the type of QTL effect and the form of the residual term influence QTL detection in multi-parent populations? A case study in the maize EU-NAM population. *Theor Appl Genet*. 2017;130:1753–64.
14. R Core Team. R: a language and environment for statistical computing. 2016. <https://www.r-project.org/>. Accessed 26 Sept 2017.
15. Kempthorne O. An introduction to genetic statistics. New York: Wiley; 1957.
16. Hackett CA, Bradshaw JE, McNicol JW. Interval mapping of quantitative trait loci in autotetraploid species. *Genetics*. 2001;159:1819–32.
17. Luo ZW, Maliepaard CA, Leach L, Zhang R, Bradshaw J, Kearsey M, et al. Constructing genetic linkage maps under a tetrasomic model. *Genetics*. 2005;172:2635–45. <https://doi.org/10.1534/genetics.105.052449>.
18. Liu W, Reif JC, Ranc N, Della Porta G, Würschum T. Comparison of biometrical approaches for QTL detection in multiple segregating families. *Theor Appl Genet*. 2012;125:987–98. <https://doi.org/10.1007/s00122-012-1889-4>.
19. Giraud H, Lehermeier C, Bauer E, Falque M, Segura V, Bauland C, et al. Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic QTL for hybrid performance in the flint and dent heterotic groups of maize. *Genetics*. 2014;198:1717–34. <https://doi.org/10.1534/genetics.114.169367>.
20. Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping. *Genetics*. 1994;138:963–71.
21. Whittaker IC, Thompson R, Denham MC. Marker-assisted selection using ridge regression. *Genet Res*. 2000;75:249–52.
22. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 2001;157:1819–29. <https://doi.org/10.1093/genetics/157.4.1819>.
23. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome J*. 2011;4:250. <https://doi.org/10.3835/plantgenome2011.08.0024>.
24. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*. 2010;42:348–54. <https://doi.org/10.1038/ng.548>.
25. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet*. 2010;42:355–60. <https://doi.org/10.1038/ng.546>.
26. Voorrips RE, Maliepaard CA. The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinform*. 2012;13:248. <https://doi.org/10.1186/1471-2105-13-248>.
27. Bourke PM, Voorrips RE, Kranenburg T, Jansen J, Visser RGF, Maliepaard C. Integrating haplotype-specific linkage maps in tetraploid species using SNP markers. *Theor Appl Genet*. 2016;129:2211–26.
28. Haley CS, Knott SA. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* (Edinb). 1992;69:315–24.
29. Browning BL, Browning SR. A fast, powerful method for detecting identity by descent. *Am J Hum Genet*. 2011;88:173–82. <https://doi.org/10.1016/j.ajhg.2011.01.010>.
30. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. *Nat Rev Genet*. 2011;12:703–14. <https://doi.org/10.1038/nrg3054>.
31. He D, Saha S, Finkers R, Parida L. Efficient algorithms for polyploid haplotype phasing. *BMC Genom*. 2018;19:110. <https://doi.org/10.1186/s12864-018-4464-9>.
32. Thérèse Navarro A, Tumino G, Visser RGF, Voorrips RE, van de Weg WE, Maliepaard C. Multiparental QTL analysis: can we do it in polyploids? In: *Acta Horticulturae*. International Society for Horticultural Science; 2020. p. 55–64. <https://doi.org/10.17660/ActaHortic.2020.1283.9>.
33. Motazed E, de Ridder D, Finkers R, Baldwin S, Thomson S, Monaghan K, et al. TriPoly: haplotype estimation for polyploids using sequencing data of related individuals. *Bioinformatics*. 2018;34:3864–72. <https://doi.org/10.1093/bioinformatics/bty442>.

34. Berger E, Yorukoglu D, Peng J, Berger B. HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data. *PLoS Comput Biol.* 2014;10: e1003502. <https://doi.org/10.1371/journal.pcbi.1003502>.
35. Garg S, Martin M, Marschall T. Read-based phasing of related individuals. *Bioinformatics.* 2016;32:i234–42. <https://doi.org/10.1093/bioinformatics/btw276>.
36. Garin V, Wimmer V, Borchart D, Malosetti M, van Eeuwijk F. The influence of QTL allelic diversity on QTL detection in multi-parent populations: a simulation study in sugar beet. *bioRxiv.* 2020. <https://doi.org/10.1101/2020.02.04.930677>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

