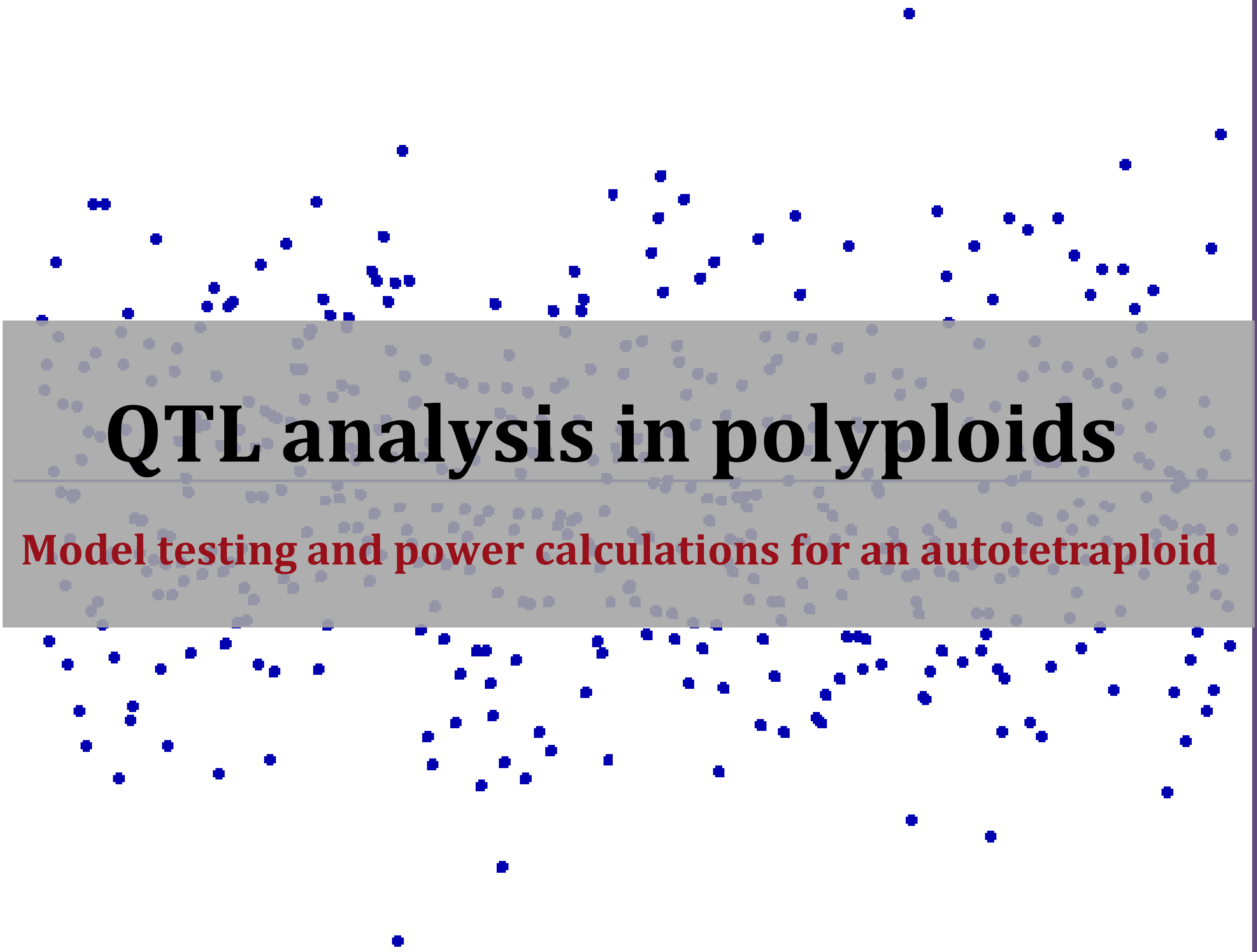


QTL analysis in polyploids

Model testing and power calculations for an autotetraploid

Peter Bourke

Laboratory of Plant Breeding, Wageningen University and Research Centre,
Wageningen, The Netherlands.

QTL analysis in polyploids

Model testing and power calculations for an autotetraploid

Peter M. Bourke

820728112100

Minor Thesis (MSc) Plant Breeding

PBR 80424

Supervisors:

Dr. Chris Maliepaard

Dr. Roeland Voorrips

July 2014

Laboratory of Plant Breeding, Wageningen University and Research Centre,
Wageningen, The Netherlands.

Abstract

The area of quantitative trait locus (QTL) analysis has provided the tools and methodology to identify and locate genetic factors underlying quantitative trait variation in diploid species for many years. More recently, attempts have been made to extend the methods to polyploid species which had previously been considered intractable. In particular, the use of co-dominant markers such as single nucleotide polymorphisms (SNPs) has enabled much denser marker sets to be generated which can help in resolving multiple homologues into distinct haplotypes. In this project two QTL models for an autotetraploid were tested using SNP dosage scores from an F1 mapping population as the starting point. A power study was performed to compare the power of detection of unlinked QTL using both models. Major factors which influence the power were identified as the QTL size and heritability, the size of the mapping population and the proximity of the closest flanking marker. Significance thresholds were derived using permutation tests, a subject which is examined in further detail in this report.

Key words: Polyploid, quantitative trait loci, QTL mapping, QTL analysis, statistical power, autotetraploid, SNP dosage, permutation test, F1 mapping population.

Acknowledgements

I would like to acknowledge the help and mentoring provided by both my supervisors Dr. Chris Maliepaard and Dr. Roeland Voorrips, without whom I would neither have started nor completed this work. I appreciate greatly the time that was taken during our weekly meetings which sometimes ran well into the second hour – these sessions challenged me to think more deeply about the various aspects of this topic and revise my methods when they were found wanting. I have learnt a lot during the course of this project and would like to personally thank both of you for that. I would like to thank Annelies Loonen and Marjon Arens for their kindness in allowing me to use the computer and desk in their office – it was a very pleasant place to sit and work during the past four months. I would like to thank Richard Finkers for providing me access to the PBR computer cluster and not afbreken any of my jobs "op het moment dat zaken uit de hand lopen", which they may have at times. I would also like to thank Ram Kumar Basnet for allowing me to use his 64-bit computer for simulations, without which many of the results reported would still be pending. Finally I would like to thank my promoter Prof. Richard Visser for providing me opportunity to conduct this research.

Table of Contents

1. Introduction

1.1. General Introduction – A brief history of polyploids

Ploidy is one of those terms not often heard in ordinary conversation, although from a biological perspective it is a rather important concept. Indeed, the word itself is probably derived from the more commonly-used terms ‘haploid’ and ‘diploid’ which refer to the state of having one (n) and two ($2n$) sets of chromosomes, respectively. Ploidy can be defined as the number of copies of the entire chromosome set in a cell (Acquaah, 2012). A polyploid is an organism that carries more than the diploid set of chromosomes in its (somatic) cells, thus having more than two copies of each chromosome (Acquaah, 2012). The focus of this report will be for the particular case of $n = 4$, the tetraploids, which are an important subgroup within the polyploids, at least in terms of plants that are of direct use to mankind.

Previous estimates put the frequency of polyploids at 70% of the angiosperms (Masterson, 1994), but more recent molecular studies have proposed revising this estimate to 90% or more (Cui et al., 2006). Many of these plants have since “re-diploidised” and would no longer be classified as polyploids, although the signature of their ancestral genomes can still be recognised (Leitch and Leitch, 2008). In the lineages of most flowering plants, polyploidisation appears to be more the norm than the exception. The current view is that genomes often undergo polyploidisation followed by sub-functionalisation or neo-functionalisation of genes (Adams and Wendel, 2005a), after which a gradual process of re-diploidisation occurs (Wang et al., 2005). Recent advances in genomics and sequencing have shed more light on this phenomenon through the analysis of orthology and synteny. However, as early as 1911 it was recognised that maize (*Zea mays*) was likely to have formerly been a tetraploid after two sets of paralogous chromosomes were identified in its karyotype (Kuwada, 1911). Haldane also held the view that duplication events might favour an organism by avoiding the effects of deleterious mutations in important genes (Haldane, 1932). Polyploidy is not only a condition found among plants but also occurs among fish, amphibians, reptiles, insects and fungi (Hieter and Griffiths, 1999; Otto and Whitton, 2000; Gallais, 2003; Comai, 2005).

1.2. The origins of polyploids

A polyploid individual may arise as a result of either somatic or zygotic chromosome doubling, or by a non-reduction event in the formation of gametes (deWet, 1980). The latter of the two processes is considered to be the more important in the majority of cases (Harlan and deWet, 1975). In this case, an autotetraploid can result either directly from the fusion of two unreduced gametes, or through the initial formation of a triploid as a bridging step to increase the frequency of diploid gametes and thus likelihood of a tetraploid forming (Ramsey and Schemske, 1998).

1.3. Auto- and Allopolyploids

Generally-speaking, polyploids are classified as either autopolyploid or allopolyploid. An autopolyploid is a polyploid whose genome displays polysomic inheritance (so that all homologues of a chromosome can potentially pair during meiosis) and whose basic chromosome sets are homologous (Gallais, 2003). An allopolyploid is generally the result of a hybridization between the genomes of different (albeit closely-related) species and exhibits disomic inheritance (so there is preferential pairing between pairs of homologous chromosomes, as opposed to random pairing). In terms of inheritance patterns and homologue pairing behaviour, allopolyploids essentially behave like diploids (Milbourne et al., 2008).

A third group can also be identified – segmental allopolyploids, which exhibit a mixture of autopolyploid and allopolyploid inheritance patterns (Stebbins, 1947; Gallais, 2003). Recent studies have shown that in certain species, loci do not follow the same mode of inheritance among different tetraploid individuals, suggesting that segregation in tetraploid hybrids are not predictable (Stift et al., 2008). It is the mode of inheritance and pairing behaviour that is the more important consideration for researchers working in the area of QTL mapping rather than the origin of the sub-genomes *per se* (Doerge and Craig, 2000), while for those interested in the evolution of genomes the opposite is probably true, *e.g.* Ramsey and Schemske (1998). There is still some discussion over the precise meanings of the terms or to the most pertinent aspects in their definition (Ramsey and Schemske, 1998).

1.4. Polyploidy in plant breeding

Polyploids commonly occur among cultivated plants. These include important food crops such as potato, wheat and banana, fibre crops such as cotton, fodder crops such as alfalfa, commodity crops such as coffee, sugarcane and oilseed rape and ornamental crops such as rose and begonia. A list of some of the more well-known polyploid crops and their most commonly-occurring ploidy levels is given in Table 1. A simple classification of some of the ornamentals like Begonia is not possible given the complexity of the genus, where chromosomal number ranges from 2*n* = 16 up to 2*n* = 156 (Dewitte et al., 2011).

In some crops the mode of inheritance has been well-established, such as autotetraploid potato or allohexaploid wheat. In others, there is still debate over the mechanisms involved *e.g.* sugarcane (Milbourne et al., 2008) or rose (Koning-Boucoiran et al., 2012). However, it is only a matter of time before all such issues are finally resolved due to the increasingly widespread use of molecular markers which can be used to investigate segregation patterns in more detail (Wu et al., 2001; Stift et al., 2008). In terms of their frequency, polyploids appear among domesticated crops at approximately the same frequency as they would had they been selected at random from their respective plant families (Hilu, 1993). Autopolyploids often display an increased size of certain organs, an effect termed 'gigas' feature (Acquaah, 2012). This has been exploited in ornamental plant breeding, where the prolonged flowering period of polyploids coupled with an increase in flower size and complexity can be desirable. Triploid banana is seedless, a commercially desirable trait. In forage production, autopolyploid red clovers and ryegrasses have been bred to produce larger, more succulent leaves, again which may be due to the gigas feature. The introduction of triploid sugar beet (with mono-germ seeds) had a large impact on the sugar beet industry (Acquaah, 2012).

Table 1. Some commonly-cultivated polyploid crops and their most common ploidy levels

Crop name	Species	Auto- / Allo-	Ploidy level

Potato	*Solanum tuberosum*	Autotetraploid	$2n = 4x = 48$
Sugarcane	*Saccharum officinarum*	Auto-octaploid	$2n = 8x = 80$
Durum wheat	*Triticum durum*	Allotetraploid	$2n = 4x = 28$
Bread wheat	*Triticum aestivum*	Allohexaploid	$2n = 6x = 42$
Leek	*Allium porrum*	Autotetraploid (Pink, 1993)	$2n = 4x = 32$
Banana	*Musa spp.*	Auto- & Allotriploids	$2n = 3x = 33$
Strawberry	*Fragaria × ananassa*	Interspecific octoploid hybrid	$2n = 8x = 56$
Oat	*Avena sativa*	Allohexaploid	$2n = 6x = 42$
Peanut	*Arachis hypogaea*	Autotetraploid	$2n = 4x = 40$
Sweet potato	*Ipomoea batatas*	Autohexaploid	$2n = 6x = 90$
Cotton	*Gossypium herbaceum*	Allotetraploid	$2n = 4x = 52$
Oilseed rape	*Brassica napus*	Allopolyploid	$2n = 4x = 38$
Alfalfa	*Medicago sativa*	Autotetraploid	$2n = 4x = 32$
Coffee	*Coffea arabica*	Allotetraploid (Lashermes et al., 1999)	$2n = 4x = 44$
Rose	*Rosa × hybrida*	Autotetraploid	$2n = 3x = 21$ $2n = 4x = 28$
Petunia	*Petunia hybrida*	Autotetraploid	$2n = 4x = 28$
Chrysanthemum	*Chrysanthemum spp.*	Various	$2n = 36, \ldots, 75$
Begonia	*Begonia spp.*	Various	$2n = 16, \ldots, 156$

1.5. Introduction to QTL analysis

Many traits of agronomic importance are known to be controlled by multiple genes, in contrast to simple Mendelian traits which are controlled by allelic variants of one gene at a single locus. The analysis and identification of these multiple genetic loci has been a major area of research in both plant and animal science for almost a century, when a model of how quantitative traits could be explained by Mendelian inheritance at multiple loci was first proposed (Fisher, 1919). More recently, rapid advances in molecular techniques have given researchers access to vast amounts of genotypic information with which to investigate the possible underlying causes of variation in quantitative traits (Doerge, 2002). The putative location of such an underlying factor (usually assumed to be a single gene) has been termed a quantitative trait locus (QTL), or collectively referred to as quantitative trait loci (also denoted QTL).

1.6. Difficulties with QTL analysis in (auto)polyploids

There are a number of difficulties in QTL mapping and analysis in autopolyploids which have hampered the development and utilisation of modern molecular resources in these crops to date. Firstly, much of the research and methodological development in the area of QTL mapping has focussed on diploid species up to now (Sax, 1923; Lander and Botstein, 1989; Jansen, 1992). This may not be a direct hindrance as such but has perhaps diverted the attention of researchers away from the area. The fact that humans are diploid may have also edged the bias towards their predominance somewhat more. Somewhat related to this is the fact that the techniques used in diploid crops may be applied quite easily to allopolyploids, which occur at least as frequently as autopolyploids (Milbourne et al., 2008). For arguably the most important autopolyploid crop species, potato, closely-related diploid potato can be used as a substitute in mapping studies (Doerge and Craig, 2000) and indeed already has been with some success (Leonards-Schippers et al., 1994; Van Eck et al., 1994; Collins et al., 1999).

The meiotic behaviour of polyploid species is not always known (whereas that in diploids is generally predictable), making it harder to predict segregation patterns in the offspring of an experimental cross. A general methodology for predicting meiotic behaviour has been developed (Wu et al., 2001) but even when molecular markers are used, it may still be difficult to distinguish some segregation patterns, particularly in the case of dominant markers (Gallais, 2003). There are also a much larger number of genotype possibilities at any marker / QTL locus. At its most extreme, a locus in the F1 offspring of a cross between two tetraploid parents may have one of up to eight possible alleles, with 36 possible combinations of parental alleles in total.

As well as this, the increased heterozygosity of polyploids can also pose difficulties in developing inbred lines (due to inbreeding depression as well as the impossibility of reaching homozygosity through selfing). Again, the result is that traditional QTL approaches involving inbred lines most often used in diploids cannot be applied to polyploids. Therefore, QTL mapping must be done in genetically-diverse populations. Difficulties may also arise in resolving marker dosages in both parents and offspring in which case reconstructing the haplotypes may prove extremely difficult. It is also possible that at higher dosages a masking of recombination behaviour can occur, since markers become less informative at intermediate dosages about possible recombination events (Doerge and Craig, 2000).

In polyploids with quadrivalent formation double reduction may also occur which results in seemingly impossible configurations appearing among the progeny (*e.g.* offspring carrying a duplex dose of a marker allele when the parental segregation was Simplex x Nulliplex). If double reductions are not correctly identified, they may adversely affect the construction of linkage maps and falsify subsequent QTL analyses. In general however, the rate of double reductions tends to be small: a maximum rate of ¼ was recently proposed (Luo et al., 2006) while some authors prefer the original limit of 1/6 proposed by Mather (1935); in any case, experimental studies have shown it to be relatively small, *e.g.* Stift et al. (2008).

1.7. Motivation for this work

Probably the biggest motivation for the use of QTL mapping in polyploids is to benefit from the advantages conferred through the use of marker information in breeding programmes. The use of marker-assisted selection (MAS) has already proven to be extremely effective in both selection and backcrossing schemes at the diploid level and it is hoped that at higher ploidy levels these benefits can also be enjoyed.

Despite the hurdles previously mentioned, the possibilities for conducting QTL analyses in polyploid crops are finally being realised due to a number of factors. With the advent of high-density markers such as single-nucleotide polymorphisms (SNPs) it is possible to generate high-quality, reliable, co-dominant marker information at a fraction of the cost of previous methods and with sufficient coverage and density that even small-effect QTL may be detected.

In time, genotyping by sequencing will likely provide even more informative data on polyploid haplotypes. The availability of vastly more powerful computers and methodologies for constructing linkage maps which exploit iterative algorithms are providing numerical solutions to previously intractable problems. It is therefore increasingly feasible to perform QTL mapping in polyploids given the available technology. There has also been an increased interest from breeding companies who have experienced the benefits of MAS in diploid plants and are willing to make the efforts required to conduct QTL analyses in polyploids in the hope of developing selectable markers for useful traits at the polyploid level.

Apart from these practical motivations, another reason why it may be preferable to perform experimental work at higher ploidy levels is that there may be differences in the expression of certain genes that cannot be simply accounted for by dosage number. For example, it was shown in experimental lines of yeast (*Saccharomyces cerevisiae*) that some genes were strongly induced while others were strongly repressed in a set of identical yeast lines with different ploidy levels (Galitski et al., 1999). In maize, it was found that when expression levels of eighteen genes were studied in an experimental ploidy series (1x, 2x, 3x, and 4x), there were a number of important deviations from a model in which gene expression was a simple multiple of the expression level for the monoploid case (Guo et al., 1996). A similar finding was reported for an autopolyploid series in potato (Stupar et al., 2007) using a microarray representing over 9000 genes. It appears that gene silencing is a common occurrence following polyploidisation (Adams and Wendel, 2005b). As well as that, up- and down-regulation of duplicated genes in polyploids was found to be organ-specific in certain cases (Adams and Wendel, 2005b). In other words, the genetic and molecular dynamics of a polyploid may not be fully captured by using diploid plants as an experimental model in a mapping population.

1.8. QTL mapping

There are a number of steps involved in QTL mapping, which can broadly be categorised as *detection*, *location* and *estimation*. That is to say, we must detect the presence of segregating genetic factors, locate them relative to some known markers and estimate their effects and possible interactions (Churchill and Doerge, 1998). QTL mapping in polyploids is no different to diploids in this respect.

At its most basic, QTL mapping involves comparing the means of two groups in a mapping population, one of which contain one form of a marker and the other an alternative form (or in the case of multi-allelic markers, comparison of means between multiple groups). Standard statistical tests like Student’s *t*-test are then used to see whether there is a statistically significant difference between the means of the two groups for some trait of interest for which phenotypic data is available. When multiple markers are available multiple comparisons need to be made, leading to the use of more stringent testing criteria so as to control the rate of type I errors (false positive results).

The testing of differences in means is similar to a regression-like or ANOVA approach, where the emphasis is no longer on means but the variance explained. Marker locations that “explain” the greatest amount of variance in the phenotypic data are said to be more likely linked to a QTL than other markers. Again, the approach relies on grouping the data in a particular way and observing when the variance between the groups becomes large in comparison to the variance within the groups, which forms the basis for an F-test of significance.

Another approach taken in QTL mapping uses maximum likelihood estimations and has probably been the more common approach to QTL mapping in diploids since it was proposed for a single-marker analysis (Weller, 1986). Whether in a single-marker or interval-mapping approach, a ratio is formed between the maximum likelihood scores under the alternative and null hypotheses of QTL linkage. In general, the $\log_{10}$ of this ratio (the LOD score) is used as a measure of the strength of an association between a location and an effect. In other cases, the Wald score can be used in place of the LOD as a test statistic (Chen, 2014).

1.9. QTL mapping in polyploids

To date, there have been a number of linkage maps of autopolyploid crops published which include potato (Meyer et al., 1998; Bradshaw et al., 2004), alfalfa (Brouwer and Osborn, 1999), sweet potato (Ukoskit and Thompson, 1997; Kriegner et al., 2003), sugarcane (Da Silva, 1993; Ripol et al., 1999; Ming et al., 2001) and rose (Gar et al., 2011; Koning-Boucoiran et al., 2012). Other authors have presented methodologies for linkage mapping in autopolyploids, including using the assumption of disomic inheritance (Ripol et al., 1999; Hackett et al., 2001) and tetrasomic inheritance (Luo et al., 2006). Some groups have simplified the problem by using simplex markers which segregate in a 1:1 ratio in the offspring (Wu et al., 1992; Meyer et al., 1998; Ming et al., 2001).

The creation of a linkage map is not an end in itself, but it usually a precursor to a QTL analysis for some trait of interest. For example, the focus of mapping studies in sugarcane has tended to be the identification of genes which control the sugar content or sugar yield (Ming et al., 2001; Grivet and Arruda, 2002; Ming et al., 2002). Therefore, linkage map construction can also be thought of as an integral part of the methodology of QTL analysis which may offer constraints or benefits depending on such factors as the marker type, marker coverage and density and level of map integration across marker segregations and across different homologues.

More recently, genotyping by sequencing (GBS) using next-generation sequencing has been proposed an approach to identify SNPs and haplotypes in more genetically diverse populations than traditional mapping populations (Elshire et al., 2011). Genome-wide association studies (GWAS) aim to identify allelic variants in panels of diverse material using high-density SNP arrays. To date, there have been reports of mixed success with the methodology (Visscher et al., 2012), although this remains a promising area of continuing research.

The earliest efforts at QTL mapping in polyploids used regression models to compare trait means for different phenotypes (Hackett et al., 2001). For example, a QTL analysis into potato late blight resistance at the tetraploid level by Meyer *et al.* (1998) tested for associations between AFLP markers and resistance scores using the non-parametric Mann-Whitney test, with significance thresholds determined by permutation tests with N = 1000 and $\alpha = 0.05$ (Churchill and Doerge, 1994). However, it was noted that “simple tests based on the presence/absence of a band are insufficiently powerful for QTL mapping in outbreeding polyploids, and that a more powerful QTL model needs to be developed” (Meyer et al., 1998). A more advanced model was proposed by members of the same group three years later (Hackett et al., 2001). In this approach, interval mapping was performed using maximum likelihood estimates for the combined trait values and QTL probabilities using an iterative approach on genotype probabilities until the likelihood ratio converged. This approach was compared to a single weighted regression of the QTL genotype probabilities derived from the marker data only (without taking trait values into consideration, and with a single iteration). It was found that there was better estimation of the QTL allele effects using the iterative approach, and that estimates of the percentage variance accounted for and residual variance were also better using the iterative versus the non-iterative approach (Hackett et al., 2001).

1.10. QTL mapping using SNP dosage data

The advent of single nucleotide polymorphism (SNP) markers has resulted in a big step forward in molecular studies as it offers researchers the possibility of reliable and reproducible co-dominant markers which can be easily, quickly and (increasingly) cheaply scored at a much higher density along the chromosome than previous marker types. Currently, SNP markers are the marker type of choice, although given the pace of developments it is likely that they will be replaced by another advancement within the coming years. SNP markers are generally bi-allelic and therefore less informative than RFLP or microsatellite markers that are multi-allelic (Acquaah, 2012). This need not be a serious issue because of their abundance along the chromosome; SNP data has been used for more than a decade in diploids to generate haplotype information by considering collections of SNP markers as a block (Gabriel et al., 2002).

One of the more recent publications on QTL mapping in polyploids has been in potato using SNP dosage data (Hackett et al., 2013), amended from the approach proposed previously by Hackett *et al*. (2001). This single marker approach was adapted and simplified from an earlier model proposed by Kempthorne which included interaction effects between the alleles (Kempthorne, 1957). However, Hackett *et al.* extend Kempthorne’s model to an interval mapping approach by considering the genotype probabilities at a grid of points along the chromosome (thus becoming independent of the marker positions).

SNP dosage scores are inferred from allele intensity ratios of SNP array data using normal mixture models. Dosage information is then incorporated in a linkage analysis for map construction and to generate genotype probabilities from which an iterative weighted regression on the QTL genotype probabilities is used. This publication has served as a useful guide in this project and as a source of ideas and inspiration. However, the application to QTL mapping remains incomplete as the method was only tested with theta scores (where theta refers to the polar coordinates of the signal intensity from the SNP array). It is therefore unknown how well this method works when applied to real (or even simulated) phenotypic data.

1.11. QTL models considered in this project

Two separate models were considered based on a regression approach rather than a maximum likelihood approach – although had there been more time a maximum likelihood approach may also have been implemented. Regression models have a somewhat simpler formulation and therefore represent a first attempt at modelling QTL effects.

1.11.1. Simple Regression on dosage scores

A very basic approach is to perform a single-marker regression (or ANOVA) analysis using marker dosage scores as the predictive variable. The basic model to be fitted is

$$y_j = \mu + \beta x_j + \varepsilon_j$$

where x_j is a dummy variable for dosage scores according to:

$$x_j = \begin{pmatrix} -2 & \text{.... for aaaa} \\ -1 & \text{.... for Aaaa} \\ 0 & \text{.... for AAaa} \\ +1 & \text{.... for AAAa} \\ +2 & \text{.... for AAAA} \end{pmatrix}$$

μ corresponds to the mean genotypic value in the duplex case. It should be noted that this model is purely additive and ignores dominance effects.

Hypothesis testing can be performed using an F-statistic, the ratio of the residual mean squares for the full model and the reduced model (Liu and Knapp, 1997). It is often considered more convenient to consider the p-value of the F-statistic, which is the probability of the F-statistic occurring under the null hypothesis (H_0). In this case, the null hypothesis is that there is no difference in the means between the full model and reduced model, *i.e.* that the marker is not linked to a QTL. We may choose to reject H_0 if p is less than a certain limit, or alternatively, if its negative logarithm ($-\log_{10} p$) exceeds a particular threshold. This quantity ($-\log_{10} p$) is at times referred to as the LOP score in this document (for logarithm of p-value). Arbitrarily setting significance thresholds runs the risk of introducing type I errors (when no linked QTL exists but we incorrectly declare that there is a significant linkage present) due to the multiple comparisons that are performed, which will be examined in more detail in §3.5.

1.11.2. Regression using parental genotype probabilities

An alternative approach to model a quantitative trait is to consider the contributions made by each of the eight possible parental alleles that may be present at a locus. Each full-sib offspring of a cross between two parents P1 and P2 inherits two alleles from P1 and two from P2 (assuming no double reduction occurs), resulting in 36 possible combinations of alleles at a locus. At any locus, the contribution from the two parents can be represented using a dummy variable X_i, where $i \in (1,2,\ldots,8)$, such that $X_i = 0$ if allele i is not present in the offspring, and $X_i = 1$ if allele i is present. At positions where the parental identity of the homologues is not fully known, intermediate values between 0 and 1 are used to represent the probabilities of that homologue being present.

The model can be thus written as

$$y_j = \mu_1 + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5 + a_6X_6 + a_7X_7 + a_8X_8 + \varepsilon_j$$

where y_j is the phenotypic value for the jth individual, μ_1 is the overall mean and ε_j is the residual error term. For any individual, we also have the constraint that $X_1 + X_2 + X_3 + X_4 = 2$ and $X_5 + X_6 + X_7 + X_8 = 2$, since two alleles are inherited from each parent. We can therefore replace X_4 and X_8 in the above equation, and re-name the coefficients to yield

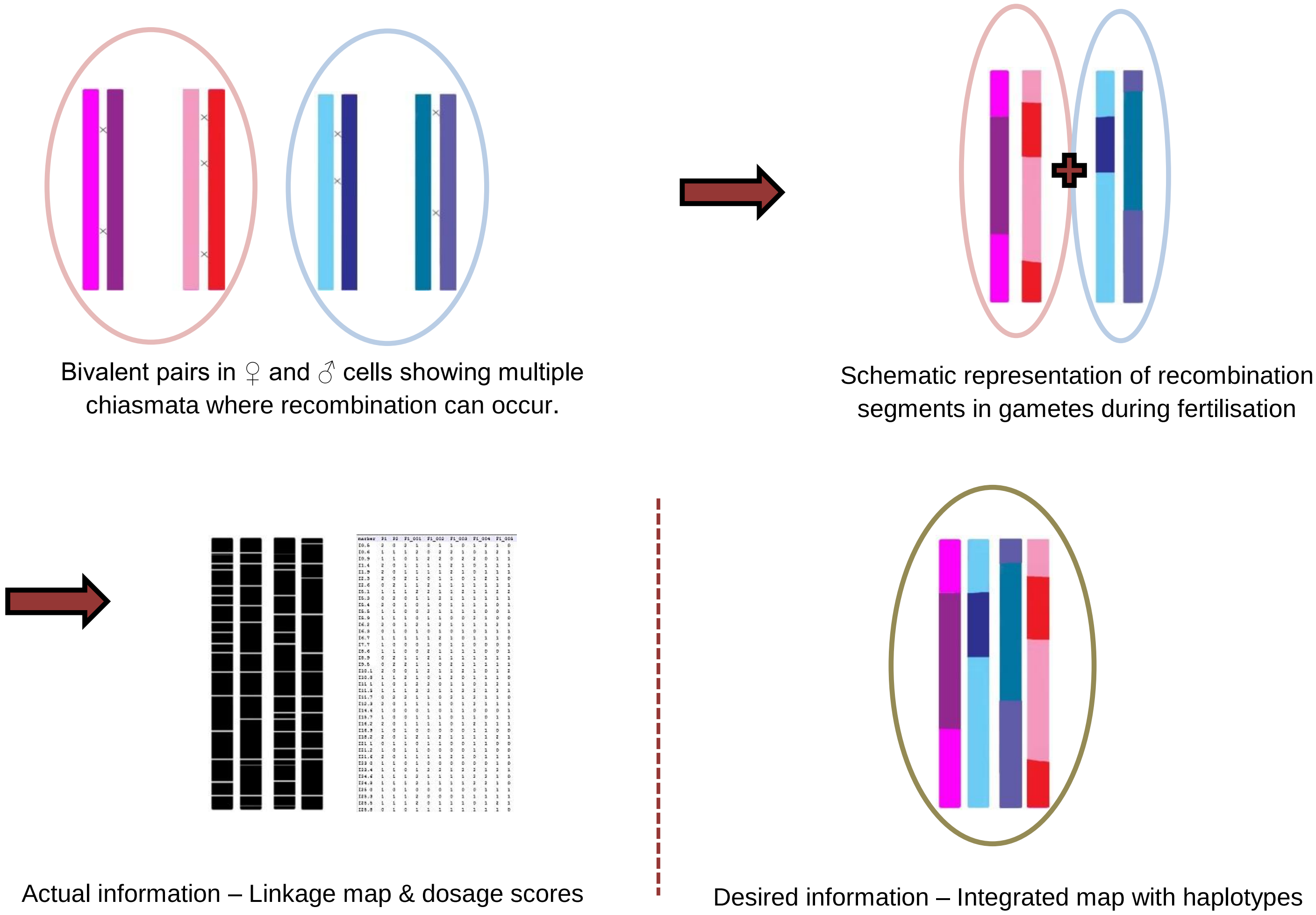
$$y_j = \mu + \alpha_1X_1 + \alpha_2X_2 + \alpha_3X_3 + \alpha_5X_5 + \alpha_6X_6 + \alpha_7X_7 + \varepsilon_j$$

where $\mu = \mu_1 + 2(a_4 + a_8)$, $\alpha_1 = a_1 - a_4$ *etc*. This model is termed the genotype probabilities model, or GP model for short.

1.12. Genotype probabilities

The model just described relies on having information on the identity of each allele (from which of the eight parental homologues it actually came from) at a set of locations along each chromosome. These locations are not necessarily at the marker positions, but usually the marker positions are used to begin with, and interpolation extends the probabilities to a grid of positions (Hackett et al., 2013). Ideally, we would like to be able to track all recombination events and thus reconstruct the full parental homologue identities in any individual of the mapping population (schematic representation in Figure 1). In practice, the data available to us is a linkage map containing the order and distance between marker positions and the dosages scores of the progeny at those positions. It is desirable to have a consensus map which integrates different marker categories (at least including S x N, S x S and D x N markers) as well as integrating the four homologues into a single chromosomal map. For example in a previous mapping project in rose, the length (in cM) of different homologues of the same chromosome were not equal (Santos Leonardo, 2013) possibly due to poor marker coverage in certain regions. This may raise some concerns over the assignment of unique positions to all marker on a potential consensus map. For species where a reference genome is available, anchoring of marker positions to a map position (in Mbp) may be feasible.

Various methods for recovering the identity of the parental homologue identities have already been proposed including using a branch and bound algorithm (Hackett et al., 2001) as well as a more recent approach using a hidden Markov model (Hackett et al., 2013).



Bivalent pairs in ♀ and ♂ cells showing multiple chiasmata where recombination can occur.

Schematic representation of recombination segments in gametes during fertilisation

Actual information – Linkage map & dosage scores

Desired information – Integrated map with haplotypes

Figure 1. Schematic drawing of bivalent pairing and gametes for a single chromosome in an autotetraploid.

1.13. Significance thresholds

The topic of significance thresholds has received a lot of attention since QTL mapping was initiated, and rightly so. A QTL is only detected when a linked marker exceeds such a threshold, meaning that significance thresholds are fundamental to any sort of QTL analysis. If significance levels are set too high there is a risk of missing smaller-effect QTL that fall just below the bar; on the other hand, if the threshold is set too low there is a risk that areas of the genome which harbour no true QTL will be incorrectly identified.

The basis of QTL mapping is to determine and quantify where there is an increased likelihood of locating a genetic factor that influences a particular trait of interest. All QTL mapping methods require significance thresholds in order to discriminate between areas of higher significance and areas of lower significance. If the trait values are normally distributed (in terms of their residuals), it is relatively straightforward to determine a critical threshold for the test statistic based on a stringency level α. However, it is often the case that trait values are not normally-distributed. In these situations, there are a number of different options that have been proposed to set a significance threshold (reviewed in Doerge (2002)).

When highly complex models are involved which may include epistatic interactions between multiple loci, one option is to use computer simulations to generate thresholds (Rebai et al., 1994). This has the advantage of being relatively fast, although it relies on a prior knowledge of the expected distribution of the test statistic. In many cases this is not known, or the experimenter does not want to assume that all the information has been fully captured by the model. In these instances non-parametric tests such as permutation or bootstrap methods are generally used (Churchill and Doerge, 1994; Efron and Tibshirani, 1994). These methods make no prior assumptions about the distribution of the test statistic, but have the drawback that they are computationally expensive.

An approximate approach to determining a LOD threshold based on tables derived from multiple simulations (1,000,000 per population type) was published which provides a fast method of determining LOD thresholds in the four main experimental (diploid) population types (Van Ooijen, 1999). An alternative set of thresholds was more recently proposed which derived the significance threshold from theoretical considerations (Piepho, 2001). This method has the slight advantage of being applicable to all population types, although is still confined to a diploid species. It was demonstrated that theoretical and re-sampling based thresholds are approximately equal in situations where biological and statistical effects are minimised (Van Ooijen, 1999); such scenarios include minimal segregation distortion, adequate sample size, low scoring errors or incomplete data and high heritability (Doerge, 2002).

So far there has been little attention paid to the issue of statistical thresholds in polyploids (to the best of my knowledge). For this project, experiment-wise permutation tests were taken as the simplest method of arriving at an approximate threshold level which make no prior assumptions about the distribution of the data.

1.14. Investigations in this project

The basis of this work is to investigate the conditions and criteria under which QTL may be identified in an F1 mapping population of an autotetraploid plant species (such as rose, potato or alfalfa). A number of models will be identified and tested after which a range of topics will be investigated, which include:

- **Additive / Dominance effects**
  What is the difference (if any) in detection capability between a purely additive-effect QTL and one which is purely dominant? How effective is the model at identifying both these types of QTL effect?

- **Effect of QTL segregation**

What effect does the QTL segregation type have on the power of detection, under the assumption of additivity or dominance? What is the relationship between the expected variance due to a locus, the actual variance in the mapping population and the strength of the statistical association in markers linked to that locus?

- **Effect of QTL effect size and different heritabilities**
  What impact does the QTL effect size have on the power of detection? Or the level of heritability for the trait?

- **Setting significance thresholds**
  How should a threshold for significance be set? If using a permutation procedure, how many iterations are needed? What is the relationship between significance thresholds and the power of QTL detection (different type I and type II errors)?

- **Effect of marker density**
  What effect does marker density have on the power to detect QTL? How does marker density affect the accuracy of reconstructed genotype probabilities?

- **Usefulness of different marker types**
  Can we say anything about the relative usefulness / informativeness of different marker types / segregations for QTL mapping? (as opposed to their relative usefulness in linkage map construction, which is a different issue)

- **Effect of mapping population size**
  What is the relationship between the size of the mapping population, the power of detection and the significance of a detected QTL effect?

In addition to these, the topic of predicting genotype probabilities using SNP dosage data will also be dealt with to some extent, as this is a pre-requisite for a QTL mapping approach which uses regression of the phenotypic data on genotype probabilities or a likelihood method which uses QTL probabilities.

2. Materials and methods

As mentioned in the Introduction, there are three possible types of polyploid that we may consider – autopolyploids, allopolyploids and segmental allopolyploids. The remainder of this report will focus on the first of these three categories, and in particular, the autotetraploids, which include such crops as potato, leek, alfalfa, petunia and rose. Given that this project was conducted solely *in silico*, this section will describe the model development and choice of model to use, the assumptions that were made and the steps that were taken in investigating these models.

For an autotetraploid species, an F1 mapping population has been the most widely-used in previous mapping studies (Meyer et al., 1998; Hackett et al., 2001; Luo et al., 2001; Bradshaw et al., 2004; Bryan et al., 2004; Hackett et al., 2013). An F2 mapping population was used in a QTL analysis into cotton quality-traits (Jiang et al., 2000), although it relied on the development of near-isogenic lines to generate one of the crossing parents (Thomson et al., 1987) and one round of selfing after the first cross. An F1 mapping populations was taken as the more general mapping population in this context, which does not require tolerance to inbreeding.

2.1. Pedigree Simulation

One of the key tools in conducting this study was the ability to simulate meiosis in a tetraploid population through the use of the PedigreeSim software programme, written in Java (Voorrips and Maliepaard, 2012). A number of assumptions were made when choosing the settings under which the simulation was run. Firstly, it was assumed that there was completely random pairing between all homologues (tetrasomic behaviour) and that bivalents rather than quadrivalents are formed during meiosis. As discussed in Hackett et al. (2001), the evidence from cytological studies in potato suggests that bivalents predominate but that quadrivalents, trivalents and univalents may also occur in low frequencies (Swaminathan and Howard, 1953). In alfalfa, bivalents were also found to predominate in metaphase 1 (Bingham and McCoy, 1988). Therefore, this simplification is justified for some of the more important autotetraploid crop species. Bivalent formation also rules out the possibility of double reduction occurring (Milbourne et al., 2008). However, the analysis could be extended to include quadrivalent formation, or a combination of bivalent and quandrivalents (Voorrips and Maliepaard, 2012). It is also possible to extend the analysis from random chromosomal pairing to fully preferential pairing, associated with allopolyploidy, or partially preferential pairing for segmental allopolyploidy.

2.2. R

All of the subsequent simulation work, data analysis and plotting was performed in R version 3.0.3 or earlier (R_Core_Team, 2014). The output files of PedigreeSim (in particular, the '_out_founderalleles.dat' and '_out_alleledose.dat' files) were used as input files in subsequent scripts to reconstruct QTL genotype probabilities using the marker dosage information. An example script is provided in Appendix X.

2.3. Simulation studies

QTL were simulated as either fully additive or fully dominant throughout all analyses in this project. Thus, partial dominance and recessiveness were not considered (although there is nothing to prevent extending the analysis to include these situations). A script written in R to assign phenotypic scores at a given locus was used to determine the total genetic contribution of the QTL to an individual's trait value (Voorrips 2014, unpublished).

Following calculation of the genetic contributions, an estimate for the population-wide genetic variance σ_G^2 was determined.

Recalling the definition of broad-sense heritability, $h^2 = \frac{\sigma_G^2}{\sigma_P^2} = \frac{\sigma_G^2}{\sigma_E^2 + \sigma_G^2}$ (Acquaah, 2012), the environmental variance σ_E^2 was calculated as

$$\sigma_E{}^2 = \left(\frac{1 - h^2}{h^2}\right)\sigma_G{}^2$$

Heritabilities of zero and one were excluded. Environmental effects were simulated by randomly sampling from a normal distribution with mean 0 and variance σ_E^2. These were added to the genetic contributions (and the overall trait mean μ) to generate a phenotypic score.

2.4. Model implementation

Both the simple regression model on marker dosage scores and the regression on the parental genotype probabilities were implemented in R. Using the '_out_founderalleles.dat' output from PedigreeSim, complete information for the mapping population in terms of the identity of the parental alleles was known. However, a QTL study would not have this information to begin with and therefore an attempt was made to reconstruct the parental homologue identities using marker dosage information.

It is assumed that a linkage map already exists and that the identity of the parental haplotypes is fully known (so information on the phase of the complete linkage group is known). For example, if parent 1 has four homologues A, B, C and D and parent 2 has homologues E, F, G and H of a certain chromosome, it is known which of the eight homologues harbours the "minor" allele (or "1" allele) in a Simplex x Nulliplex situation *etc*. Simplex SNPs provide the framework from which the identity of all four homologous chromosomes may be determined; other SNP markers may then added relative to these by maximising the LOD scores for different allele pairs (Hackett et al., 2013).

2.5. Reconstruction of parental homologue identities in offspring

We require information on the parental identity of all four homologues in order to implement the GP model. A simple algorithm for assigning the probability of the presence or absence of each of the eight parental homologues at the marker positions using marker dosage data was implemented in R. This was tested under various scenarios and adjusted until it was found to approximate the results obtained when the full parental homologue identities were known.

The starting point for this approach was to use markers for which the parental identity of at least one of the four homologues was known. From these positions, probabilities were assigned to nearby linked loci according to the Kosambi mapping function . The Kosambi mapping function was chosen over the Haldane mapping function as it models crossover interference and can be simply expressed in closed form, which is not true of the more complicated Carter-Falconer or Felsenstein mapping functions (Chen, 2014).

The Kosambi mapping function relates the distance *d* (in Morgans) to the recombination fraction *θ* (Kosambi, 1943):

$$\theta = \frac{1}{2}\left(\frac{e^{4d} - 1}{e^{4d} + 1}\right)$$

Two situations were considered – the first being where we know the identity of one of the homologues at a marker position. The second situation is where we know with certainty that one of the parental homologues is *not* present. Certain marker positions may provide both types of information in an individual. We already have information on the estimated recombination fraction between marker positions (during map construction), which are alternatively expressed as genetic distances through a mapping function. These may be converted into a probability of linkage in coupling (so, a probability that no recombination actually occurred) at a nearby position on a homologue by considering the what the probability of linkage in coupling is at a recombination fraction of 0 and at a recombination fraction of 0.5 (the so-called boundary conditions).

In the first case where we know the identity of one of the homologues, the boundary conditions are *p*(0) = 1 and *p*(0.5) = 0.5, where *p*(*θ*) is the probability of linkage in coupling when the recombination fraction is *θ.*

We may therefore write

$$p_1(\theta) = 1 - \theta = \tfrac{1}{2}\left(\tfrac{e^{4d}+3}{e^{4d}+1}\right)$$

In the second case (absence) the boundary conditions are *p*(0) = 0 and *p*(0.5) = 0.5 and thus

$$p_0(\theta) = \frac{1}{2}\left(\frac{e^{4d} - 1}{e^{4d} + 1}\right)$$

These functions are referred to as the Kosambi-1 and Kosambi-0 functions, respectively. They express the probability of linkage in coupling at a distance *d* from a marker position.

To implement the algorithm, probabilities are initially assigned at all marker positions in the offspring of the mapping population using marker dosage information as a guide. Some examples of how probabilities were assigned in the case of Simplex x Nulliplex, Simplex x Simplex and Duplex x Nulliplex markers are given in Table 2. It will be noted that the condition that each set of four X_i probabilities should sum to 2 is in some cases violated – this is corrected through normalisation at the end of the procedure.

Table 2. Example starting probabilities for SxN, SxS and DxN markers according to dosage scores.

Marker segregation	Dosage	Minor allele(s)	Genotype probabilities							
		(e.g.)	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
S x N	0	homologue 7	0.5	0.5	0.5	0.5	0.5	0.5	**0**	0.5
	1	homologue 7	0.5	0.5	0.5	0.5	0.5	0.5	**1**	0.5
S x S	0	homologues 3, 5	0.5	0.5	**0**	0.5	**0**	0.5	0.5	0.5
	1	homologues 3, 5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	2	homologues 3, 5	0.5	0.5	**1**	0.5	**1**	0.5	0.5	0.5
D x N	0	homologues 7, 8	0.5	0.5	0.5	0.5	**1**	**1**	**0**	**0**
	1	homologues 7, 8	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	2	homologues 7, 8	0.5	0.5	0.5	0.5	**0**	**0**	**1**	**1**

Note: The segregation patterns and assigned probabilities for SxN and NxS markers are equivalent. The term SxN here refers to the category of marker with Simplex x Nulliplex segregation; it does not necessarily imply parent 1 is simplex unless otherwise explicitly stated. *c.f.* §3.3 for further clarification.

The algorithm begins at the first marker on the first chromosome and locates individuals for which there was a '1' probability at that position. Moving in both directions away from the marker position, probabilities are updated if it is found that the Kosambi probability at that distance exceeds the probability that was already assigned. The process terminates if a probability of less than 0.5, or greater that the Kosambi-derived probability is encountered. This procedure is repeated for all marker positions along the chromosome, for all simulated chromosomes.

The procedure is then repeated, targeting the '0' probabilities and assigning linked loci with the Kosambi-derived probability if that is smaller. Again, further searching along the chromosome in an individual terminates when a smaller probability is encountered, or a probability greater than 0.5.

Subsequently, the probabilities are normalised so that $\sum_{i=1}^{4} X_i = \sum_{i=5}^{8} X_i = 2$ at all positions. Details of the normalisation procedure can be found in Appendix I. It may be noted that this procedure can easily be extended to any grid of loci along the chromosome in an interval-mapping like approach; use of the marker positions alone was chosen for simplicity for this project.

2.6. Effect of additivity, dominance and QTL segregation

All possible QTL segregations were simulated for a set of combinations of effect sizes and heritabilities as given in Table 3, for a purely additive QTL or a purely dominant QTL on the first chromosome. A second QTL with purely additive effect was simulated on a separate chromosome both times. The maximum test statistic was recorded in each case and the means and standard deviations of these values were determined per QTL segregation type.

Table 3. Simulated combinations of effect sizes and heritabilities to investigate effect of QTL segregation.

QTL1 : Additive / Dominant	**QTL2 :** Additive	
Effect size	**Effect size**	$\boldsymbol{h}^2$
5	5	0.2
5	25	0.2
25	5	0.2
5	5	0.7
5	25	0.7
25	5	0.7

The expected means and variances of a purely additive QTL given an effect size *a* can be worked out by calculating the probability of transmission of alleles to the offspring. These expectation values are provided for reference in Appendix III.

2.7. QTL effect size and heritabilities

In the previous section, different QTL effect sizes were simulated and compared. A second approach to generating different QTL effect sizes is to generate different numbers of equal-sized QTL. When greater numbers of QTL are present, their individual contributions to the genetic variance decreases and so the power to detect each individual QTL diminishes (Beavis, 1998). A power study was conducted for different numbers of QTL and different heritabilities, for a range of mapping population sizes and marker densities to generate power curves for QTL detection. Both additive and dominant-effect QTL were considered, as well as different thresholds for significance ($\alpha = 0.05$ and $\alpha = 0.2$).

The method of determining the power of detection was adapted from a previous simulation study in a diploid crop (Beavis, 1998). For each combination of experimental conditions (#QTL, h^2, flanking marker proximity, population size) 50 linkage groups were generated of 20cM length, with QTL of effect size 5 assigned to the centres of a random subset of these linkage groups. Markers were positioned at the ends of each linkage group (both 10cM away), or by moving one of the makers closer to the QTL (to 2cM away), providing a comparison for the effect of marker distance. In order to declare a marker location as significant, a suitable threshold is required. Therefore, a permutation test with N = 1000 permutations was performed for each set of experimental conditions (marker proximity, number of QTL, population size and heritability). A random set of phenotypic values was generated with mean μ and variance σ_i^2, for some $i \in (1,2,\dots,1000)$. A SNP set was generated with no QTL present (using the same marker proximity), and the *maximum* test statistic (F-value) from all 100 marker positions was recorded for each of the 1000 shuffles of the phenotypic values. At a significance level of α, the $100(1-\alpha)$ percentile of the ordered list was taken as the critical test statistic F_{crit} to control the experiment-wise type I error rate to less than or equal α. This threshold was used in all 1000 replications (for the same experimental conditions).

In the case where *both* markers on a linkage group were found to be significant, the number of QTL identified was counted as one, the same as would occur had only one of the markers been significant. The power, $1-\hat{\beta}$ is defined as the proportion of simulated linkage groups with QTL present that were correctly identified. This approach was also taken in the power study previously mentioned (Beavis, 1998). For example, in a simulation of 3 QTL per marker set, the total number of simulated QTL was 3000 and the total number of linkage groups with no QTL present was 47,000. If (*e.g.*) 789 linkage groups were correctly found to have at least one significant marker, the power estimate for that set of conditions was 789/3000 = 0.263.

The rate of false positives was monitored in an analogous way to the correctly-identified markers – both one or two false positives were recorded as a single linkage group incorrectly identified. This approach tended to slightly under-estimate the number of false hits but was chosen so that the derivation of $\hat{\alpha}$ and $1-\hat{\beta}$ would be consistent.

In this analysis, heritability is taken in the broad sense – as the ratio between the genetic variance and the phenotypic variance. This is acceptable since in many cases polyploid crops are vegetatively propagated (and thus dominant effect loci are of importance) and also because a direct comparison could then be made between the additive and dominant cases.

2.8. Setting significance thresholds

A standard procedure to determine an experiment-wise significance threshold using a permutation test was used (Churchill and Doerge, 1994). Some investigations into the distribution of test statistics from this permutation procedure were conducted (up to 500,000 permutations of data, both with and without QTL present) as well as monitoring the type I error rate in the power studies ($\hat{\alpha}$) and comparing them to the pre-assigned type I error rate α.

2.9. Usefulness of different marker categories

In a tetraploid organism the possible segregations at a locus are Simplex x Nulliplex, Simplex x Simplex, Duplex x Nulliplex, Duplex x Simplex and Duplex x Duplex (these segregations are denoted SxN, SxS, DxN, DxS and DxD respectively). A marker that segregates as Nulliplex x Simplex is also referred to as Simplex x Nulliplex for simplicity *etc.* It should be noted that in this report, these terms refer to the ‘minor’ allele dosages for some locus in the two parents of an F1 mapping population. Other possibilities involving higher dosages (*i.e.* triplex or quadruplex) display the same segregation patterns as members of this set.

The relative usefulness of each marker category in re-constructing the genotype probabilities (using the method described in §2.5) was estimated through a series of simulations.

For each marker category, one hundred marker sets were randomly simulated containing 100 equally-spaced markers at 1cM spacing per chromosome (2 chromosomes in total). The genotype reconstruction function was applied in each case and the resulting arrays were compared to the true probabilities based on the full haplotype information.

A measure of closeness to the “true probabilities” X_i was defined as

$$\delta = \frac{1}{n}\left(\sum_{m=1}^{n}\left(\frac{1}{8}\sum_{i=1}^{8}\left|X_i - \widehat{X_\iota}\right|\right)\right)$$

where $\widehat{X_\iota}$ is the estimated probability of the presence of homologue *i* at a marker position *m* and *n* is the total number of markers (in this case, $n = 200$ since each chromosome had 100 markers). Note that the term “true probability” refers to the value of X_i when the identity of homologue *i* at a position is fully known – so taking either a ‘0’ or a ‘1’ value for absence or presence, respectively.

For each simulated marker set the mean deviation *δ* was recorded. An overall mean for each marker category from the 100 replications was determined to give an estimation of the ‘usefulness’ of that marker category in reconstructing the haplotype probabilities (for a mapping population of 150 F1 individuals). A Tukey’s test was performed to test the significance of the differences found, following an ANOVA on the deviation data using marker category as a treatment to estimate the residual mean square error necessary to calculate Tukey’s yardstick. This relies on the independence of treatments which we may tentatively assume. An alternative test would be an independent 2-sample t-test with the Bonferonni or Šidák correction for multiple comparisons

(using $\alpha' = \frac{\alpha}{n}$ or $\alpha' = 1 - (1-\alpha)^n$ for *n* comparisons, respectively).

The mean deviations *per position* over the 100 replications were also estimated (since each replication consisted of a random SNP set with markers spaced at 1cM, enabling the comparison).

2.10. Effect of marker density

The effect of different levels of marker density was investigated in two manners. Firstly, a dense SNP set over two chromosomes was generated, with 198 markers per chromosome randomly distributed over a length of 100cM – thus 66 markers of each of SxN, SxS and DxN. A single Simplex x Nulliplex QTL was located at the centre of chromosome I and genotype probabilities were reconstructed as previously described (§2.5). Using the model described in §1.11.2, a regression was performed on the parental genotype probabilities and the peak LOP score was recorded. Five markers chosen at random from each chromosome were removed from the marker set after which the process was repeated. Following this, ten markers were chosen at random and removed from each chromosome, and so on in multiples of five until the full marker set was exhausted. The process was repeated 50 times in order to see the effect of marker density on the size of the QTL peak (independent of marker category or position).

A second approach to investigate the effect of marker density was based on the method outlined in §2.9 for estimating marker usefulness. In this approach, the mean deviations were estimated (from 100 replications) per marker category, but with different numbers of markers per chromosome (so, different marker densities). The details of this simulation are given in Table 4.

Table 4. Details of simulated marker sets outlining different levels of marker density investigated. The conditions in each row were replicated over 100 simulations to estimate means.

Nr. Chromosomes	Chromosome length	Nr. Markers	Inter-marker distance	Marker range
	(cM)		*(cM)*	*(cM)*
2	100	10	9.1	9.1 - 91
2	100	20	4.8	4.8 - 96
2	100	30	3.2	3.2 - 96
2	100	40	2.4	2.4 - 96
2	100	60	1.6	1.6 - 96
2	100	80	1.2	1.2 - 96
2	100	100	1	1 - 100

Simulations were run for all five marker categories as well as a mixed marker set containing 20% of each of the five marker categories (randomly assigned to a grid of positions along the chromosomes).

2.11. Effect of mapping population size

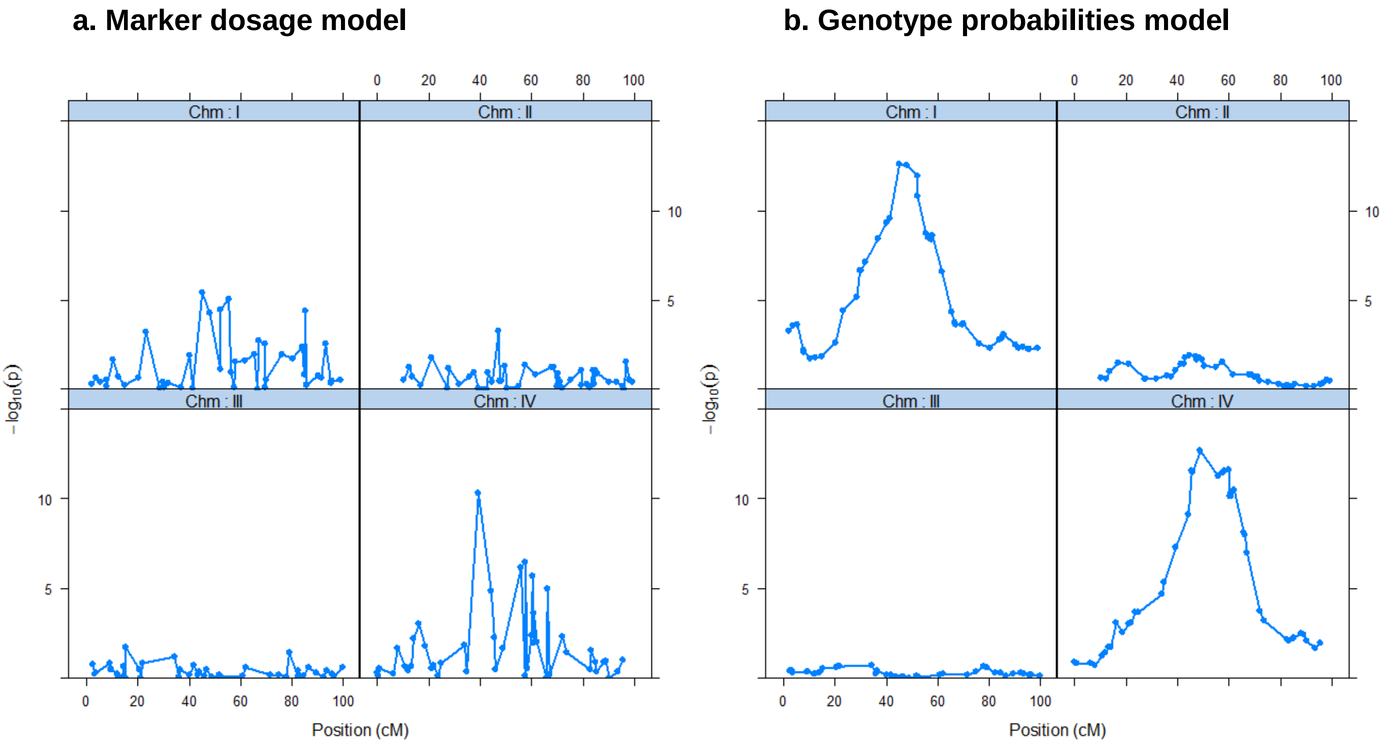
The effect of mapping population size on the power of detection of QTL as well as the maximum LOP score (from a flanking marker) was investigated. The first of these investigations has already been described in §2.7. Given its practical importance, population size was taken as the main variable in the power studies, ranging from a minimum of 50 individuals to a maximum of 600, in increments of 50. A slightly more detailed study for the case α = 0.05 and additive-effect QTL was performed with a range of F1 = 30, 60, ..., 600 individuals in increments of 30.

The relationship between mapping population size and the peak LOP score was also investigated for population sizes of 100, 200, ..., 1000. In particular, it was of interest to determine whether the relationship between the peak LOP score and the number of individuals was linear. For each population size, 1000 replicate marker sets were simulated, each containing three Simplex x Nulliplex QTL of equal additive effect size, randomly allocated to separate chromosomes over 10 chromosomes of length 100cM (so all QTL were unlinked). A (broad-sense) heritability of 0.2 was also chosen. The maximum LOP score over all chromosomes was saved per run, allowing an averaging over the 1000 replicates. It was therefore possible to plot the peak LOP score versus population size (for this set of conditions) and perform a simple linear regression to check the strength of association between the two.

3. Results

3.1. Model implementation

Both models described in §1.11 were implemented in R and tested under various scenarios (number of QTL, QTL effect size, heritabilities, QTL segregation *etc.*). It became clear that a simple regression on marker dosages is too simplistic a model for this type of analysis, although interestingly, it can be shown to be equivalent to a regression of the genotype probabilities in the case where no attempt is made to improve the probabilities from their initial base assignments (*c.f.* Table 2, §2.5). A comparison between the results of both models in a simple scenario with two unlinked QTL on four chromosomes is given in Figure 2. The simulated mapping population size was 120 F1 individuals.



a. Marker dosage model

b. Genotype probabilities model

Figure 2. Comparison between results of regression using **a.** marker dosage scores, **b.** genotype probabilities.

Note: Simulated marker-set had 2 QTL of equal additive effect; 45 markers / chm. using SxN, DxN and SxS markers. h^2 = 0.7

It is perhaps worth noting that in certain situations where a close-flanking marker is linked in coupling phase to the QTL '+' alleles, the Dosage model can actually give a higher peak than the Genotype Probabilities model (GP model). The slight increase in power is a result of the higher degrees of freedom available using the simpler Dosage model over the GP model. However, this relies on the marker and QTL segregation matching, with a very small distance between them (~ 1cM or less).

For the rest of this report, the GP model will be the main subject of investigation.

3.2. Reconstruction of parental homologue identities

In order to apply the GP model, it is first necessary to have a good approximation to the genotype probabilities. As mentioned in §1.12 in the Introduction, various methods have already been proposed and implemented to derive these probabilities from the marker information. It wasn't the primary focus of this project to develop a new or improved method for reconstructing parental homologue identities from marker data, nor to compare different methods of doing this. Therefore, a rudimentary approach was adopted which used the Kosambi mapping function to model probabilities around marker positions where the parental identity of the homologue was known with certainty (presence or absence).

Some refinements were made during the process of implementing this approach, with the final version able to reproduce the results using the full information relatively well (assuming that we have sufficient marker density). The algorithm was applied to the simulated marker-set from the previous section with the resulting significance profile shown in Figure 3.

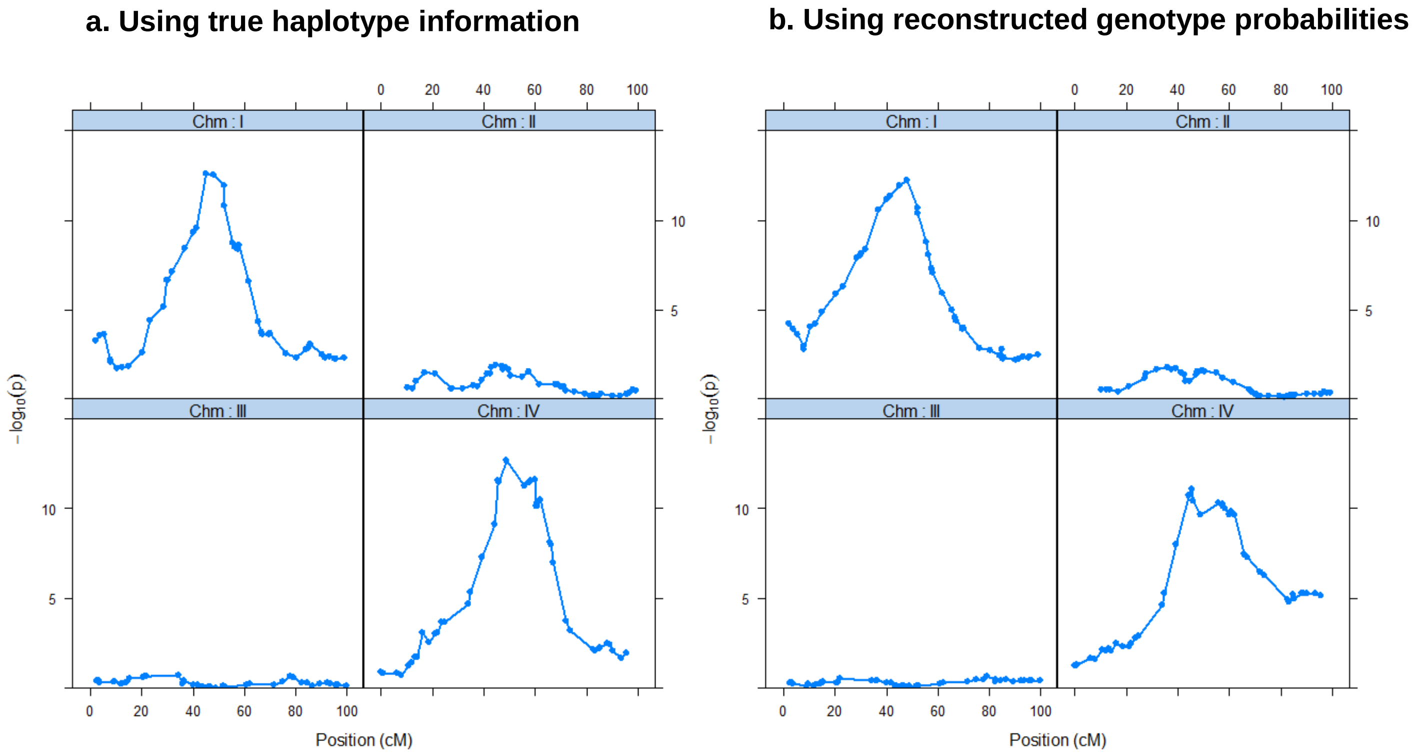


Figure 3. Comparison between results of QTL analysis using **a.** true genotype probabilities, **b.** reconstructed genotype probabilities, using a simulated F1 mapping population of 120 individuals.
Note: Simulated marker-set had 2 equal additive QTL; 45 markers/chm. using SxN, DxN and SxS markers. h^2 = 0.7

The desired profiles are not fully recovered – an area of false significance has appeared near the telomere of chromosome IV that cannot be accounted for. As well as that, the strength of association (judged by the height of the peak) has slightly decreased, particularly on chromosome IV. This has an implication for the power of detection of smaller-effect QTL. Nevertheless, it is clear that the results are quite an improvement over those when the raw dosage scores were used (Figure 2.a). The effect of order of application of genotype reconstruction functions (so using '0' entries first or '1' entries first) is shown in Appendix II for this particular example. It appears that the power decreases and the results disimprove if the ‘0’ information is applied last. For the rest of this report, applying the ‘1’ information last (as shown in Figure 3.b above) was the approach taken. This and other related points are dealt with to a greater extent in the discussion section §4.2.

3.3. Effect of additivity, dominance and QTL segregation

A simulation was run to determine the effect of QTL mode of action and segregation on the detectability of the QTL, producing results that were broadly consistent with those expected. In a regression approach to QTL mapping, a higher level of significance results when a larger amount of variance in the phenotypic data can be explained by the variables in the model. For example, when the gene action is purely dominant, all offspring which have at least one dose of the ‘+’ allele will enjoy the same phenotypic advantage as any other with that allele. This results in only two groups: those with the ‘+’ allele and those without. Depending on the parental configurations (so whether the QTL is SxN or DxD for example), we expect a certain segregation among the offspring which results in a different genetic variance associated with that loci. The size of this variance has a direct impact on our ability to detect the QTL. In contrast, at an additive locus the possibilities for intermediate values increases the number of groups segregating in the offspring for most QTL segregations, leading to higher associated variances. The expected variances based on different QTL segregation patterns and gene actions are given in Appendix III.

The impact of QTL segregation on the ability to detect a QTL was investigated (*c.f.* §2.6). For now this was taken as the magnitude of the peak LOP score near that locus (the topic of power of detection is treated separately in §3.4). If we compare the resulting peak LOP scores it is clear that variance plays at least as important a role as effect size in modulating our ability to detect a QTL. Consider Figure 4.a where a major dominant QTL (effect size = 25) is located on chromosome I and a minor additive DxN QTL (effect size = 5) is located on chromosome II. It should be noted that the legend and colour scheme refer to the QTL segregation on chromosome I – this is the only factor that varies in this data.

It is immediately obvious that we are unable to detect a dominant QTL for any of the segregations which have a triplex or quadruplex parent (since all offspring will inherit at least one of the ‘+’ alleles which is enough to have full expression of the gene effect). In the situations where we can detect the dominant QTL, some of the results are not easy to account for based on a consideration of the expected variance. For example, in the case where QTL1 is dominant and DxD (effect size 25) and QTL2 is additive and DxN (effect size 5), the expected variances (referring Appendix III) due to each of the loci are:

$$\sigma_d^2 = \frac{35d^2}{1296} = 16.88 \qquad \text{and} \qquad \sigma_a^2 = \frac{a^2}{3} = 8.33.$$

In other words, we expect a bigger variance component due to the dominant locus and thus a more significant association. However, we see from Figure 4.a that this wasn't the case – the additive QTL had on average a higher power of detection. It must be remembered though that there is a difference between the expected variance and the actual variance. The proportion of offspring expected to inherit 0000 in a DxD situation is $^{1}/_{36}$. With a mapping population of 150, that equates to four individuals. It is quite possible that fewer than four individuals inherited the 0000 configuration, thus reducing the variance due to that locus and the power of

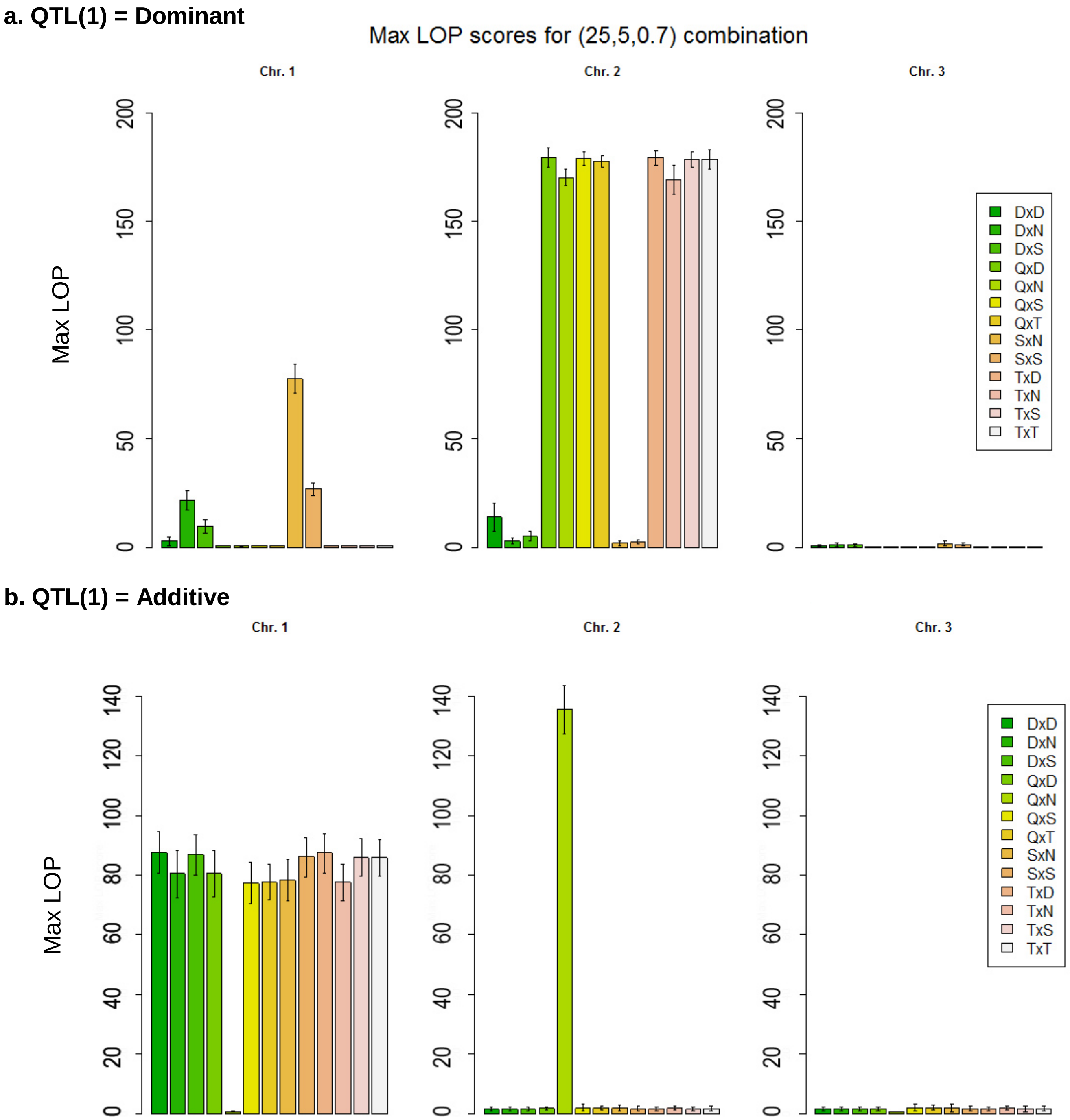


Figure 4. Comparison between the peak LOP score in a two QTL scenario with a major QTL on chromosome I and a minor QTL on chromosome II: **a.** when QTL on chromosome I is dominant, b. when QTL on chromosome I is additive.

Note: Segregation key refers to the segregation of the QTL on chr. I only. The QTL on chromosome II was fixed at DxN (additive).

detection. Although Figure 4 is the result of averaging the effects over all different permutations of a QTL segregation (in the DxD case there are 36 such permutations), a much larger simulation study may have resulted in these fluctuations averaging out (but this needs to be verified).

In the second scenario (Figure 4.b) when both QTL have additive effect there is almost no power to detect the minor QTL on chromosome II. The only exception to this occurs when the QTL on chromosome I is QxN, with no segregation in the offspring.

However, it is unlikely that major-effect QTL would be so treated, since the advantages of including significant markers as cofactors in a model are well-established (Jansen, 1993; Zeng, 1993). There wasn’t time to implement an ‘analysis of covariance’-like approach although this would appear to be a good idea even from these preliminary results.

3.3.1. Relationship between genetic variance and peak LOP

The relationship between the genetic variance due to a QTL and the significance value that results from a QTL analysis using the GP model was investigated further. The results were slightly more difficult to interpret however, and raise some interesting questions of their own.

It is possible to assign QTL-like effects to any marker position in a simulation study by determining the expected phenotypic values in the offspring (assuming a certain heritability) if that marker co-locates with a QTL. This was done for a dense marker-set and an F1 population of 200 individuals with 25 markers of each of the five segregation categories (SxN, DxN, SxS, DxS and DxD) randomly shuffled, so with 125 markers every 0.8cM along the chromosome. At each marker position in turn, a value (+10) was assigned to the ‘+’ allele for that locus and a QTL analysis was performed using the GP model (with full homologue identity information). The QTL itself was excluded each time so that the peak LOP score was recorded from a nearby marker (usually flanking although not always).

The results were somewhat intriguing (Figure 5), with two things worth noting initially. In the case of a dominant QTL there appears to be a linear relationship between the genetic variance and the statistical significance for all segregations *but* SxN, where four groups emerge. These four groups also appear in the case of additivity and also display multiple linearity, although for the highest group it appears that significance is independent of variance. Indeed, the two graphs could be joined together, since in the SxN case the variance is the same under both additive and dominant gene action. The spread of actual variances in the SxN case is also the smallest, due to the fact that there are only two groups segregating in the offspring. The slopes of the lines between the two graphs do not appear to be equal – with a slope of approximately 2.5 in the dominant case (Figure 5.a) and 0.9 in the additive cases (Figure 5.b).

When the process was repeated the same pattern emerged, suggesting that this is a real effect and not a peculiarity of a particular simulation. The exercise was then performed using a simulation where recombination was supressed (all the markers and potential QTL being at the same position). It was found that all data-points in the four bands migrated to the highest band, but the linear band in the dominant scenario was unaffected (Appendix IV).

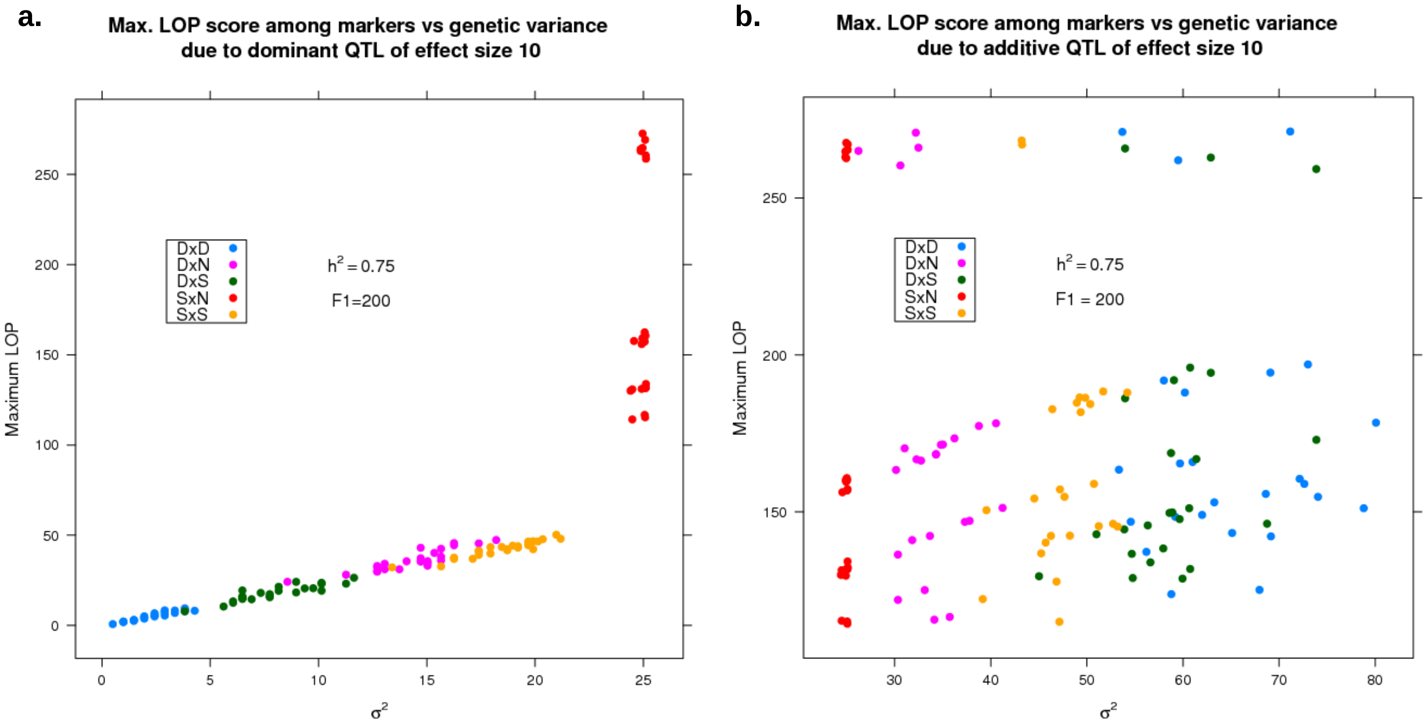


Figure 5. Investigation of the relationship between genetic variance and significance using the GP model for **a.** a dominant QTL, **b.** an additive QTL (of effect size 10).
Note: Results based on a single chromosome with markers evenly-spaced at 0.8cM. F1 = 200 and h^2 = 0.75

3.3.2. Impact of multiple alleles at a locus

Another aspect that was considered was the impact of multiple allelic effects at a locus. SNP markers are chosen to be bi-allelic (to simplify the fluorescence labelling and calling). However, when groups of SNPs are considered together as a block they may describe haplotypes with more than two combinations. One of the challenges and opportunities of QTL mapping in polyploids is the extra allelic diversity that they may offer, which may be associated with a particular (SNP) haplotype.

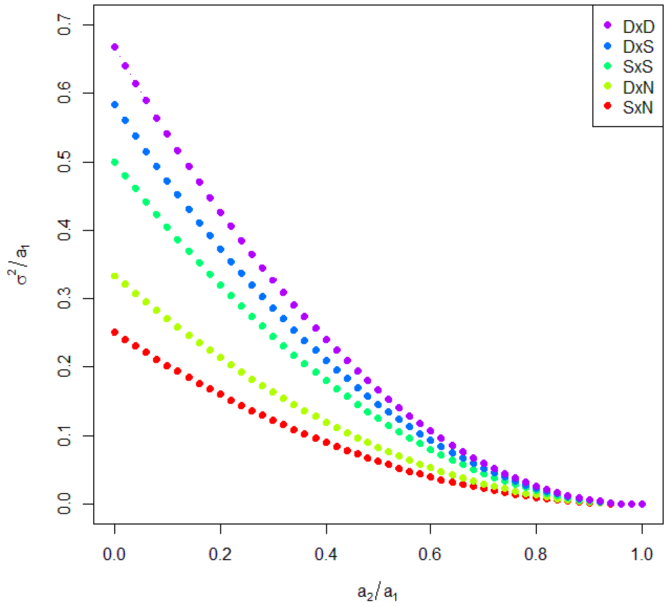


Figure 6. Plot of the variance associated with an additive locus where the effect of the major allele (a_2) varies in comparison to the minor allele (a_1). Maximum variance is realised at loci where the difference in effect size of the two alleles ($a_1 – a_2$) is greatest.

All of the simulations of QTL in this report assume that they are bi-allelic and that one of the alleles confers an effect and the other none. The greatest variance results when the difference between the effect sizes of the two alleles is greatest – so assuming non-negative effects, the case where the '-' allele confers no benefit (Figure 6).

However, at a QTL locus there can be up to eight different alleles present in an F1 mapping population. These may combine in 36 different ways in the F1 offspring. If we know the relative contribution of all eight alleles (unlikely in reality, but possible through simulation) we can easily determine what the expected variance due to that locus is in a mapping population. With 36 classes, we require a relatively large F1 population so that all of the classes will tend to be proportionately represented. A short simulation was run using eight additive effect sizes of (0, 5, 10, ... , 30, 35). Four of these values were randomly assigned to Parent 1 and four to Parent 2, and the expected variance in the offspring was determined. This procedure was repeated 1 x 10^6 times, and a histogram of the resulting genetic variances was produced. It can be seen that with an 8-allele QTL, there are a very large number of possible variances associated with the locus if all effects are assumed to be additive (*c.f.* Appendix V for the plot). In a single mapping population only one of these bars is relevant of course, but it is worth noting nonetheless if considering models which allow for up to eight different alleles at a QTL.

3.4. QTL effect size and heritabilities

It was noted in the previous section that the effect size of a QTL is only part of the picture. The genetic variance associated with different segregation patterns at a locus plays an equally important role in determining the power of QTL detection. Therefore, it might make more sense to model QTL sizes by the amount of genetic variance they explain (indeed this is often

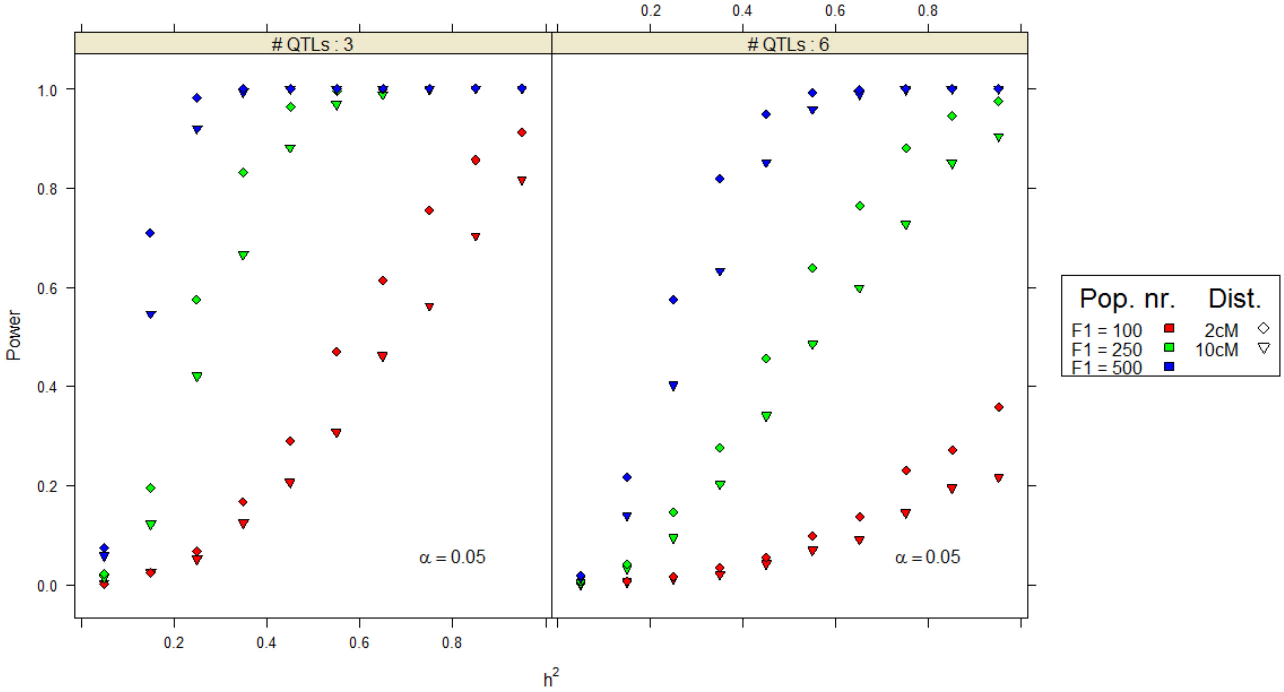


Figure 7. Power of detection of additive-effect QTL versus heritability, for different numbers of QTL, F1 mapping population sizes (Pop. nr.) and flanking marker distances (Dist.).

Note: The experiment-wise error rate was controlled at α = 0.05 by permutation tests with 1000 permutations.

the approach used in simulation studies). A simple way to do this is to simulate various numbers of QTL of equal size so that the proportion of genetic variance explained by each individual QTL is varied.

A reasonably high heritability is crucial for the detection of QTL. At low heritabilities, most of the phenotypic variance will be due to random environmental factors and thus the strength of association between a QTL and the phenotypes will be drowned out by experimental noise. From a breeding perspective heritabilities are of central importance since they determine the amount of (additive) genetic variance that may be fixed through selection (Acquaah, 2012).

The power to detect additive QTL under different heritabilities was derived through simulation, for three population sizes (100, 250 and 500 F1 individuals) and two closest flanking marker distances (10cM and 2cM). The analyses were performed by either assuming 3 QTL or 6 QTL of equal effect (Figure 7, left and right panels). The results when the experiment-wise type I error rate α was set at 0.2 are provided in Appendix VI. The simulations were run with 1000 repetitions for each combination of experimental conditions, for 50 linkage groups with QTL randomly assigned to a number of these groups (with a maximum of one QTL per linkage group, so unlinked QTL). The GP model was used using the full parental homologue identity information. Therefore, the results represent the best possible scenario – a less than 100% accuracy in the genotype probability predictions would be expected to affect the power downwards (needs to be verified). It would have been interesting to compare the power estimates when reconstructed probabilities were used but this would have taken an inordinate amount of computing time.

The topic of significance testing and threshold setting will be dealt with in the next section (§3.5) – for now, it will just be noted that an experiment-wise threshold was set using a permutation test with N=1000 permutations.

We can see that in the 3 QTL case where the heritability is 0.5 and the closest flanking marker to every QTL is 10cM, the power of detection ranges from 25.7% when F1 = 100 individuals, to 92.5% with 250 individuals and 100% with 500 individuals. With 3 QTL and a heritability of 0.5 each QTL explains 16.67% of the phenotypic variance. In the 6 QTL case with $h^2 = 0.5$ each QTL explains 8.33% of the phenotypic variance. With 10cM between the QTL and the flanking markers, the power estimates become 5.6% when F1 = 100, 41.3% when F1 = 250 and 90.5% when F1 = 500 individuals.

Increasing the significance threshold by allowing a type I error rate of 0.2 had the effect of slightly raising these figures (*e.g.* in the last estimate for F1=500 individuals and a distance of 10cM to the nearest marker, the power to detect a QTL that explains 8.33% of the variance was increased from 90.5% to 96.8%)

3.4.1. Power using the Dosage model

The same approach was taken with the Dosage model for an additive QTL using $\alpha = 0.05$.

This model is clearly inferior to the GP model (Figure 9). In particular, it was noted that very significant F-values were produced during the permutation of random phenotypic data, resulting in high thresholds: ~ 13 in comparison to ~ 4 for the GP model (Figure 8). The power to detect true QTL was thus significantly diminished (Figure 9).

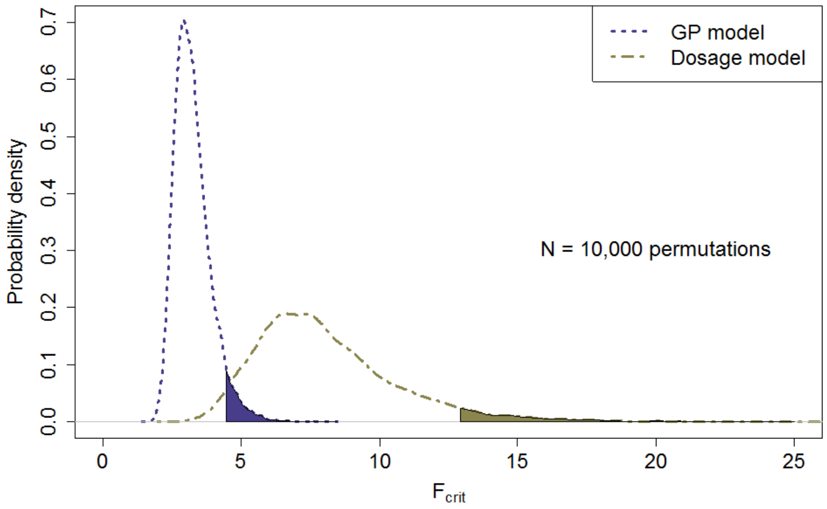


Figure 8. Distribution of the (maximum experiment-wise) test statistic from 10,000 permutations of random trait values using GP model and Dosage model. Shaded regions represent 5% tail.

Apart from the drastic reduction in power, the power curves appear to fluctuate, suggesting that many QTL produced test statistics may have been close to the detection threshold of 12 or 13 and therefore lost due to the high stringency imposed through $\alpha = 0.05$. The results for $\alpha = 0.2$ are included in Appendix VI for reference.

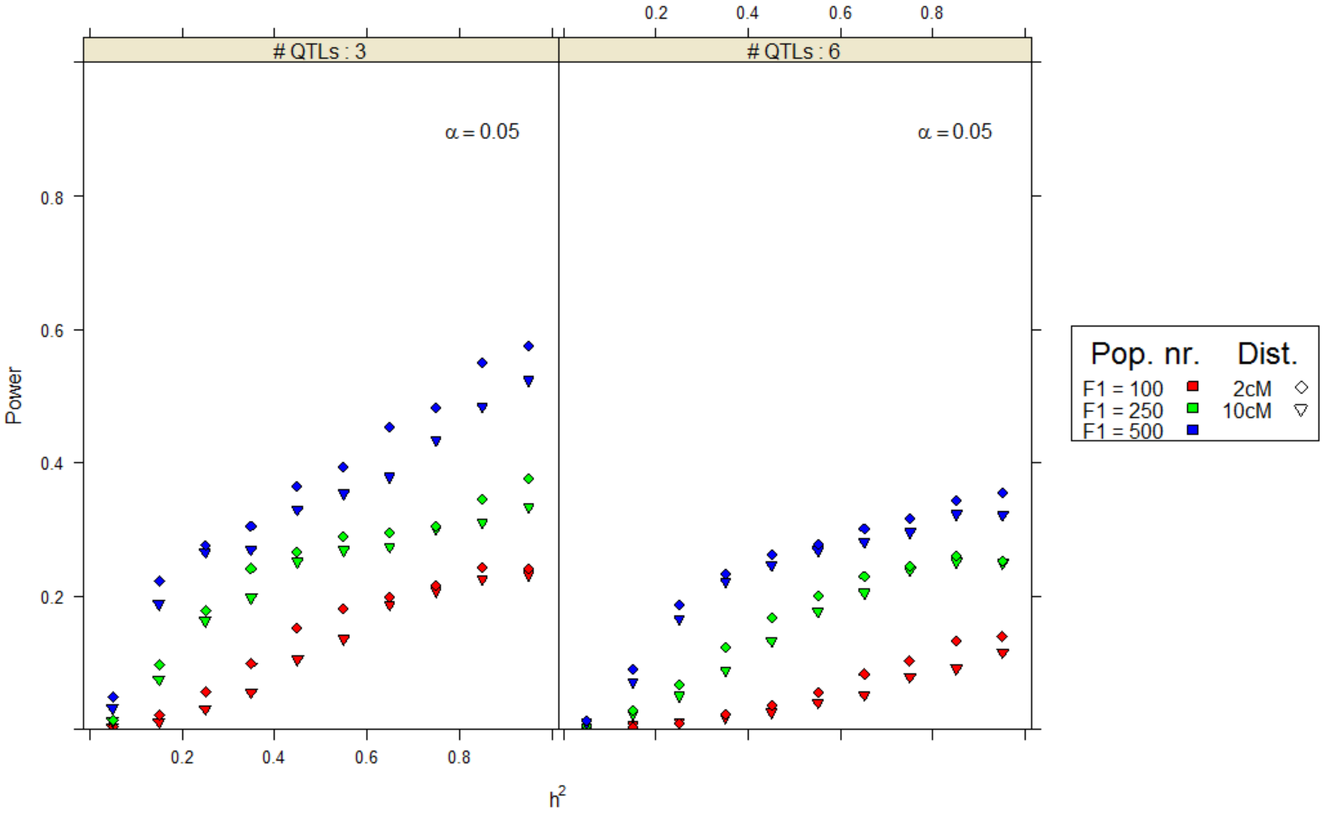


Figure 9. Power of detection of additive-effect QTL versus heritability using the dosage model, for different numbers of QTL, F1 mapping population sizes (Pop. nr.) & flanking marker distances (Dist.)
Note: The experiment-wise error rate was controlled at $\alpha = 0.05$ by permutation tests with 1000 permutations.

3.5. Setting significance thresholds

As outlined in the introduction (§1.13), there are quite a number of different methods for setting significance thresholds, but the method that was adopted for this project was the permutation test (Churchill and Doerge, 1994). The main characteristics have already been highlighted: a permutation test is a non-parametric test which is distribution-free, very simple to apply and has been shown to provide good estimates, but may be computationally-demanding. Two particular topics were investigated in the course of this project:

1. How many permutations are needed;
2. What is the distribution of the test statistics.

3.5.1. Number of permutations (N)

One of the more important questions in permutation testing is "How much is enough?". The ability to answer this question has implications both on the feasibility of using permutation tests in an experiment as well as the number of QTL versus false positives that are likely to be detected. The general procedure in conducting a permutation test is to permute the trait values and randomly assign them to individuals, thus breaking the association between QTL and phenotypes while retaining the mean and variance of the phenotypic data. An alternative is to calculate the mean and variance of the phenotypic data and generate a random sample of phenotypes from a Normal distribution with this mean and variance. Of course, the latter is assuming normality of the data which the former does not. Both were tested and compared to see what difference this made, if any.

The number of permutations that were needed before the critical test statistic stabilised in a single run of 500,000 permutations was examined (Figure 10). There are a few features worth pointing out. Firstly, it appears that the F_{crit} values using the original data were in general higher than the stable estimate (F_{crit} = 3.50) whereas the opposite was the case when random phenotypic data was used. Both of these tended towards the same stable value of F_{crit} eventually, which occurred after 50,000 – 55,000 permutations. There was very little difference in the results for 50,000 versus 500,000 permutations, although 50,000 is still a

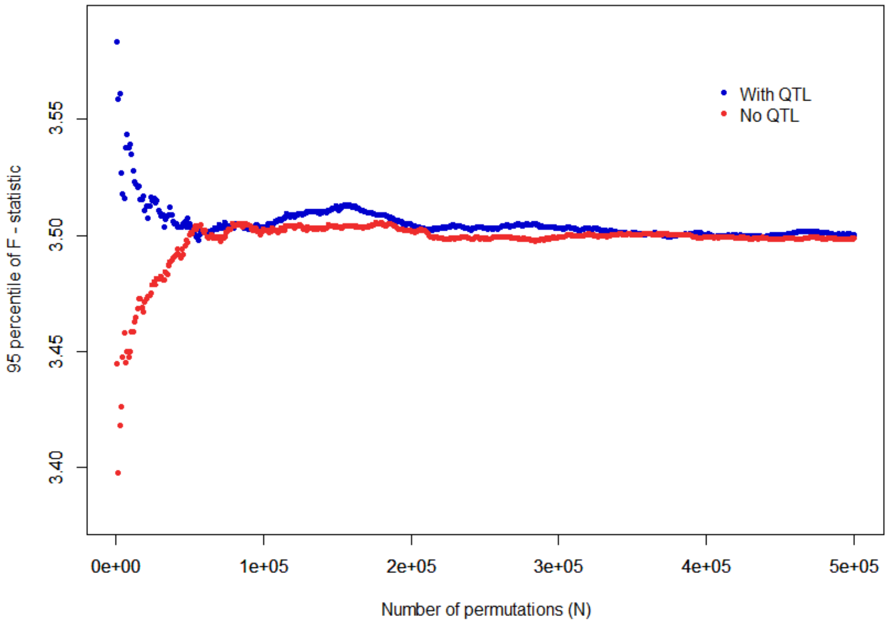


Figure 10. Behaviour of 95-percentile of the F-statistic as the number of permutations increases.

great deal larger than the N = 1000 permutation recommended for estimating the 95-percentile. It is also interesting that the curves for the data with and without QTL present converge towards the stable estimate at more or less exactly the same value of N. It is not true that the curves always tend towards the asymptote from different directions – when the same approach was taken using the 99-percentile, both curves approached the asymptote from above, although not as smoothly (Appendix VIII).

One may ask – does this actually matter? The range of values of F_{crit} for the number of permutations N between 1000 and 500,000 was (3.498, 3.583) for this particular simulated data-set, with a final value of approximately 3.5 (for the case 'With QTL'). If N=1000 permutations were used as recommended (Churchill and Doerge, 1994), the critical test statistic would have been 0.08 higher than the stable estimate. The question then is – how good is good enough? If we are willing to accept an inaccuracy of 2.3% above the stable value, then we have gained a big efficiency over running to 50,000 permutations (Figure 8). This topic is discussed further in the discussion section (§4.4) with some recommendations for further work.

3.5.2. Distribution of the test statistic

It would also be useful to know the distribution of the maximum F-values, particularly if N=1000 permutations are to be used, so that we may assign levels of significance smaller than p=0.001. Given that we are looking at a set of F-values, it was first assumed that these values might be F-distributed under the null hypothesis. However, by simulation it was found that the peak of the probability density function appeared to lie somewhere close to 2. It can be shown analytically that the peak of any F-distribution cannot exceed 1 (Appendix IX). Van Ooijen (1999) remarked that "the distribution of the maximum of a series of LOD scores cannot be determined in a straightforward manner". However, there are some good fitting routines that have been since implemented in R that may assist us in this (Delignette-Muller et al., 2014). Using the decision tree provided in a quick guide to distribution fitting (Damodaran, 2007), it was noted that the distribution followed the path:

Continuous data → Asymmetric → Outliers mostly positive → Lognormal / Gamma / Weibull distribution.

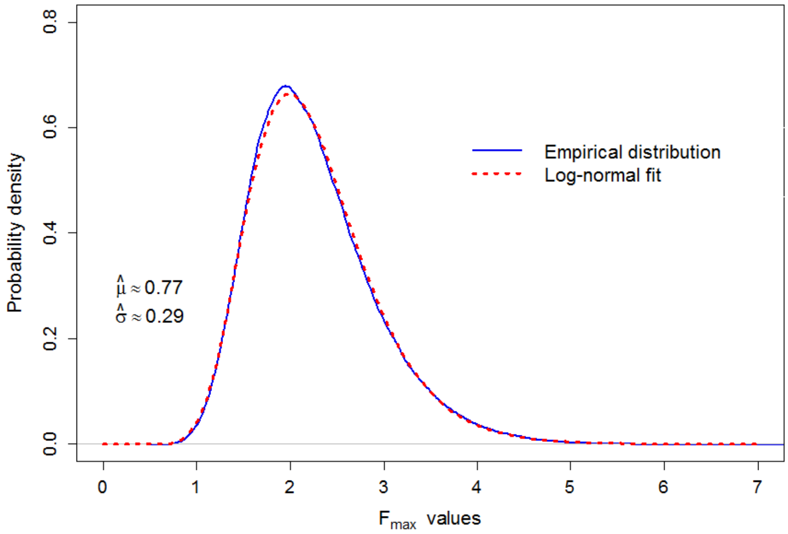


Figure 11. Empirical distribution of the critical F-values from 500,000 permutations with a fitted log-normal distribution shown (dashed line).

Using the maximum likelihood approach, a log-normal distribution was fitted to the data, producing an extremely good fit (Figure 11) with a log mean estimate of 0.7704 ± 0.0004 and a log standard deviation of 0.2887 ± 0.0003. It therefore appears that for this data-set at least, the maximum test statistics follow a log-normal distribution.

3.6. Usefulness of different marker categories

Different marker categories provide different amounts of information, and this is especially true near a QTL where issues such as linkage in coupling or repulsion are important. As part of the process of reconstructing genotype probabilities, we would like some measure of the relative usefulness of different marker categories in correctly predicting these probabilities. As mentioned in the Materials and Methods section (§2.9), a measure for the accuracy of the genotypic prediction was defined as the average of the absolute deviations from the true probabilities. This quantity (referred to as the mean deviation, δ) was determined when the marker set consisted of only one type of marker for each of the five marker categories already introduced, as well as a sixth marker set where equal numbers of these five were randomly mixed together.

Running simulations over 100 randomly-generated marker sets, it was found that on average the best markers to use (for the method described in §2.5) to reconstruct the genotype probabilities were Duplex x Nulliplex markers (Figure 12).

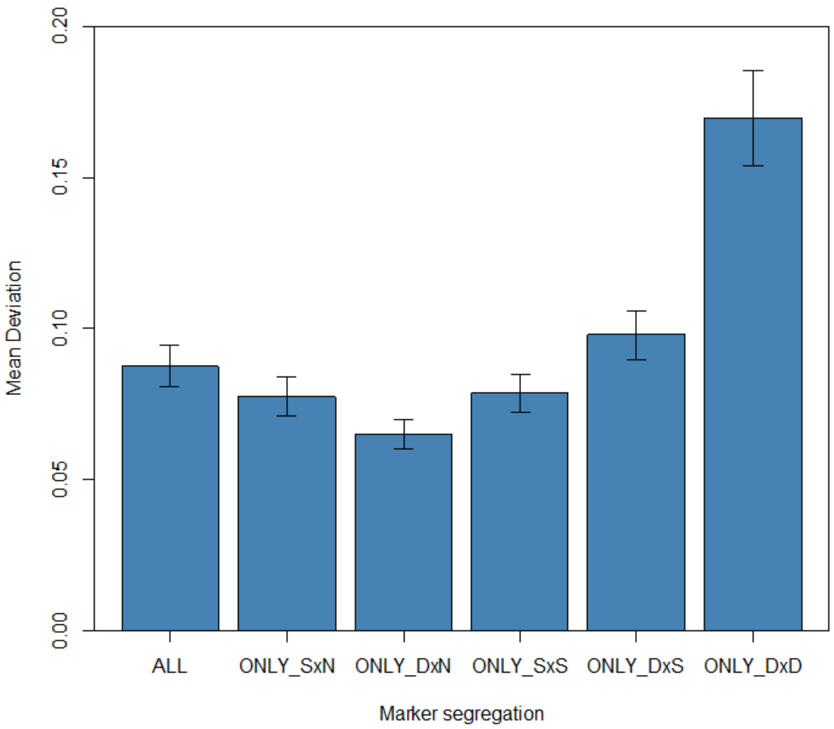


Figure 12. Average deviation from true probability over 100 replicated marker sets

Tukey's test was used to test the significance of the differences between the means for each marker category (Table 5). It was found that DxN markers performed significantly better than all other groups, whereas DxD markers performed significantly worse. The usefulness of each marker category can also be directly inferred by consideration of the number of ‘0’ and ‘1’ entries expected in the initial assignment of genotype probabilities based on the raw dosage information (Table 6). With a more sophisticated method of reconstructing the genotype probabilities, it is likely that the mean deviation estimates could be reduced further (*c.f.* Discussion §4.3).

Table 5. Differences between mean deviations per marker category, showing Tukey’s yardstick. *Note*: significant differences are highlighted in red.

	No DxN	No SxS	No SxN	All	No DxS	No DxD
No DxN	0.000					
No SxS	0.013	0.000			Tukey's W:	0.003
No SxN	0.012	-0.001	0.000			
All	0.023	0.009	0.010	0.000		
No DxS	0.033	0.019	0.020	0.010	0.000	
No DxD	0.105	0.091	0.092	0.082	0.072	0.000

Table 6. Expected frequency of base probabilities in X_i arrays for different marker types

Marker segregation	Proportion of ‘0’ or ‘1’ probs. assigned	% of ‘0’ or ‘1’ probs.
Duplex - Nulliplex	1/6	16.67%
Simplex - Nulliplex	1/8	12.5%
Simplex - Simplex	1/8	12.5%
Duplex - Simplex	5/48	10.42%
Duplex - Duplex	1/18	5.56%

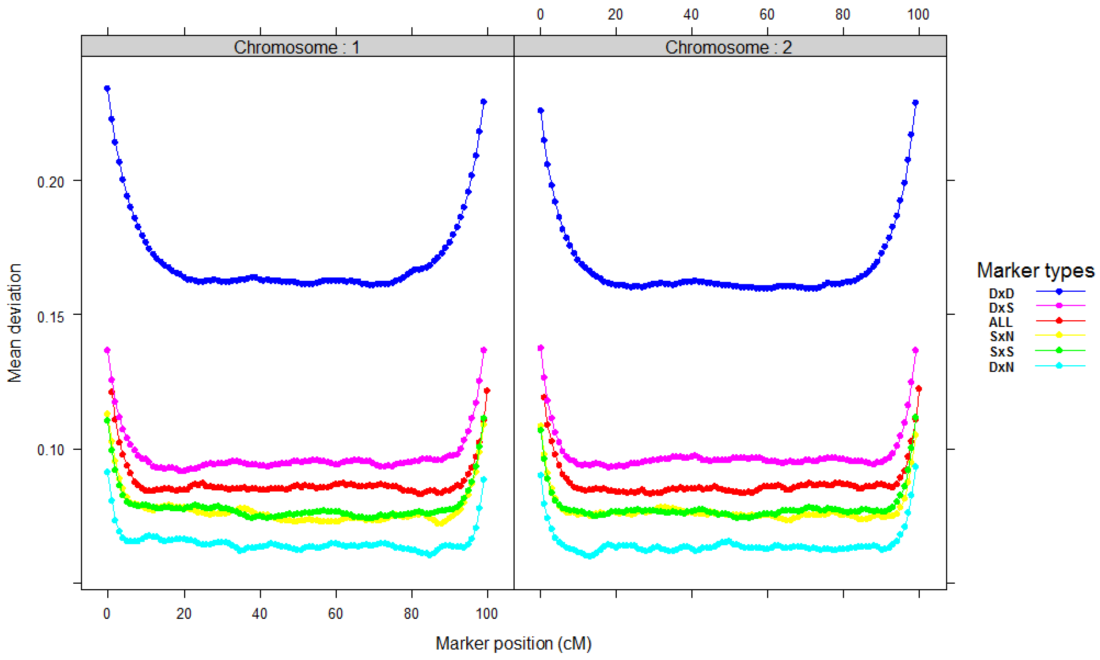


Figure 13. Mean deviations from true homologue identities (100 replications) versus position on the chromosome. *Note:* simulated population size of 150 individuals.

It is also interesting to consider how this deviation measure varies with position along the chromosome. To this end, the simulations were repeated, but rather than estimating a single mean over all marker positions, the means per position were derived. This was possible because for each replication, random marker configurations were assigned to a grid of positions along 2 chromosomes, allowing comparison over the 100 replications (Figure 13).

As we can clearly see for all marker categories, there is a sharp drop in the accuracy of prediction towards the ends of the chromosomes. For example in the case of DxD markers (dark blue line) on chromosome I, the mean deviation at the telomeres was 0.2467, almost twice the mean deviation at the centre (0.127). This reflects the fact that at the centre of the chromosome there is essentially twice as much information (coming from both sides) as there is at the telomeres. Possible improvements in the strategy to reconstruct genotype probabilities will be dealt with in more depth in the discussion (§4.3).

3.7. Effect of marker density

Marker density is a somewhat misleading term, since we may have a 'dense marker map' and yet still have large gaps where no markers are present. However, the term is loosely employed here to refer to the number of markers per cM with the possibility of gaps allowed. Two investigations were performed to investigate the effect of marker density on our ability to apply the techniques already outlined – the effect of marker density on the size of the peak LOP score when a QTL is present, and the effect of marker density on the accuracy of the reconstructed genotype probabilities (which will in turn have an impact on the LOP score, as higher accuracy will result in a greater amount of variance being explained by markers near the QTL).

3.7.1. Effect on peak LOP score

For the first of these approaches, a very dense SNP map was simulated as already described in the Materials and Methods (§2.10), with almost 200 markers / chromosome and a single QTL on chromosome I. Markers were removed in multiples of 5 (independent of previous

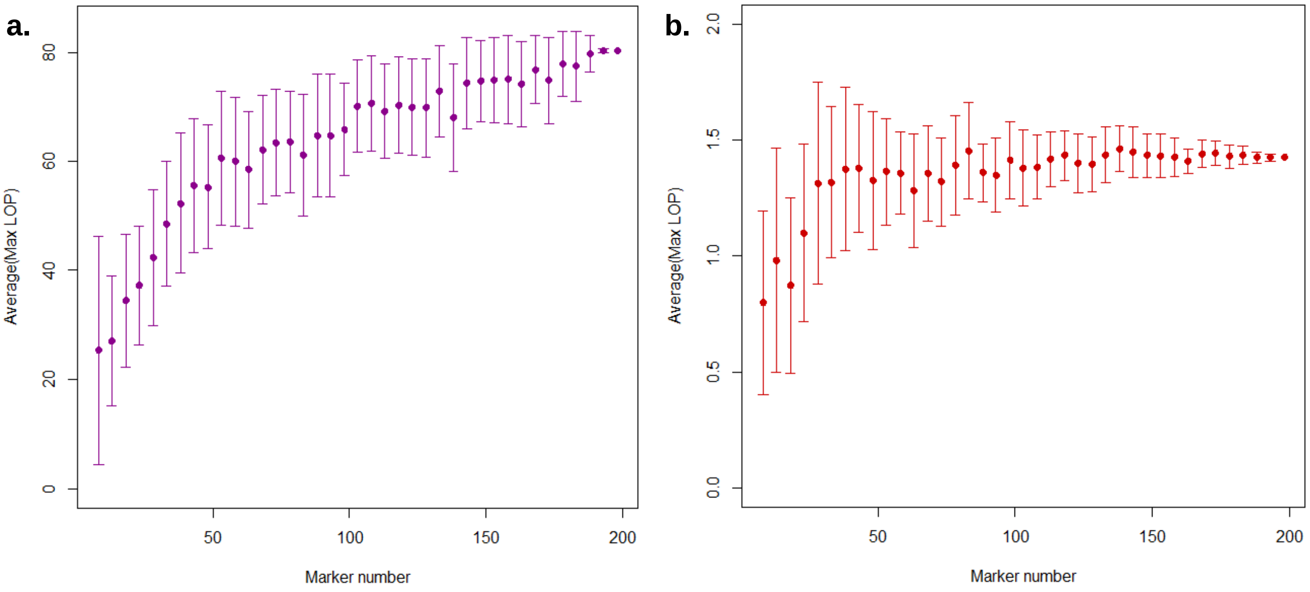


Figure 14. Peak LOP scores v's marker density: **a.** Chm. I with single additive SxN QTL at centre, **b.** Chm. II, with no QTL present. $h^2 = 0.6$. *Note:* SxN, SxS and DxN markers make up the marker sets.

selections) and the peak LOP score was determined through regression on the reconstructed genotype probabilities.

We can see that in the case of chromosome I, there is a gradual and steady increase in the peak LOP score as the numbers of markers increase (Figure 14). The rate of increase appears to be maximal in the range 0 – 60 markers, suggesting that most gains in QTL power and precision will occur when marker numbers are increased within in this range. On chromosome II it is reassuring to see that as marker density increases there is no corresponding increase in the peak LOP score which implies that the false positive rate does not increase despite the increased numbers of markers (and hence increase in the number of separate tests that are performed). It should be remembered that this data was generated from a single ultra-dense marker set with markers being removed from the full set rather than added. To gain further insight into the relationship between marker density and the accuracy of QTL detection a second, slightly different approach was adopted.

3.7.2. Effect on genotype probability estimates

As already described in §2.10, different marker densities were simulated and the mean deviation in the genotype probabilities was calculated from 100 replicated marker sets (Figure 15). It was found that with higher numbers of markers (so with a smaller inter-marker distance), the mean deviation from the true genotype probabilities decreases. This is not surprising, and confirms what was seen in Figure 14. Indeed, there appears to be a similarity in the shape of the curves in both figures. Using the data for SxS markers as an example, it was found that if marker numbers were log-transformed there was a linear relationship between the deviation and the log of the marker numbers ($R^2 = 0.994$).

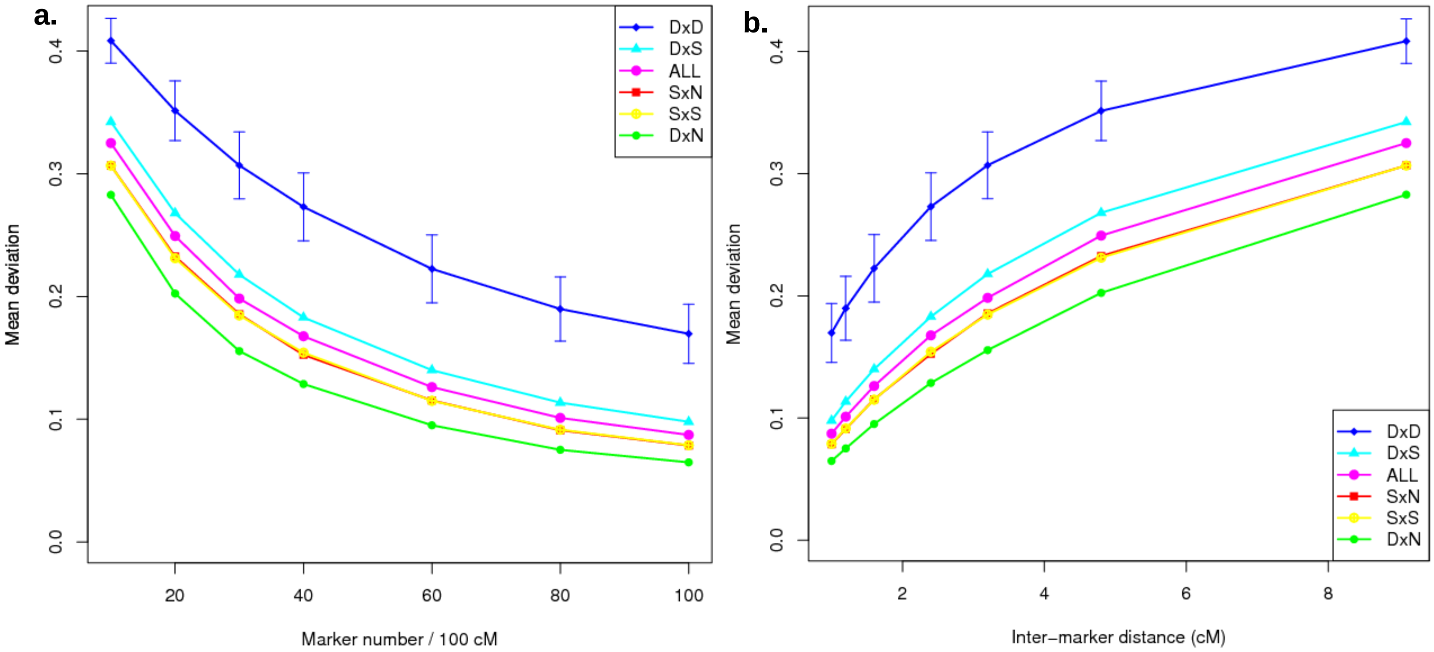


Figure 15. a. Mean deviation from true genotype probabilities (for different marker categories) versus number of markers per Morgan. **b.** Mean deviation from true probabilities versus inter-marker distance.

Note: Error bars in DxD case represent standard deviations from 100 replicated marker sets over 2 chromosomes. Other error bars are omitted for clarity, but are of comparable size

The resulting linear equation (in the SxS case) was

$$y = -0.1006(\log_e N) + 0.532$$

where N is the number of markers.

A series of other possible functions were tested but this produced the best fit to the data. This equation predicts that the mean deviation should equal zero when N = 198. The same approach was used for the other marker types, with the predicted numbers of markers needed before the genotype probabilities tend to their true values is given in Table 7.

Table 7. Predicted number of markers per Morgan needed to accurately reconstruct genotype probabilities using the methods described in this report

Marker category	Regression line	R^2	x-intercept	N_0 / M
S x N	$y = -0.101 \log_e(N) + 0.534$	0.994	5.283	197
S x S	$y = -0.101 \log_e(N) + 0.532$	0.994	5.288	198
D x N	$y = -0.095 \log_e(N) + 0.49$	0.984	5.145	172
D x S	$y = -0.109 \log_e(N) + 0.59$	0.996	5.429	228
D x D	$y = -0.107 \log_e(N) + 0.665$	0.995	6.196	491

These figures should be taken merely as a general guide to the required marker densities needed to accurately reconstruct the genotype probabilities. In reality, it is unlikely that a precisely zero deviation will ever be reached using the methods described, but that we can get “asymptotically” close with these numbers of markers per Morgan. This will be further explored in the Discussion (§4.5).

3.8. Effect of mapping population size

3.8.1. Effect on power of detection

Knowing the correct population size to use is of paramount importance when designing an experiment. In an ideal situation, the researcher should know beforehand how many individuals are needed, given prior information on the expected heritability of the trait of interest, the size of the QTL and the density of the marker map. Of course many of these parameters will be unknown. Increasing population size is one way of increasing the power to detect QTL, but comes at a cost. Efficient experiments should take this into account in order to maximise the benefits from a mapping study.

Previously (§3.4) the results of power simulations where heritability was the main variable were presented. In this section, the size of the mapping population was used as the main varying factor. Given the time it takes to run multiple simulations, two heritabilities were chosen (h^2 = 0.2 and h^2 = 0.7) as well as two sets of QTL number (3 and 6 equally-sized unlinked QTL). For example, in the situation where h^2 = 0.2 and #QTL = 6, then 20% of the phenotypic variance is explained by 6 QTL in which case each QTL contributes 3.33% to the phenotypic variance. As before, markers were assigned either to both ends of each linkage group (of 20cM) or one of the markers was at the end and one closer to the centromere (2cM from the centre). The simulations were run with 1000 repetitions for sets of 50 linkage groups with the QTL randomly assigned to the centre of a number of these groups (unlinked QTL).

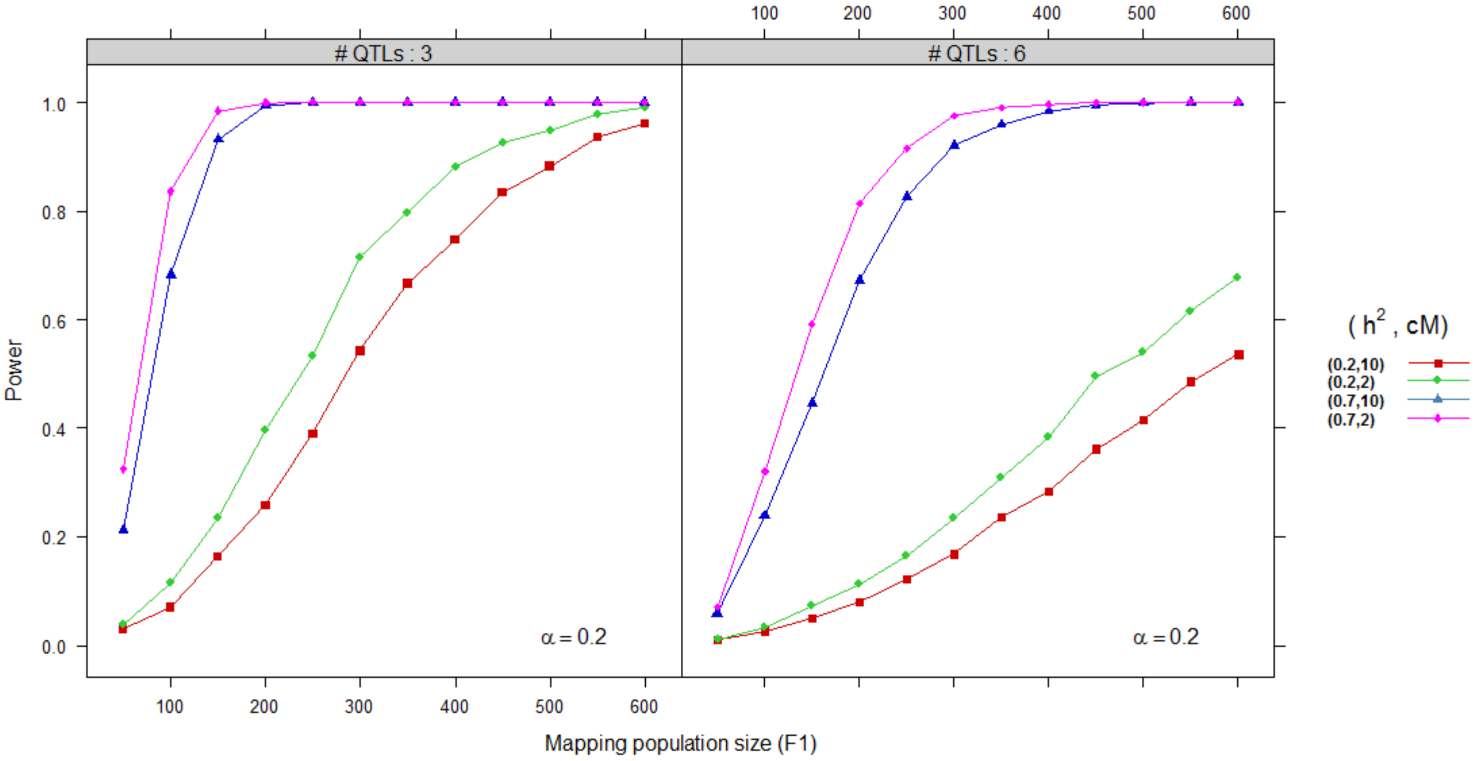


Figure 16. Power curves for additive SxN QTL versus size of the mapping population. Each data-point represent the average proportion of QTL detected over 1000 replicated simulations.

Note: Significance thresholds were set by 1000 permutations of random trait values and selecting the 80-percentile (α = 0.2) of maximum test statistics. cM in the legend refers to the distance between QTL and its closest marker.

As before, we can see that as the number of individuals in the mapping population is increased, there is a steady increase in the power of QTL detection. Some slight deviations were noted, although the shapes of the curves are pretty clear. Power curves were also determined for additive QTL when α = 0.05 and for dominant QTL when α = 0.05 (Appendix VII).

3.8.2. Effect on peak LOP score

It is also of interest to know how the peak LOP scores are affected by the size of the mapping population. Intuitively, one could imagine that as the size of the mapping population increases so too should the height of the peak in LOP values, since there is greater support (and more available degrees of freedom) from the larger population.

According to this simulation the size of a QTL peak increases linearly as a function of population size (Figure 17). It may be interesting to confirm this result for other QTL segregation types or heritabilities, but it is expected the results would be similar across different experimental conditions.

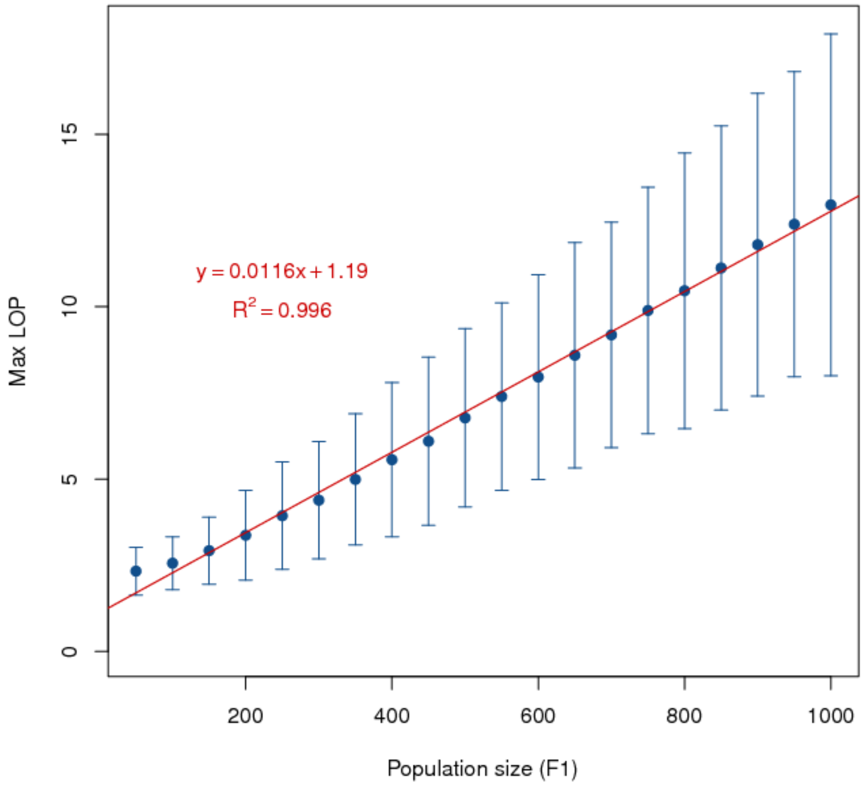


Figure 17. Peak LOP score across all chromosomes versus size of the mapping population.

Note: Each data-point represent the average from 1000 replicate simulation (error bars show standard deviation). Three SxN QTL of equal additive effect were randomly assigned to one of 10 chromosomes in each run, singly (unlinked QTL). h^2 = 0.2

4. Discussion

Although the methods used in this project were relatively simplistic, some insights were still gained into how one might tackle the question of QTL analysis in a polyploid. More work needs to be done in this area but for now some of the points raised during the previous sections will be dealt with.

4.1. Choice of model

When two single marker approaches were compared it was found that the results using the GP model (regressing the trait values on the genotype probabilities) were superior to those using a regression on the raw marker dosages. This superiority was evident in both the power of QTL detection as well as the ability to locate a QTL even when nearby markers are not linked in coupling phase. Indeed, one of the major advantages of using the GP model is that information contained in coupling-phase markers may be extended to other positions, increasing the likelihood of picking up QTL. This is not the case with a simple regression on marker dosages, where markers act alone in the analysis and only prove significant if they happen to be both nearby and linked in coupling phase (*c.f.* Figure 2.a, §3.1).

It should be pointed out that in extreme cases where one of the markers is closely flanking a QTL and is in coupling phase, the level of significance can be slightly higher using the Dosage model rather than the GP model (because of the extra degrees of freedom available – fewer parameters to estimate with the simpler model). It is also very fast to implement, taking a fraction of the time it takes to implement a more complex model such as the GP model (particularly so if the haplotype identities must first be reconstructed). It might therefore be a good idea to run a simple regression using the marker dosages in order to check whether any major QTL happen to be closely linked in coupling phase to a marker. These markers might then be used as covariates in a more sophisticated model. Indeed, this was the first step in the approach taken by Bradshaw *et al.* (2004).

Hackett *et al.* (2013) used an iterative weighted regression on the genotype probabilities, with much of the methodology having already been developed in an earlier paper (Hackett et al., 2001) and subsequently applied to real data for mapping resistance QTL to late blight in potato (Bradshaw et al., 2004). Given the more advanced techniques employed in reconstructing genotype probabilities with a hidden Markov model and a weighted regression, it is likely that their results are superior to what was achieved during this project.

It will be interesting to see what further developments take place in this area, particularly given that regression approaches have tended to be replaced by maximum likelihood methods in diploid analyses. Maximum likelihood estimates may be considered desirable for several reasons, including invariance of the maximum likelihood function (*i.e.* the MLE of $f\ (\theta)$ is given by $f\ (\mathrm{MLE}(\theta))$) as well as the fact that the asymptotic distribution of $\mathrm{MLE}(\theta)$ can be derived (Chen, 2014). Maximum likelihood estimates were used by Hackett et al. (2013) in their map construction for estimation of recombination fractions and LOD scores but not for QTL mapping itself. The use of maximum likelihood is not considered better *per se* (although it has certain attractive properties) but it would be interesting nonetheless to compare more models than just the basic ones implemented in this project. Moving from a single marker analysis to multiple markers, or applying composite interval mapping (Zeng, 1993) or multiple QTL mapping (Jansen, 1993) might also be worth investigating.

4.2. Power study analysis

The design of the power study was adapted from one presented by Beavis in the book ‘Molecular Dissection of Complex Traits’ (1998) in a chapter about power, precision and accuracy in QTL mapping. By design, all simulated QTL were unlinked to other QTL, which made the analysis more straightforward. Another possible approach might have been to simulate fewer linkage groups with more markers on each, closer to the situation in real life. If this approach were taken, the calculation of power would have to be changed, counting the number of significant markers within a defined distance of the QTL position. How large should this window be? Would power be calculated as the fraction of significant markers within QTL windows? Would a significant marker that fell just outside the window be considered a false positive? When such questions arise, the approach of Beavis does appear to have its advantages given that windows are created (±10cM from the QTL when it exists) as a separate linkage group with no linkage beyond its boundaries. Thus the approach taken was clear, simple, and avoided the possible issues just raised. In reality however, these issues are at play and cannot be so easily evaded.

Another approach taken in power studies has been to use the asymptotic distribution of the test statistic to estimate the power. When H_0 is false (so, when a QTL is linked to the position being tested) the test statistic follows a non-central F-distribution with p-1 and p(n-1) degrees of freedom (Cohen, 1988). The power of a test to detect the alternative hypothesis (presence of a QTL) is equal to the area under the non-central F-distribution to the right of the critical value for the test (Cohen, 1988). This approach has been implemented in R in the `pwr` package (Champely, 2009).

In QTL analysis multiple tests are performed. There are thus multiple asymptotic F-distributions to consider. It was assumed that the closest flanking marker would provide the maximum power of QTL detection - preliminary tests were carried out using this approach. However, a direct comparison with Beavis’ approach was not feasible on the PBR cluster. Packages such as `pwr` need to be installed before they can be used, making them unavailable on the cluster unless they are installed there first. Therefore, I am not in a position to say whether the derived power estimates (by simulation) and those using closed formulae are similar. As was pointed out by Beavis, analytic methods rely on knowledge of the asymptotic distribution of the test statistic and are applicable when t and F statistics are available. Other models (such as interval mapping or multiple QTL model methods) do not have known distributions; therefore Monte Carlo simulations are more appropriate for these approaches (Beavis, 1998). For this project, F tests were available and therefore analytical power estimates could have been applied. It would have been an interesting check but was not pursued in the end.

Looking now to the results from the power simulations (§3.4) it is clear that the size of the mapping population and the QTL effect size (in terms of % total variance explained) are both very important for QTL detection. QTL effect size is essentially out of the experimenter's control, although suitable choice of parents (contrasting for the traits of interest) can make a big difference in this. By carefully controlling the experimental conditions, the genotypic variance may explain a greater proportion of the phenotypic variation (by reducing the level of environmental noise) which can also amplify the power. Probably the simplest way to gain extra power is to add more individuals in the mapping population – although after a certain point this will be neither economically nor practically feasible.

It is worth pointing out that the results from these power simulations do not compare favourably with the power estimates derived by Beavis for a diploid. Using a significance threshold of 0.25, Beavis estimated that with an F2 population of 100, an additive QTL that explained 2.375% of the total variance would be detected with a power of 6%, rising to 46% when 500 individuals were used (Beavis, 1998). In contrast, using a significance threshold of 0.2, I estimated that with an F1 population of 100, an additive QTL that explained 2.5% of the total variance would be detected with a power of 2%, rising to 25% when 500 individuals were used. Beavis used likelihood-based interval-mapping in his approach which, together with a slightly higher significance threshold, probably accounts for this difference. This should serve as some motivation for looking at other QTL models which might result in some gain in power.

4.3. Reconstruction of genotype probabilities

It was noted in the results section §3.2 that the method for reconstructing genotype probabilities used throughout this project was a rather basic first attempt and could be improved upon considerably. To expand on this topic slightly, there are a few points to consider. The first is that no attempt was made to recover information from intermediate marker dosages. The problem with intermediate dosages in a tetraploid (or any higher ploidy level) is that it is impossible to predict the identity of the '+' allele using dosage information at the marker position alone. For instance, in the case of a SxS marker where an individual has a dosage of '1' we know that this is the result of inheriting one of the '+' alleles that both parents carry, but we are unable to say which one. The probability of the alleles coming from either parent is 0.5 which corresponds to having no information at all. Therefore we would like to be able to exploit the information contained in intermediate dosages in order to increase our predictive power.

The importance of intermediate dosages is also highlighted by their frequency – for example in the case of a Simplex x Simplex marker, one half of all progeny are expected to inherit one '+' allele. One may imagine probabilistic algorithms that locate the presence of markers which confirm particular homologues and then weight the possible presence of the same homologue at a second (intermediate dosage) position by its distance. The method could become progressively refined through a sequence of iterations, where weights are adjusted accordingly. The approach taken in this project merely located the positions of homologue certainty and used this information alone rather than trying to use it to improve the estimates at nearby intermediate dosage positions. In this were done, a network of homologue certainties might be built up which with sufficient density might be enough to proceed with in the QTL analysis.

One of the many shortcomings of my approach is that different results are expected depending on the order of application of the Kosambi-0 and Kosambi-1 functions (as defined in §2.5). The reason that the results are order-dependent is a consequence of the algorithm itself, relating to the domain of each function. This is particularly problematic near recombination sites where discontinuities appear in the string of parental homologue identities (*i.e.* non-smooth boundaries between '0' and '1' probabilities, and *vice versa*). An entry in the X_i probability matrix is only available to be updated by the Kosambi-1 function if it is in [0.5,1). Therefore, if that entry has already been updated by the Kosambi-0 function it will have some value in (0,0.5) and will not be available to the Kosambi-1 function, even if the probability assignment of the Kosambi-1 function would have been more appropriate.

To illustrate this problem with the data-set used in §3.1 and §3.2, the normal order (applying the Kosambi-0 function first) was reversed and instead the Kosambi-1 function was first applied followed by Kosambi-0. As can be seen, the results were different (Appendix II). It appears in this example at least that it is better to apply the Kosambi-1 function last, as was done throughout. One possibility to overcome this issue might be to use inter-marker distances as the basis for applying one or the other function. Rather than progressing from the first marker position to the last, it would make more sense to first organise the markers according to their proximity to other markers and apply the relevant function (Kosambi-1 or Kosambi-0) to marker pairs where the distance is smallest and progress along the list of paired distances. However, it would be a pity not to use the intermediate dosage information as previously advocated, in which case the Kosambi functions could be used to provide probability weights as discussed.

The application of a hidden Markov model to the problem represents a good approach at reconstructing the genotypes at the marker positions, described in Hackett *et al.* (2013). Hidden Markov models were first applied to speech recognition in the early 1970's and have since been applied to a whole range of problems including the identification of CpG islands in sequence reads (Durbin et al., 1998). It would be interesting to compare the results for a number of different approaches to identify what the optimal strategy is.

It was also noted that there was a loss of power when the reconstructed genotype probabilities were used in comparison to the true probabilities. This is hardly surprising given that there is a level of uncertainty introduced in the data when the genotype probabilities are not fully known. Indeed, the degree of this uncertainty was already investigated in the results sections §3.6 and §3.7 for the approach used here. The precise relationship between the level of uncertainty and the power and precision of a QTL analysis was not investigated but could have been. R has a built-in function (`jitter`) which adds a specified amount of noise to data (for example). It would be interesting to repeat the power analyses with different levels of inaccuracy in the genotype probability data to see the robustness of this model to deviations in the homologue identities.

4.4. Setting significance thresholds

In order to derive an experiment-wise threshold for significance it was proposed that trait values be shuffled a certain number of times (*e.g.* N = 1000 times) and each time this is done, the maximum F-statistic be recorded to assess the maximum association between the data and the trait values that would be expected to occur by chance (Churchill and Doerge, 1994). The usual next step is to find (for example) the 95 percentile of these 1000 values and declare F-values that exceed this number to be significant (at a significance level of 0.05). However, there is an inherent limitation to strictly following the prescribed methodology. For example, in a permutation test of 10,000 permutations the resulting maximum test statistics are sorted according to size. To estimate the significance of an association at a position the test statistic is compared to the ordered list from the permutation test. With 10,000 permutations even a very significant test statistic will still only have a p-value of 0.0001. We are therefore limited in our ability discriminate between different significant loci below this threshold.

One possible solution (pursued briefly in this report) is to use the results from the permutation test to construct an empirical distribution for the test statistic. This would then enable comparisons to be made at levels of significance greater than those provided by a list of 1000 or 10,000 test statistics. When a distribution was fitted to the results of a number of permutation tests in the course of this work it appeared that a log-normal distribution provided a good fit to the data and that an F-distribution did not (in fact, *could* not; *c.f.* Appendix IX). According to van Ooijen (1999), it is very difficult to predict the distribution of a collection of maximum test statistics. The procedure advocated by Doerge and Churchill (1994) to estimate an experiment-wise critical test statistic recorded the maximum test statistics over all marker positions for each permutation of the trait values. It is curious to contemplate why these values should follow such a distribution.

Log-normal distributions are an important class of distributions, defined as the distribution of a random variable whose logarithm is normally distributed (Crow and Shimizu, 1988). The distribution appears in many diverse settings, from the growth of organisms or the size of incomes (Crow and Shimizu, 1988) to the number of words written in sentences by G. B. Shaw (Weisstein, 2014). It would have been interesting to see whether a log-normal distribution could be fitted in all cases or whether the fits obtained during this project were circumstantial. Using real data (phenotypic as well as marker data) would be an interesting test – although an analytical derivation would in the end be needed.

An alternative approach is to work exclusively with the test statistics themselves and create a significance profile along the chromosomes from these. In theory there is no problem with this approach – it is merely another way of representing the data. When this point was investigated it was found that there was a strongly linear relationship between the F-values and the $-\log_{10}(p)$ values associated with those F-values (assuming an F-distribution under the Null hypothesis). In other words, it appears that using F-values or $-\log_{10}(p)$ values is equivalent; from my investigations there doesn't seem to be any particular reason why F-values shouldn't be used.

It was found that there was very little difference between the results obtained when the phenotypic values were permuted versus a random sample. This is perhaps unsurprising, as environmental noise was generated by sampling from a normal distribution (and so we would expect the residuals to be normally-distributed). It would be interesting to compare the results using skewed data (perhaps real data) to see whether this is also the case in such situations.

Permutation tests and the choice of N

The recommendation for the number of permutations that would suffice in a typical QTL analysis by the original authors was that the higher the desired degree of significance, the higher the number of permutations needed (Churchill and Doerge, 1994). For example, if a significance threshold of $\alpha = 0.05$ is required, N=1000 permutations may be enough, whereas at $\alpha = 0.01$ up to N=10,000 permutations or more could be needed (Churchill and Doerge, 1994). However, the authors were careful not to specify the numbers required but rather gave these figures as indications only. I was interested to see whether something could be said about the number of permutations needed for QTL mapping in polyploids as this has implications for the power of the QTL study (a very high threshold will result in fewer detections, as was demonstrated in §3.4.1).

The topic of how many times one should permute was indirectly dealt with in a seminal work on a more general type of non-parametric test referred to as Monte Carlo procedures (Hope, 1968). It was shown that the number of permutations (the size of the “reference set”) depends on α and the procedure adopted for testing significance (Hope, 1968). Monte Carlo procedures, of which permutation tests are one example, are particularly useful when little is known about the distribution of the test statistic. If one can make assumptions concerning the distribution of the test-statistic there seems little point in applying non-parametric tests in the first place (Fisher, 1935).

It was interesting to observe the behaviour of the significance threshold when higher numbers of permutations were used (§3.5.1). In the example given, the value of the significance threshold (the 95-percentile) settled to a steady value after approximately 50,000 permutations. Of course, the data was generated in a single run of 500,000 permutations and repeated inspections of subsets of this data generated Figure 11. Therefore the data are not independent – there is no suggestion that from multiple simulations 50,000 is the point where the value of F_{crit} will always converge. However, there *is* a suggestion that after 1000 permutations F_{crit} had not converged. As was noted in §3.5.1, there remains the question of how far one should take convergence if there are only minor fluctuations from the final value after 1000 permutations. Do all permutation runs produce the same steady state value of F_{crit}? We would hope this to be true, otherwise too much stochasticity would be introduced by the use of permutation tests. Does the estimate of the ‘true’ F_{crit} get better with increasing numbers of permutations (assuming it exists after all possible permutations have been exhausted)? This would appear to be assumed, although we would like to check this to be sure.

If the data used to generate Figure 11 is examined in a different way, we have a 'random' sample of 500 possible sets of 1000 permutations of the same data (Figure 18).

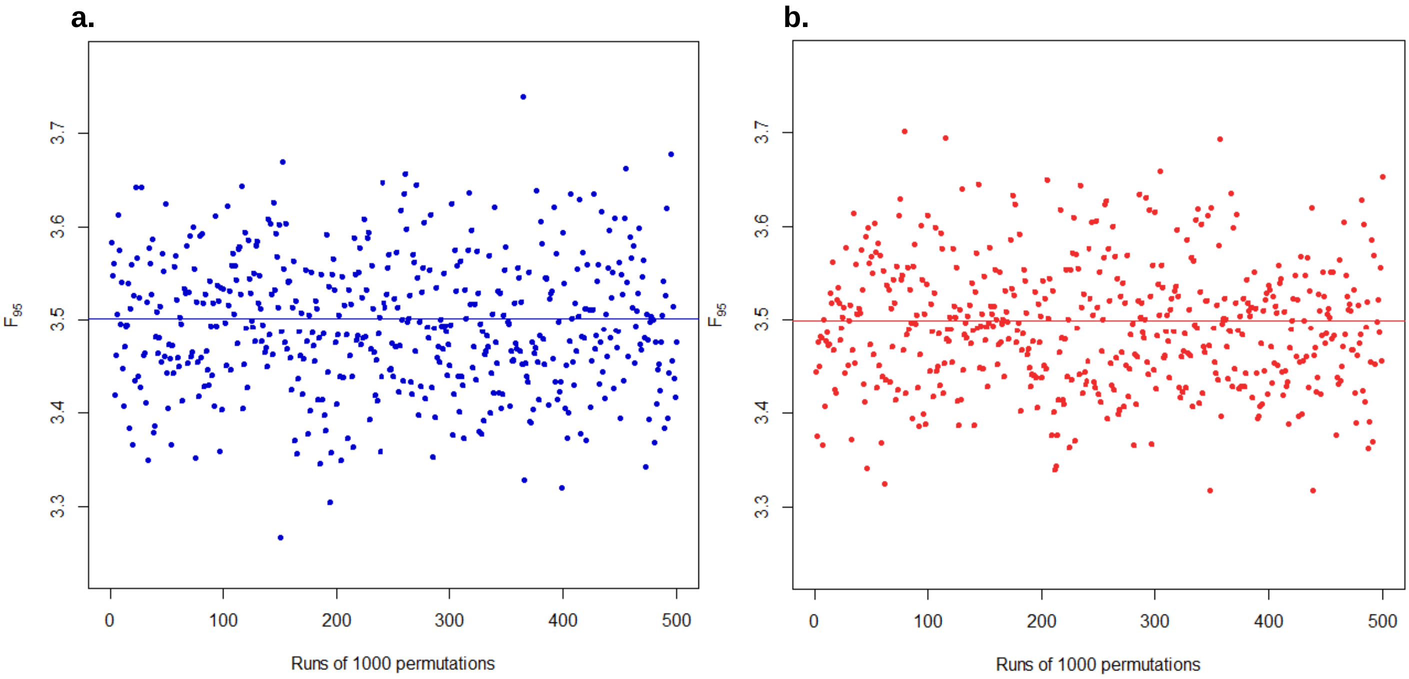


Figure 18. Alternative representation of the data used previously in Figure 10. 95-percentiles from 1000 permutations of phenotypic data when **a.** QTL is present, **b.** no QTL is present. Horizontal lines show the value of the 95 percentile when all data is used together.

The values for the 95 percentile appear to be evenly distributed around the steady value that was identified previously (~3.5) when 50,000 or more permutations were considered. The variance in both cases is approximately 0.005 which is relatively small. Indeed, this variance gives some measure of how good a sample of size N is at approximating the 'true' value of the significance threshold – if this variance is small we can assume that N permutations are sufficient. Of course it may not be feasible to estimate this variance when N is very large because of the time it would take to produce repetitions of this magnitude. It is more useful to consider this variance for smaller N.

One assumption that has apparently been neglected at times in the consideration of Permutation Tests is that the data to be permutated must be exchangeable (Anderson and ter Braak, 2003; Churchill and Doerge, 2008). Exchangeable data arises if the treatments are randomly allocated to experimental units although it must be assumed in observational studies (Anderson and ter Braak, 2003). In the context of QTL mapping we can assume that the experiment set-up included randomisation and thus trait values can be fully exchanged. In more complex experimental designs which include nested treatment structures the consideration of exchangeability is more important.

It has been proposed that in certain situations, more statistical power may be gained through a permutation of residuals rather than a permutation of the raw data itself (Anderson and ter Braak, 2003). Permuting residuals of the full model (ter Braak, 1992) or residuals of the reduced model (Freedman and Lane, 1983) give approximately the same results, although Freedman and Lane's method was shown to be closest to the conceptually exact test (Anderson and Robinson, 2001). For example, in a crossed design with two factors A and B and interaction term AB, a permutation test using the residuals of the reduced model had more statistical power than permuting the raw data when testing for main effects (Anderson and ter Braak, 2003). The main advantage of the exact test is that type I error will be controlled at the chosen significance level, although in different simulated scenarios (different error distributions) permutation of the residuals did not result in an inflation of the type I error (Anderson and ter Braak, 2003).

Ultimately, what we would like to know is whether permutation tests are fit for purpose for the case of QTL analysis in polyploids. One may assume that there is no reason why not, since they have been tried and tested in the diploid case for at least 20 years already. From the investigations performed here, permutation tests appear to be a good method to use (albeit somewhat slow).

4.5. Effect of marker density

We saw that through simulation we were able to predict the density of markers needed to accurately reconstruct the genotype probabilities. It was predicted that even when using only DxN markers (the best for this approach) a marker every 0.6cM is required to reach a relatively low level of error. This is clearly an unrealistic aspiration. Fewer markers still produce good results of course (a DxN marker every 2.5cM also worked well), but the quality of prediction increases with more markers.

The predictions for the 'optimum' marker densities per marker category were tested by simulation, and unsurprisingly they did not produce the perfect results that were tentatively claimed (Figure 19). In other words, there was still some error in the homologue identity probabilities even when a saturated marker set was used. It is clear that the deviation from the true homologue identities tends to zero as the number of markers increased, but only 'reaches' zero in the limit of very large numbers. Transforming the

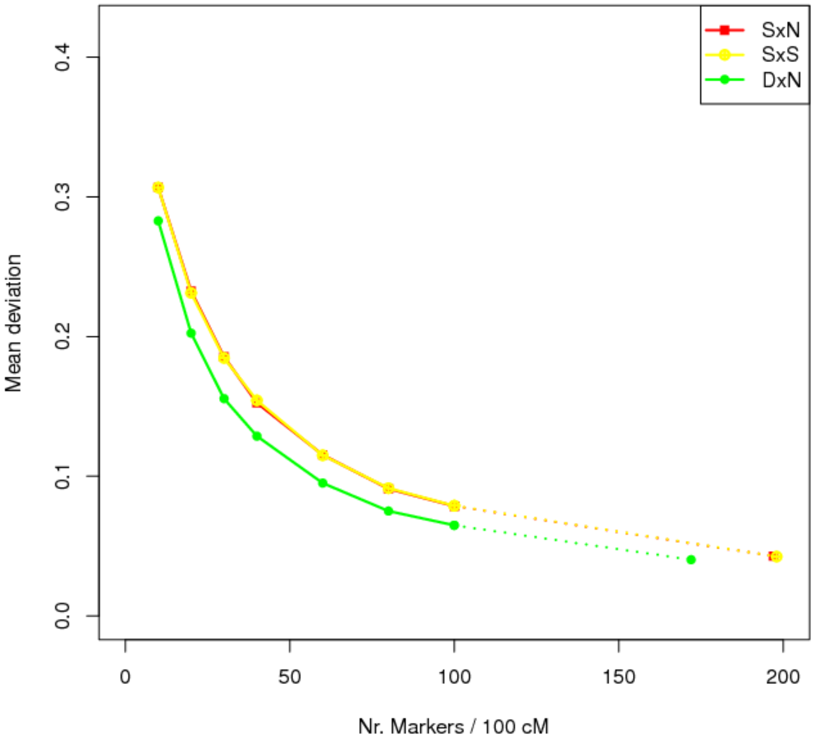


Figure 19. Extension of the results shown previously in Figure 15.a, where the predicted number of markers N_0 to achieve zero deviation were used. Zero was not reached.

data and fitting a linear function was a shortcut used to see approximately what that large number might be in each case – but as an approximate estimate only. It turned out that these estimates fell somewhat short. But it is clear that the method used falls *far* short of what we would hope to achieve. This may at least motivate us to develop better methods for genotype probability reconstruction as previously discussed.

4.6. The effect of multiple QTL alleles

The principal reason for using homologue identities (*e.g.* the GP model) is that it simplifies the issue of polyploid inheritance. In a diploid mapping population with two founding parents (such as a backcross or RIL population *etc.*) there can be at most four QTL alleles (A1 and A2 in parent 1 and A3 and A4 in parent 2) which can produce a maximum of four combinations in the offspring. It is often assumed that there are only two alleles present (A or a) which simplifies the situation further. QTL in a tetraploid may contain up to eight alleles in a bi-parental cross, increasing the number of possible combinations in the offspring to 36. Including all 36 possible QTL genotypes in a model for the offspring results in too many parameters to estimate, leading to a loss of power and poor results. Using genotype probabilities instead, an eight-allele QTL is still admissible but now with only eight variable parameters to estimate (or six if the boundary conditions that each parent contributes two alleles are taken into account; estimation of the mean is common to both approaches).

It is interesting to speculate that different interactions may occur between QTL alleles in ways not anticipated through a simple extension of the diploid concepts of additivity, recessiveness and dominance. In the introduction (§1.7) it was mentioned that previous studies have shown that the polyploid transcriptome can be influenced in non-additive and non-random ways (Leitch and Leitch, 2008). One might therefore ask whether certain combinations of alleles may produce a phenotypic effect that cannot be predicted by dissecting the effects of its component alleles. If such interactions occur, it may help to explain part of the difficulty in breeding at the polyploid level where such favourable allele combinations are less likely to re-assemble in the offspring by chance. How one might screen for such interactions is an interesting issue – perhaps necessitating highly controlled experiments with a set of clonal material that differ at a single locus only. From a breeding perspective the most important aspect will likely remain the identification of sources of desirable genetic variation and combining these in the offspring of a breeding programme.

4.7. Genetic variance and peak LOP

One unresolved question that still perplexes is what the exact relationship between the genetic variance due to a QTL and the significance of the resulting test statistic at a nearby marker is. In particular, I did not find an easy answer to why in the case of a fully dominant QTL, the genetic variance was proportional to the peak test statistic for all marker categories except for SxN markers where the pattern resembled that of the additive case.

Why do we see a qualitative difference between the results for SxN markers and all other marker segregations when the nearby QTL is dominant? What are the four groups that seem to emerge when distance between the marker and QTL or genetic variance in the mapping population cannot explain the differences? What is the factor which defines these groups?

One option was to consider the number of recombination events between the marker and the QTL. Indeed, when recombination was repressed completely (in which case the genetic distance was zero between markers) the effect disappeared – all markers achieved the highest level of significance (*c.f.* Appendix IV). Nevertheless, the difference between SxN and other marker segregations remained in the dominant case. It is also hard to understand how recombination could be the source of this pattern since each individual is assumed to be the product of an independent meiosis with its own recombination signature. In other words, grouping should not occur over a population. These result remain as something to think about – unfortunately, no definite answer was determined.

5. Conclusions

5.1. Summary of key findings

Two simple models were compared for QTL mapping in polyploids by regressing phenotypic data either using SNP dosage information at the marker positions (dosage model) or the homologue parental identities at these positions (GP model). It was found that using raw marker dosage information alone resulted in a very low power of detection (or a high false positive rate if a less stringent threshold was used) as well as an inability to detect QTL without a nearby marker in coupling phase.

In contrast, regression using homologue parental identities proved to be a relatively powerful approach if the heritability is sufficiently high and a suitable mapping population size is used (~100 individuals or more). In order to use this model it is first necessary to reconstruct the parental homologue identities using SNP dosage information. A basic approach to do this was implemented in R but it was found that a dense marker map is needed to accurately reconstruct the homologue identities. Certain marker categories provide more information than others: DxN markers were found to be the most informative for the methods employed, since dosages of ‘0’ and ‘2’ provide full information on four of the eight homologue identities at that position. This coupled with the relatively high proportion of ‘0’ and ‘2’ dosages expected under normal segregation ($^1/_3$) makes them the best marker type for this approach.

Some interesting behaviour was observed when the relationship between the genetic variance due to a QTL was compared with the significance of the test statistic from the most informative flanking marker. For the case of a dominant QTL there was a strongly linear relationship between the two quantities for all marker categories except SxN markers, where four bands appeared that could not be explained by the level of associated genetic variance. These four bands were also found when the simulated QTL had a purely additive effect. It is thought that recombination may have something to do with this segregation in the results, but this was not confirmed.

Significance thresholds for marker testing were implemented using a permutation test in R. This provided the basis for detecting QTL in the power studies, but was also investigated to try to determine how many permutations are needed and whether we can specify the resulting distribution. No conclusive findings were reported; it was found that up to 50,000 permutations (or more) might be needed before a stable estimate of the critical test statistic emerges, but that smaller numbers of permutations give good approximations to this value. When different functions were used in a fitting approach, a log-normal distribution was found to fit the distribution of (maximum) test statistics best, although it is not known whether this is generally the case.

5.2. Recommendations

By way of conclusion, a number of recommendations for further work are offered, many of which have already been alluded to during this report. There is clearly a need for further work in the area of QTL mapping in polyploids. Much of the work presented here was under certain assumptions which might not always be applicable using real data. For example the mode of inheritance may not always be known in polyploids (Stift et al., 2008). It was originally planned to compile information on the segregation behaviour among some common polyploid crops (whether the crop species mostly forms bivalents, trivalents, quadrivalents or a mixture of these *etc.*) but this information was not readily available from literature. In this project it was assumed that bivalents were exclusively formed. In the end it made little difference to the analyses since the method used to reconstruct homologue parental identities did not consider what these possible bivalents might have been. However, it would be interesting to explore the area further so that methodologies could be tailored to the segregation behaviour of individual crops (or mapping populations even) after a diagnostic test using markers (such as that according to Stift *et al.* (2008) for example) has been used.

It was also assumed that the segregation in the offspring was regular whereas segregation patterns of certain markers may in fact be skewed. Any approach to QTL mapping should have ways of dealing with this problem (at the very least, identifying skewed data and excluding those markers) – such issues were not considered in this report but should not be ignored in further studies.

The method of reconstructing the homologue parental identities was far from ideal and could be improved upon greatly. Indeed, this seems to be at the core of any approach to QTL analysis (for polyploids) if SNP dosage data is used. Methods using a branch-and-bound algorithm and hidden Markov models have already been described in the literature. It would be interesting to pursue these and other strategies to compare how each of them perform under various circumstances and conditions.

It would be worthwhile to consider other models as these may offer some advantages over regression approaches. In particular, it would be interesting to see whether maximum likelihood models could be used or whether multiple-marker or composite-interval models might also be applied. Reconstructing the homologue parental identities already uses information from other markers and in a sense extends the information content from markers to their flanking regions. Interval mapping approaches have been shown to be a more powerful strategy for QTL detection in the diploid case; testing this in the case of polyploids would be a worthwhile endeavour.

Finally, although this project relied on simple models, it was still possible to investigate some important questions such as the effect of heritability, QTL size, population size and distance to the nearest marker to the power of QTL detection. Future work in this area will no doubt revisit these important topics to deliver further insights, methods and recommendations for QTL mapping in polyploids.

6. References

Acquaah, G. (2012). Principles of Plant Genetics and Breeding. (Wiley-Blackwell).

Adams, K.L., and Wendel, J.F. (2005a). Polyploidy and genome evolution in plants. Current opinion in plant biology **8,** 135-141.

Adams, K.L., and Wendel, J.F. (2005b). Novel patterns of gene expression in polyploid plants. Trends in Genetics **21,** 539-543.

Anderson, M., and ter Braak, C. (2003). Permutation tests for multi-factorial analysis of variance. Journal of statistical computation and simulation **73,** 85-113.

Anderson, M.J., and Robinson, J. (2001). Permutation tests for linear models. Australian & New Zealand Journal of Statistics **43,** 75-88.

Beavis, W.D. (1998). QTL Analyses: Power, Precision and Accuracy. In Molecular Dissection of Complex Traits, A.H. Paterson, ed (CRC Press), pp. 145 - 162.

Bingham, E.T., and McCoy, T.J. (1988). Cytology and Cytogenetics of Alfalfa. In Alfalfa and Alfalfa Improvement, A.A. Hanson, D.K. Barnes, and R.R. Hill, eds (American Society of Agronomy, Crop Science Society of America, Soil Science Society of America), pp. 737-776.

Bradshaw, J.E., Pande, B., Bryan, G.J., Hackett, C.A., McLean, K., Stewart, H.E., and Waugh, R. (2004). Interval mapping of quantitative trait loci for resistance to late blight [*Phytophthora infestans* (Mont.) de Bary], height and maturity in a tetraploid population of potato (*Solanum tuberosum* subsp. *tuberosum*). Genetics **168,** 983-995.

Brouwer, D., and Osborn, T. (1999). A molecular marker linkage map of tetraploid alfalfa (*Medicago sativa* L.). Theoretical and Applied Genetics **99,** 1194-1200.

Bryan, G.J., McLean, K., Pande, B., Purvis, A., Hackett, C.A., Bradshaw, J.E., and Waugh, R. (2004). Genetical dissection of H3-mediated polygenic PCN resistance in a heterozygous autotetraploid potato population. Molecular Breeding **14,** 105-116.

Champely, S. (2009). pwr: Basic functions for power analysis. R package version 1.1. 1. In The R Foundation, Vienna, Austria.

Chen, Z. (2014). Statistical Methods for QTL Mapping. (CRC Press).

Churchill, G., and Doerge, R. (2008). Naive application of permutation testing leads to inflated type I error rates. Genetics **178,** 609-610.

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

Churchill, G.A., and Doerge, R.W. (1998). Mapping Quantitative Trait Loci in Experimental Populations. In Molecular Dissection of Complex Traits, A.H. Paterson, ed (CRC Press), pp. 31 - 41.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (Lawrence Erlbaum Associates, Inc).

Collins, A., Milbourne, D., Ramsay, L., Meyer, R., Chatot-Balandras, C., Oberhagemann, P., De Jong, W., Gebhardt, C., Bonnel, E., and Waugh, R. (1999). QTL for field resistance to late blight in potato are strongly correlated with maturity and vigour. Molecular breeding **5,** 387-398.

Comai, L. (2005). The advantages and disadvantages of being polyploid. Nature Reviews Genetics **6,** 836-846.

Crow, E.L., and Shimizu, K. (1988). Lognormal distributions: Theory and applications. (M. Dekker New York).

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., and Barakat, A. (2006). Widespread genome duplications throughout the history of flowering plants. Genome research **16,** 738-749.

Da Silva, J.A. (1993). A methodology for genome mapping of autopolyploids and its application to sugarcane (Saccharum spp.). (Cornell University, August).

Damodaran, A. (2007). Probabilistic approaches, scenario analysis, decision trees and simulations (http://people.stern.nyu.edu/adamodar/pdfiles/papers/probabilistic.pdf Accessed 29/06/2014).

Delignette-Muller, M.L., Pouillot, R., Denis, J.-B., and Dutang, C. (2014). fitdistrplus: help to fit of a parametric distribution to non-censored or censored data (R package version 0.1-3, URL http://CRAN.R-project.org/package=fitdistrplus).

deWet, J.M.J. (1980). Origins of Polyploids. In Polyploidy (Springer US), pp. 3-15.

Dewitte, A., Twyford, A., Thomas, D., Kidner, C., and Van Huylenbroeck, J. (2011). The origin of diversity in Begonia: genome dynamism, population processes and phylogenetic patterns. The dynamical processes of biodiversity—Case studies of evolution and spatial distribution. Tech Open Access**,** 27-52.

Doerge, R., and Craig, B.A. (2000). Model selection for quantitative trait locus analysis in polyploids. Proceedings of the National Academy of Sciences **97,** 7951-7956.

Doerge, R.W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. Nature Reviews Genetics **3,** 43-52.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. (Cambridge University Press).

Efron, B., and Tibshirani, R.J. (1994). An introduction to the bootstrap. (CRC press).

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one **6,** e19379.

Fisher, R.A. (1919). The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh **52,** 399-433.

Fisher, R.A. (1935). The design of experiments.

Freedman, D., and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. Journal of Business & Economic Statistics **1,** 292-298.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., and Altshuler, D. (2002). The Structure of Haplotype Blocks in the Human Genome. Science **296,** 2225-2229.

Galitski, T., Saldanha, A.J., Styles, C.A., Lander, E.S., and Fink, G.R. (1999). Ploidy Regulation of Gene Expression. Science **285,** 251-254.

Gallais, A. (2003). Quantitative genetics and breeding methods in autopolyploid plants. (Institut national de la recherche agronomique).

Gar, O., Sargent, D.J., Tsai, C.-J., Pleban, T., Shalev, G., Byrne, D.H., and Zamir, D. (2011). An autotetraploid linkage map of rose (*Rosa hybrida*) validated using the strawberry (*Fragaria vesca*) genome sequence. PLoS One **6,** e20463.

Grivet, L., and Arruda, P. (2002). Sugarcane genomics: depicting the complex genome of an important tropical crop. Current Opinion in Plant Biology **5,** 122-127.

Guo, M., Davis, D., and Birchler, J.A. (1996). Dosage effects on gene expression in a maize ploidy series. Genetics **142,** 1349-1355.

Hackett, C., Bradshaw, J., and McNicol, J. (2001). Interval mapping of quantitative trait loci in autotetraploid species. Genetics **159,** 1819-1832.

Hackett, C.A., McLean, K., and Bryan, G.J. (2013). Linkage Analysis and QTL Mapping Using SNP Dosage Data in a Tetraploid Potato Mapping Population. PLoS ONE **8,** e63939.

Haldane, J.B.S. (1932). The causes of evolution. (Princeton University Press).

Harlan, J., and deWet, J.M.J. (1975). On Ö. Winge and a Prayer: The origins of polyploidy. The Botanical Review **41,** 361-390.

Hieter, P., and Griffiths, T. (1999). Polyploidy - More Is More or Less. Science **285,** 210-211.

Hilu, K. (1993). Polyploidy and the evolution of domesticated plants. American Journal of Botany**,** 1494-1499.

Hope, A.C. (1968). A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society. Series B (Methodological)**,** 582-598.

Jansen, R. (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. Theoretical and Applied Genetics **85,** 252-260.

Jansen, R.C. (1993). Interval mapping of multiple quantitative trait loci. Genetics **135,** 205-211.

Jiang, C., Wright, R., Woo, S., DelMonte, T., and Paterson, A. (2000). QTL analysis of leaf morphology in tetraploid *Gossypium* (cotton). Theoretical and Applied Genetics **100,** 409-418.

Kempthorne, O. (1957). An introduction to genetic statistics.

Koning-Boucoiran, C., Gitonga, V., Yan, Z., Dolstra, O., Van der Linden, C., Van der Schoot, J., Uenk, G., Verlinden, K., Smulders, M., and Krens, F. (2012). The mode of inheritance in tetraploid cut roses. Theoretical and Applied Genetics **125,** 591-607.

Kosambi, D.D. (1943). THE ESTIMATION OF MAP DISTANCES FROM RECOMBINATION VALUES. Annals of Eugenics **12,** 172-175.

Kriegner, A., Cervantes, J.C., Burg, K., Mwanga, R.O., and Zhang, D. (2003). A genetic linkage map of sweetpotato [*Ipomoea batatas* (L.) Lam.] based on AFLP markers. Molecular Breeding **11,** 169-185.

Kuwada, Y. (1911). Meiosis in the pollen mother cells of *Zea Mays* L. Bot Mag **25,** 163-181.

Lander, E.S., and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121,** 185-199.

Lashermes, P., Combes, M.C., Robert, J., Trouslot, P., D'Hont, A., Anthony, F., and Charrier, A. (1999). Molecular characterisation and origin of the *Coffea arabica* L. genome **261,** 259-266.

Leitch, A., and Leitch, I. (2008). Genomic plasticity and the diversity of polyploid plants. Science **320,** 481-483.

Leonards-Schippers, C., Gieffers, W., Schäfer-Pregl, R., Ritter, E., Knapp, S., Salamini, F., and Gebhardt, C. (1994). Quantitative resistance to Phytophthora infestans in potato: a case study for QTL mapping in an allogamous plant species. Genetics **137,** 67-77.

Liu, B., and Knapp, S. (1997). Computational tools for study of complex traits. In Molecular Dissection of Complex Traits, pp. 43.

Luo, Z., Hackett, C., Bradshaw, J., McNicol, J., and Milbourne, D. (2001). Construction of a genetic linkage map in tetraploid species using molecular markers. Genetics **157,** 1369-1385.

Luo, Z.W., Zhang, Z., Leach, L., Zhang, R.M., Bradshaw, J.E., and Kearsey, M.J. (2006). Constructing Genetic Linkage Maps Under a Tetrasomic Model. Genetics **172,** 2635-2645.

Masterson, J. (1994). Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science **264,** 421-424.

Meyer, R., Milbourne, D., Hackett, C., Bradshaw, J., McNichol, J., and Waugh, R. (1998). Linkage analysis in tetraploid potato and association of markers with quantitative resistance to late blight (Phytophthora infestans). Molecular and General Genetics MGG **259,** 150-160.

Milbourne, D., Bradshaw, J.E., and Hackett, C.A. (2008). Molecular Mapping and Breeding in Polyploid Crop Plants. In Principles and Practices of Plant Genomics (Vol. 2: Molecular Breeding), C. Kole and A.G. Abbott, eds (Science Publishers), pp. 355 - 394.

Ming, R., Liu, S.-C., Moore, P.H., Irvine, J.E., and Paterson, A.H. (2001). QTL analysis in a complex autopolyploid: genetic control of sugar content in sugarcane. Genome Research **11,** 2075-2084.

Ming, R., Wang, Y., Draye, X., Moore, P., Irvine, J., and Paterson, A. (2002). Molecular dissection of complex traits in autopolyploids: mapping QTLs affecting sugar yield and related traits in sugarcane. Theoretical and Applied Genetics **105,** 332-345.

Otto, S.P., and Whitton, J. (2000). Polyploid incidence and evolution. Annual review of genetics **34,** 401-437.

Piepho, H.-P. (2001). A quick method for computing approximate thresholds for quantitative trait loci detection. Genetics **157,** 425-432.

Pink, D.A.C. (1993). Leek *Allium ampeloprasum L.* In Genetic Improvement of Vegetable Crops, G. Kalloo, Bergh, B. O., ed (Pergamon Press), pp. 29 - 34.

R_Core_Team. (2014). R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria).

Ramsey, J., and Schemske, D.W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annual Review of Ecology and Systematics**,** 467-501.

Rebai, A., Goffinet, B., and Mangin, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. Genetics **138,** 235-240.

Ripol, M., Churchill, G., Da Silva, J., and Sorrells, M. (1999). Statistical aspects of genetic mapping in autopolyploids. Gene **235,** 31-41.

Santos Leonardo, T. (2013). THE GENETIC LINKAGE MAP AND THE MODE OF INHERITANCE OF TETRAPLOID ROSE. In Laboratory of Plant Breeding (Wageningen University).

Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris. Genetics **8,** 552.

Stebbins, G. (1947). Types of polyploids: their classification and significance. Advances in Genetics **1,** 403-429.

Stift, M., Berenos, C., Kuperus, P., and van Tienderen, P.H. (2008). Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to Rorippa (yellow cress) microsatellite data. Genetics **179,** 2113-2123.

Stupar, R.M., Bhaskar, P.B., Yandell, B.S., Rensink, W.A., Hart, A.L., Ouyang, S., Veilleux, R.E., Busse, J.S., Erhardt, R.J., and Buell, C.R. (2007). Phenotypic and transcriptomic changes associated with potato autopolyploidization. Genetics **176,** 2055-2067.

Swaminathan, M.S., and Howard, H. (1953). Cytology and genetics of the potato (*Solanum tuberosum*) and related species. Bibliographia Genetica **16,** 1–192.

ter Braak, C.J. (1992). Permutation versus bootstrap significance tests in multiple regression and ANOVA. In Bootstrapping and related techniques (Springer), pp. 79-85.

Thomson, N.J., Reid, P.E., and Williams, E.R. (1987). Effects of the okra leaf, nectariless, frego bract and glabrous conditions on yield and quality of cotton lines **36,** 545-553.

Ukoskit, K., and Thompson, P.G. (1997). Autopolyploidy versus allopolyploidy and low-density randomly amplified polymorphic DNA linkage maps of sweetpotato. Journal of the American Society for Horticultural Science **122,** 822-828.

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

Van Ooijen, J.W. (1999). LOD significance thresholds for QTL analysis in experimental populations of diploid species. Heredity **83,** 613-624.

Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. The American Journal of Human Genetics **90,** 7-24.

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

Wang, X., Shi, X., Hao, B., Ge, S., and Luo, J. (2005). Duplication and DNA segmental loss in the rice genome: implications for diploidization. New Phytologist **165,** 937-946.

Weisstein, E.W. (2014). Log Normal Distribution. (URL: http://mathworld.wolfram.com/ LogNormalDistribution.html accessed July 8, 2014: From MathWorld--A Wolfram Web Resource).

Weller, J. (1986). Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics**,** 627-640.

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

Wu, R., Gallo-Meagher, M., Littell, R.C., and Zeng, Z.-B. (2001). A general polyploid model for analyzing gene segregation in outcrossing tetraploid species. Genetics **159,** 869-882.

Zeng, Z.-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proceedings of the National Academy of Sciences **90,** 10972-10976.

7. Appendices

Appendix I – Normalisation after estimating genotype probabilities

At any (marker) location where homologue parental identities are estimated we consider the eight genotype probabilities (X_1, X_2, ..., X_8). Following the steps described in §2.5 normalisation of the genotype probabilities is required to ensure that the condition $\sum_{i=1}^{4} X_i = \sum_{i=5}^{8} X_i = 2$ holds for each set of genotype probabilities X_i at all positions and in all individuals of the mapping population. The exact probabilities '0' and '1' are excluded from the normalisation step as we do not wish to alter their value.

Define S_α as the sum of the non-one entries in (X_1, X_2, X_3, X_4) and S_β as the sum of the non-one entries in (X_5, X_6, X_7, X_8), where non-one refers to probabilities $\neq 1$.

The normalisation factors are given by

$$n_\alpha = {S_\alpha}\Big/{(2 - \Sigma(1))} \qquad \text{and} \qquad n_\beta = {S_\beta}\Big/{(2 - \Sigma(1))}$$

where $\Sigma(1)$ refers to the sum of the '1' probabilities (in the special case of $\Sigma(1) = 2$, normalisation is not required).

The first step in normalisation is to divide all non-one entries in (X_1, X_2, X_3, X_4) by n_α and all non-one entries in (X_5, X_6, X_7, X_8) by n_β. However, in certain cases we may have exceeded probabilities of 1 through this step, as demonstrated by an example in Table I.a.

Table I.a. Example of normalisation highlighting the problem of exceeding a probability of 1.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Non-normalised probabilities	0.0010	0.9472	0.0130	0.8407	0.2085	0.8729	0.8654	0.8066
Sum				1.8019				2.7533
Normalising factors				0.9010				1.3767
Normalised probabilities	0.0011	**1.0513**	0.0144	0.9332	0.1514	0.6340	0.6286	0.5859
Sum				2.0000				2.0000

In this example, a normalised probability of X_2 = 1.0513 results. The solution adopted was to assign a '1' whenever unity was exceeded and then re-apply the normalisation procedure.

Table I.b. Previous example continued, where the probability in X_2 has been set to 1

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Adjusted probabilities	0.0011	**1.0000**	0.0144	0.9332	0.1514	0.6340	0.6286	0.5859
Non-one Sum				0.9487				2
Re-normalising factors				0.9487				1.000
Re-normalised probabilities	0.001	1	0.015	0.984	0.1514	0.6340	0.6286	0.5859
Sum				2.000				2.000

Lemma I. The re-normalisation procedure described does not require subsequent iterations.

Proof. If we consider four homologues alone, so for example X_1 – X_4, there are two conceivable scenarios – (a) that one of the entries exceeds 1 after the first normalisation, or (b) that two of the entries exceed one. In cases where none of the entries exceed 1, re-normalisation is not required.

Suppose scenario (a) holds. Let s_1 be the sum of the other three probabilities, so $0 < s_1 < 1$ (if $s_1 = 0$ it would imply that the probability equals 2, which is impossible). We then apply the step of setting the offending probability to 1 and re-normalising. The algorithm now calculates s_1 and divides it by (2 – 1) = 1, *i.e.* our new normalising factor is simply s_1. Can we exceed unity by dividing any of the remaining three probabilities by this number?

Let us denote these three probabilities by X_i, X_j and X_k where we know that
$X_i + X_j + X_k = s_1 < 1$, and therefore (since they are non-negative numbers) each $X_{i,j,k} < 1$.

Suppose there exists $\alpha \in \{i, j, k\}$ such that $\frac{X_\alpha}{s_1} > 1$. Since $0 < s_1$ we may re-write this as $X_\alpha > s_1$. But s_1 is the sum of three X_i, therefore this clearly cannot hold. We can therefore conclude that there is no $\alpha \in \{i, j, k\}$ such that $\frac{X_\alpha}{s_1} > 1$. The algorithm will successfully terminate after a single iteration.

Suppose that scenario (b) holds. Let the four non-normalised probabilities be denoted by X_a, X_b, X_c, X_d and let s_{before} denote their sum, $s_{before} = X_a + X_b + X_c + X_d$. If X_α and X_β are the two normalised probabilities which exceed unity, then with s_{after} to indicate the sum of the normalised probabilities, we have $s_{after} \geq X_\alpha + X_\beta > 2$. The normalising factor (assuming none of the entries is '1') is $n = \frac{s_{before}}{2}$.

After normalising, $s_{after} = \frac{X_a}{n} + \frac{X_b}{n} + \frac{X_c}{n} + \frac{X_d}{n} = \frac{1}{n}(X_a + X_b + X_c + X_d) = \frac{2(s_{before})}{s_{before}} = 2$

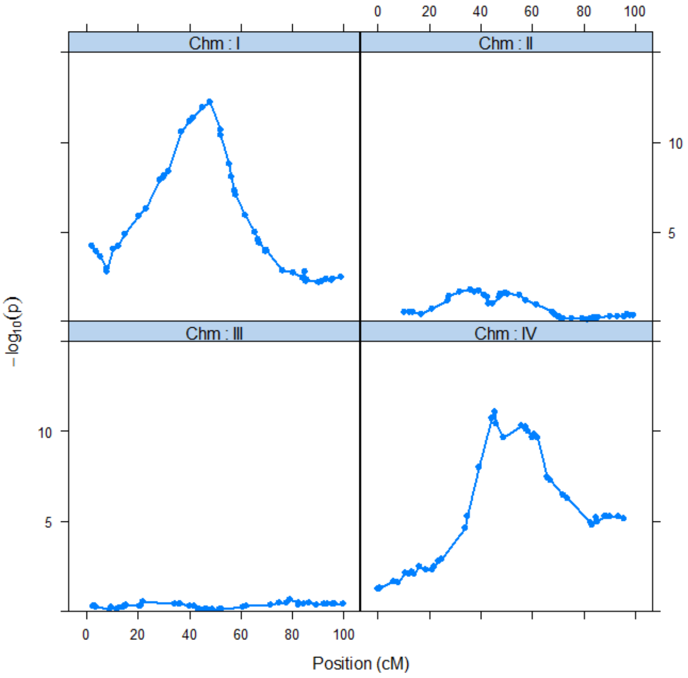
Therefore $2 = s_{after} \geq X_\alpha + X_\beta > 2$ which cannot hold, so (b) cannot be true. (If one of the entries equalled 1, then we are back to the same reasoning that concluded (a)).

Therefore in all cases, this approach terminates after a single re-normalisation iteration, as required.

Appendix II: Effect of order of application of genotype reconstruction functions

The reconstruction of probabilities method mentioned throughout this report first works using the '0' probabilities and then updates the array using the '1' probabilities

a. Applying 'Kosambi-1' function last



b. Applying 'Kosambi-0' function last

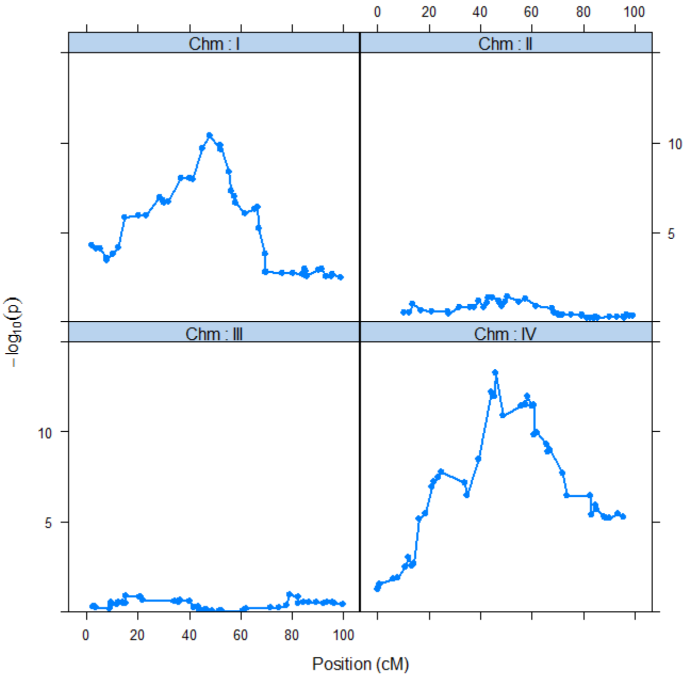


Figure II. Comparison between the results of a regression on genotype probabilities (using the example marker set from §3.1 and §3.2), when **a.** Kosambi-1 function is applied last, **b.** Kosambi-0 function is applied last.

Note: Simulated marker-set had 2 QTL of equal additive effect; 45 markers / chm. using SxN, DxN and SxS markers. h^2 = 0.7

c.

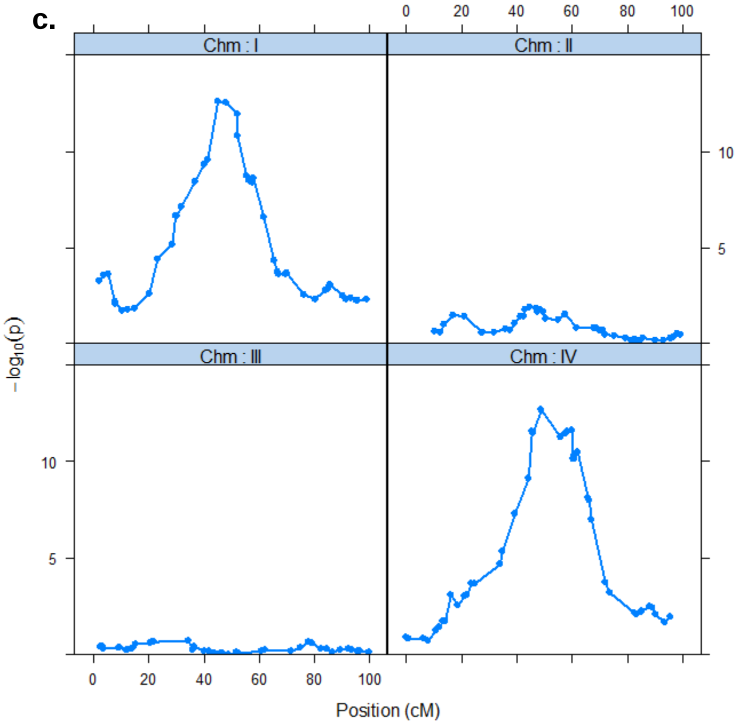


Figure II.c. Original graph using full parental haplotype information

Appendix III: Expected variances for an additive or dominant QTL

Table III.a. Expected mean and variance for different segregation types of an additive QTL.

QTL segregation	Seg. ratio	Effect sizes	Mean effect	σ^2
Q x N	1	1 = a	a	0
S x N	1 : 1	1/2 = 0 1/2 = a	a/2	$a^2/4$
D x N	1 : 4 : 1	1/6 = 0 2/3 = a 1/6 = 2a	a	$a^2/3$
S x S	1 : 2 : 1	1/4 = 0 1/2 = a 1/4 = 2a	a	$a^2/2$
D x S	1 : 5 : 5 : 1	1/12 = 0 5/12 = a 5/12 = 2a 1/12 = 3a	3a/2	$7a^2/12$
D x D	1 : 8 : 18 : 8 : 1	1/36 = 0 2/9 = a 1/2 = 2a 2/9 = 3a 1/36 = 4a	2a	$2a^2/3$

Table III.b. Expected mean and variance for different segregation types of a dominant QTL.

QTL segregation	Seg. ratio	Effect sizes	Mean effect	σ^2
Q x N	1	1 = d	d	0
S x N	1 : 1	1/2 = 0 1/2 = d	d/2	$d^2/4$
S x S	1 : 2 : 1	1/4 = 0 3/4 = d	3d/4	$3d^2/16$
D x N	1 : 4 : 1	1/6 = 0 5/6 = d	5d/6	$5d^2/36$
D x S	1 : 5 : 5 : 1	1/12 = 0 11/12 = d	11d/12	$11d^2/144$
D x D	1 : 8 : 18 : 8 : 1	1/36 = 0 35/36 = d	35d/36	$35d^2/1296$

Appendix IV: Relationship between variance and peak LOP (recombination suppressed)

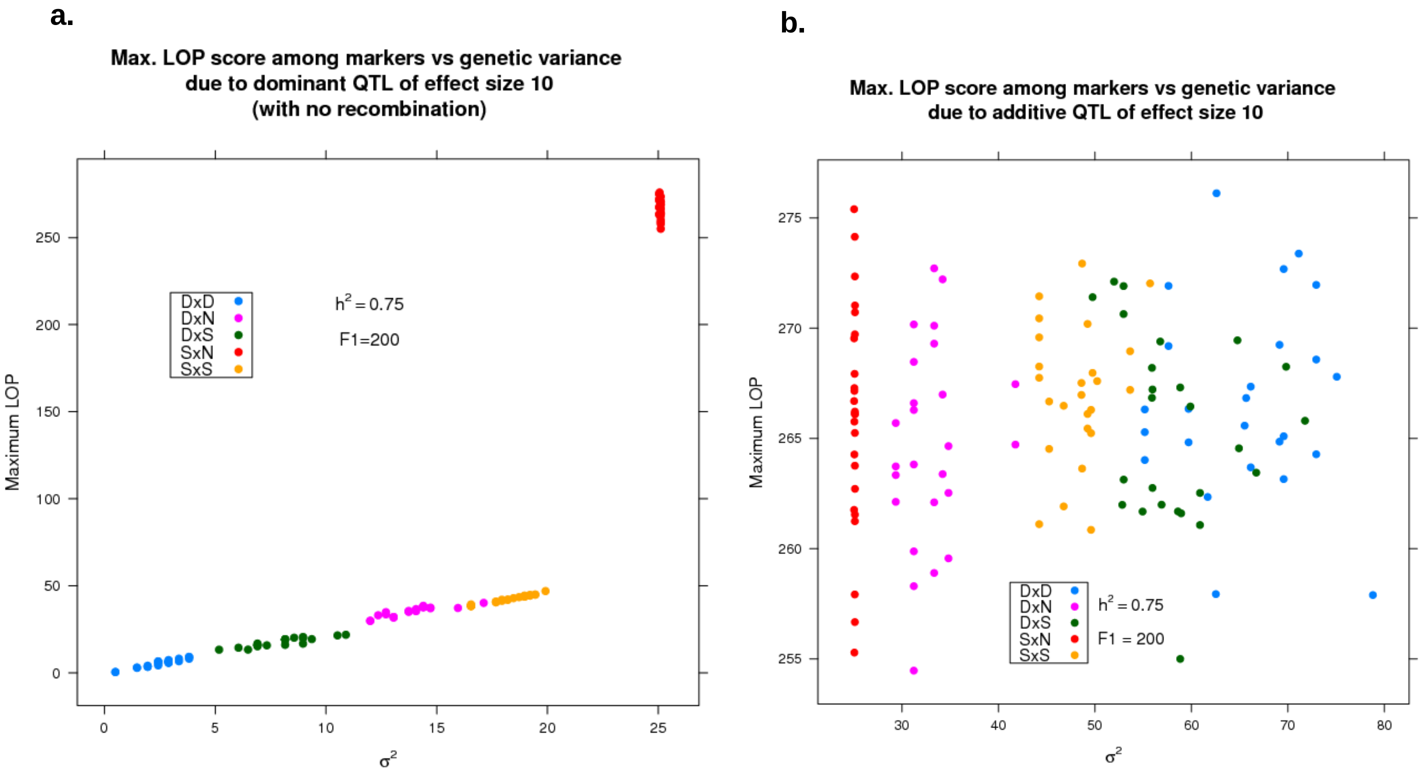
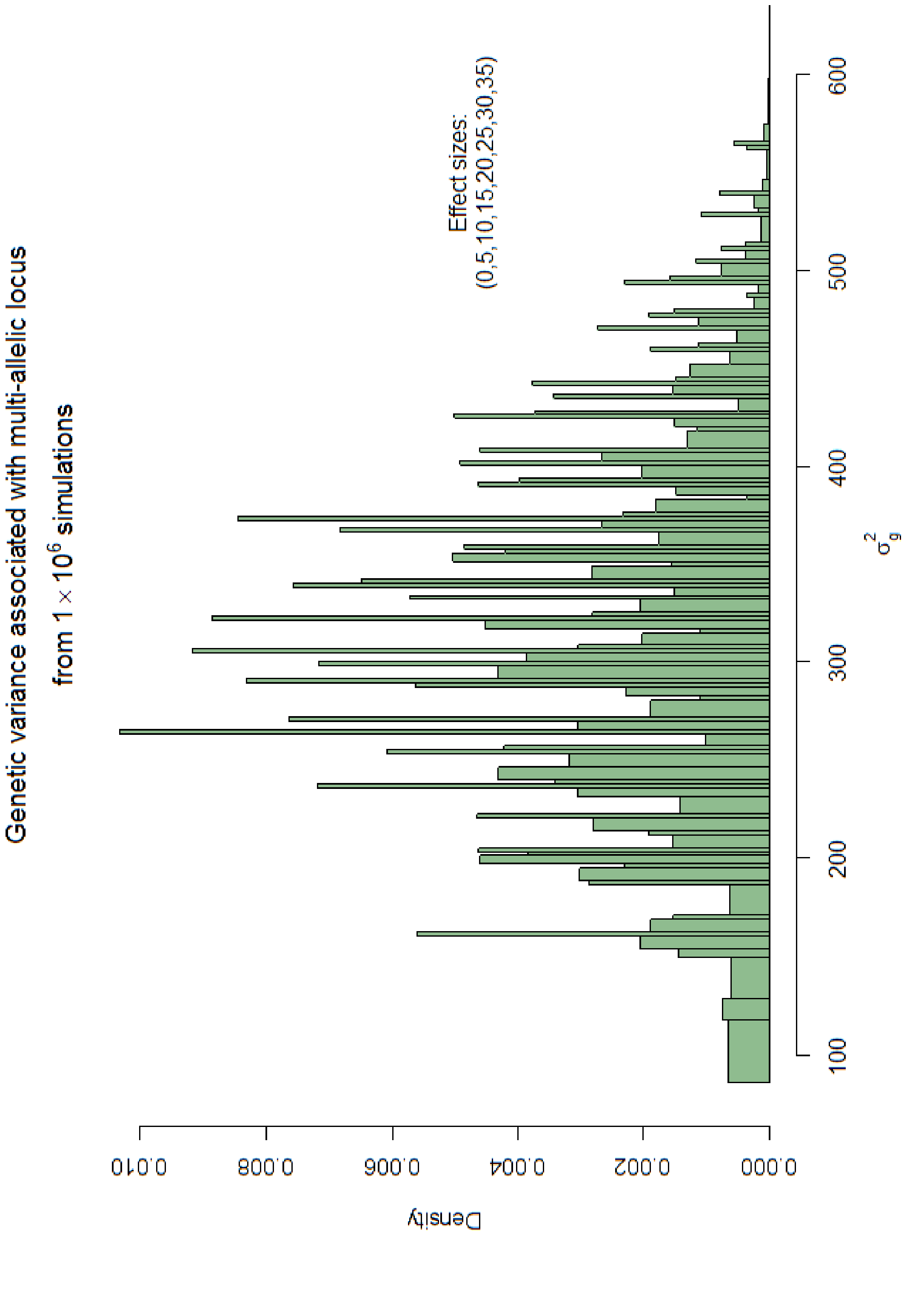


Figure IV. Investigation of the relationship between variance and significance using the GP model for **a.** dominant QTL, **b.** additive QTL (of effect size 10), where recombination has been suppressed.
Note: Results based on a single chromosome with markers evenly-spaced at 0.8cM. F1 = 200 and h^2 = 0.75

Appendix V: Genetic variances associated with an eight-allele QTL



Appendix VI: Power curves for different heritabilities

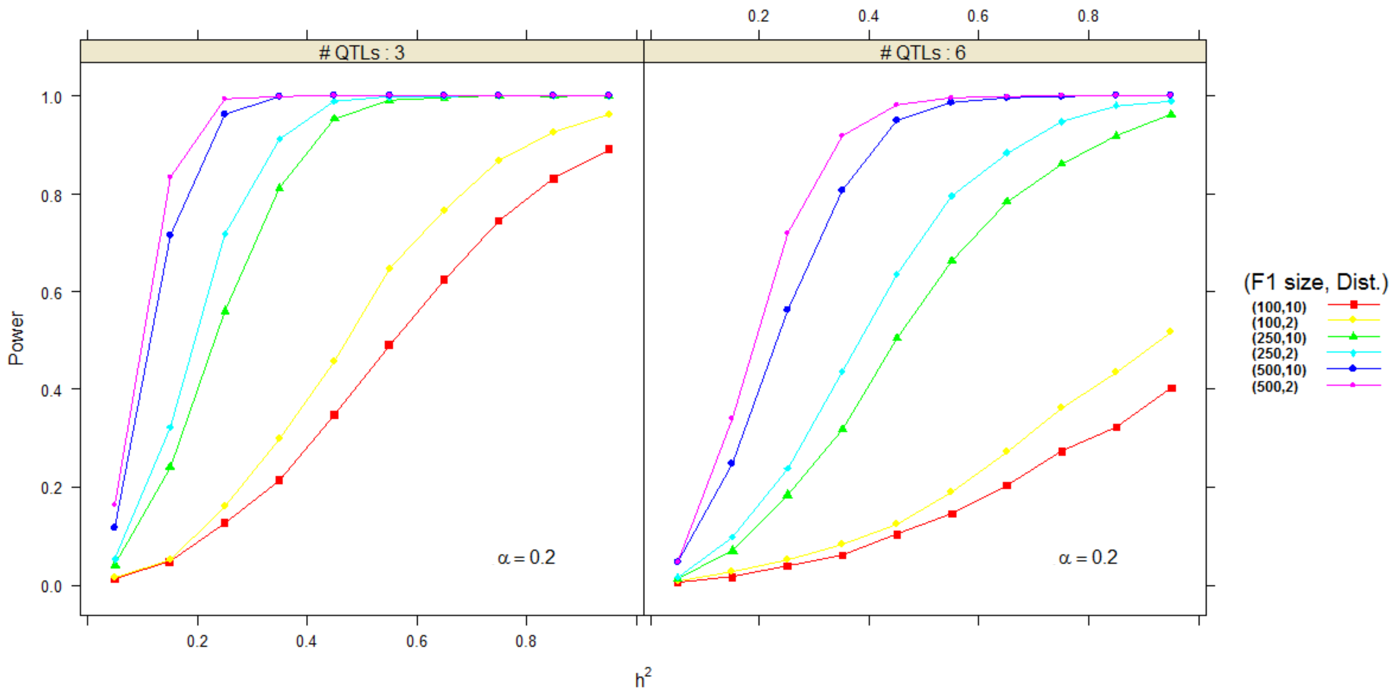


Figure VI.a. Power of detection of additive-effect SxN QTL versus heritability using the GP model, for different numbers of QTL, F1 mapping population sizes and flanking marker distances.

Note: The experiment-wise error rate was controlled at *α* = 0.2 by permutation tests with N=1000 permutations.

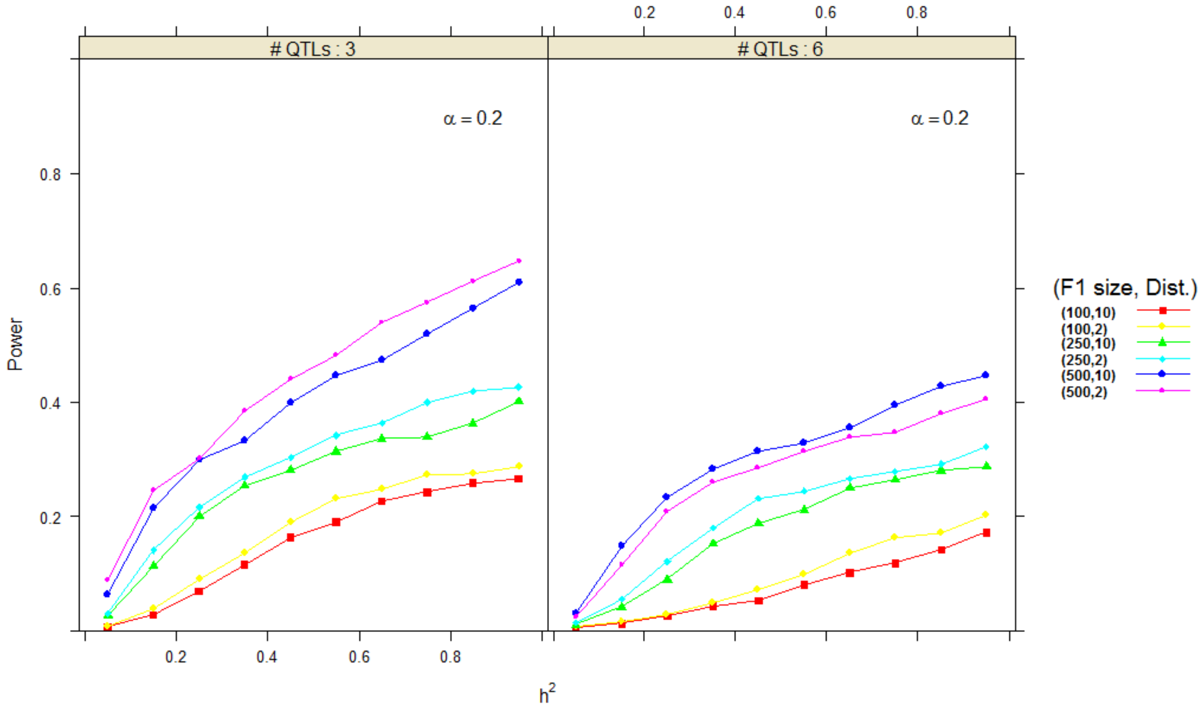


Figure VI.b. Power of detection of additive-effect SxN QTL versus heritability using the dosage model, for different numbers of QTL, F1 mapping population sizes and flanking marker distances.

Note: The experiment-wise error rate was controlled at *α* = 0.2 by permutation tests with N=1000 permutations.

Appendix VII: Power curves for mapping population size

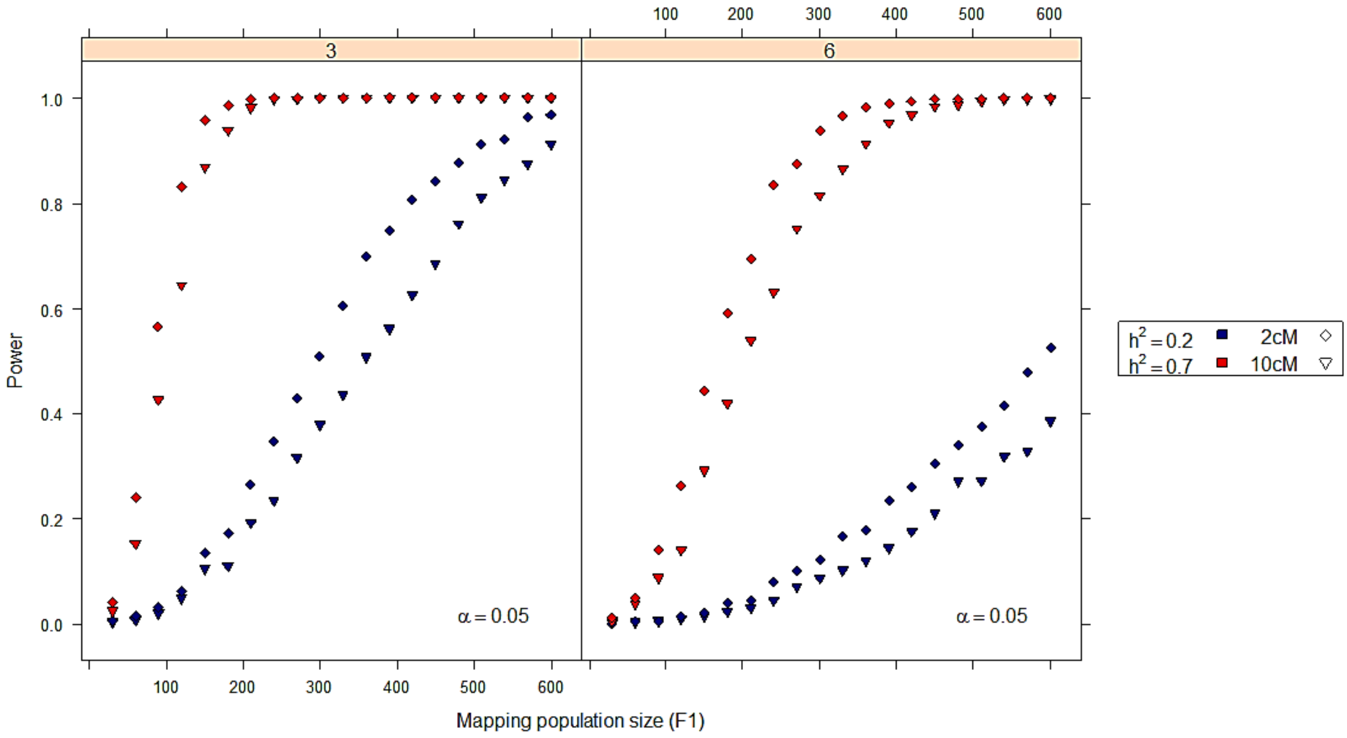


Figure VII.a. Power curves for detection of an additive QTL (α = 0.05) versus size of the mapping population using GP model. Each data-point represent the average proportion of QTL detected over 1000 replicated simulations. *Note:* Distances refer to proximity of closest marker to QTL.

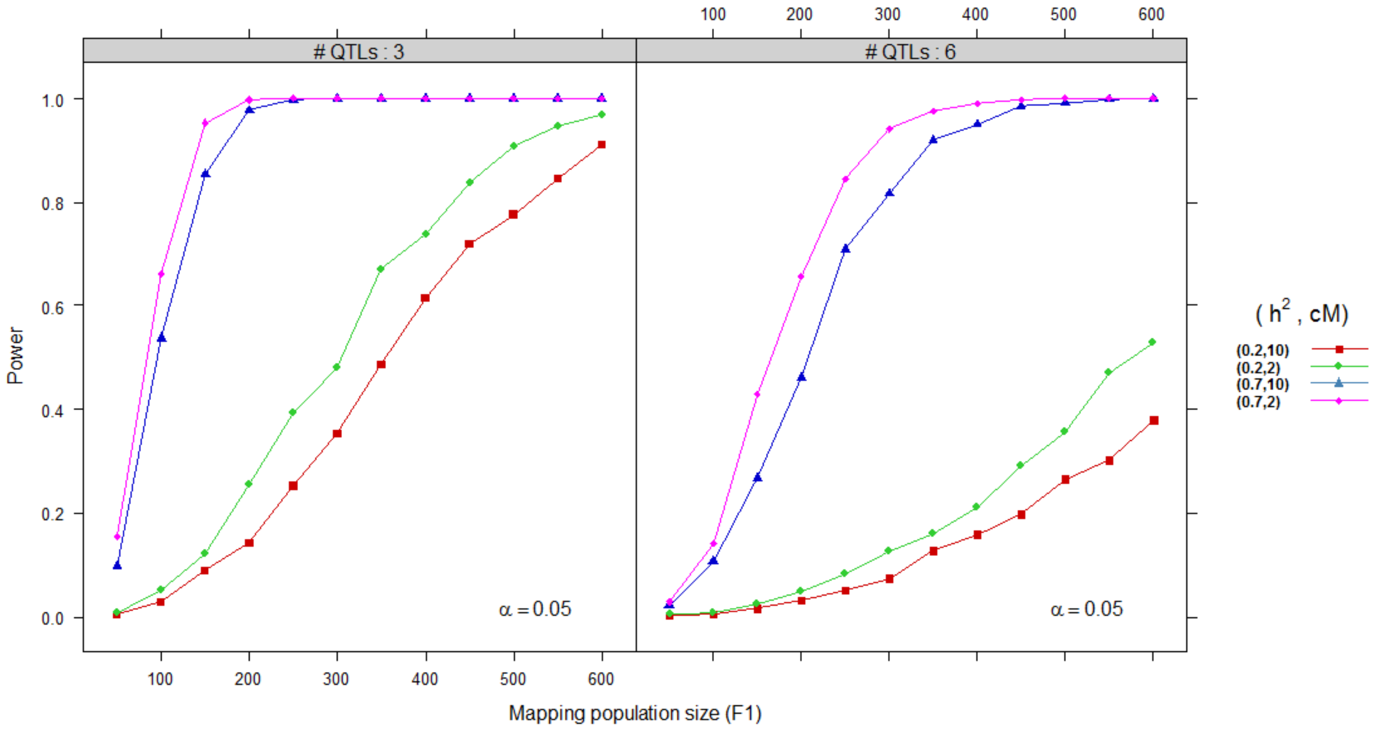


Figure VII.b. Power curves for dominant SxN QTL (α = 0.05) using GP model

Appendix VIII. Permutation Testing – supplementary graphs

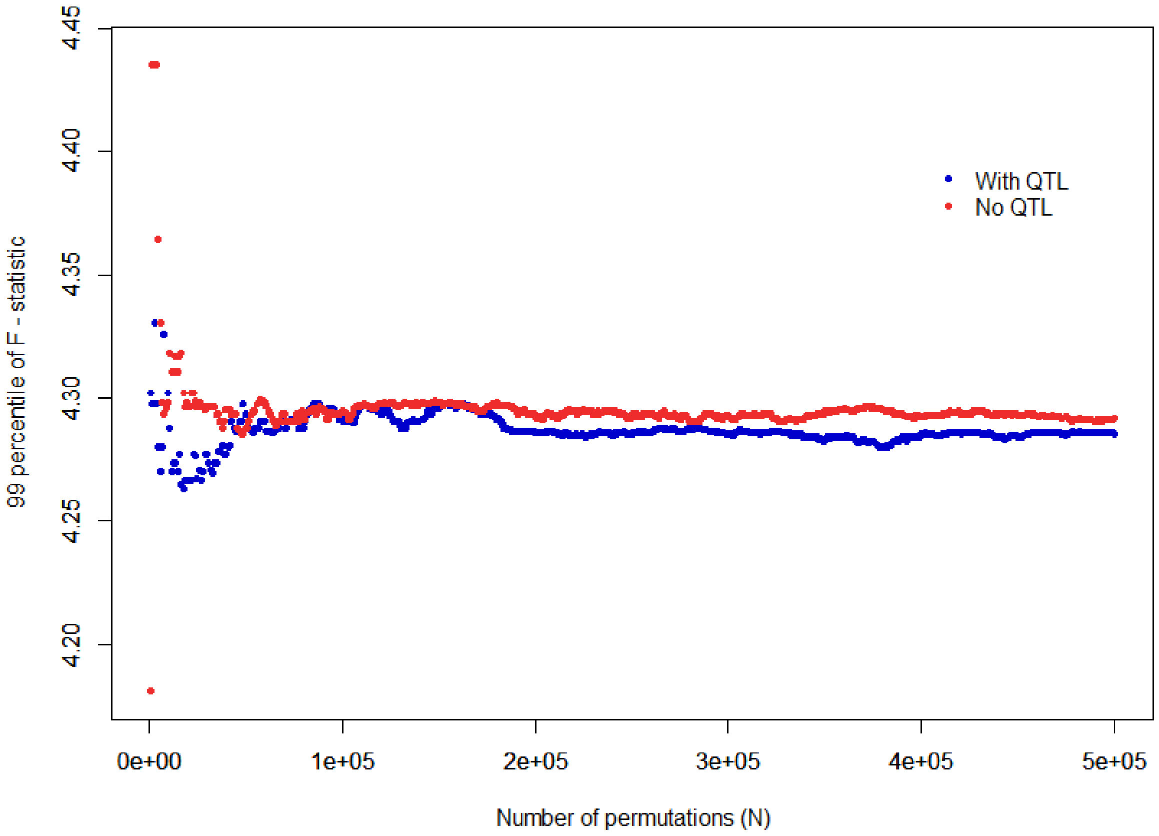


Figure VIII. Behaviour of the 99-percentiles of the F-statistic with increasing number of permutations

Appendix IX. Position of the peak in the pdf of an F-distribution

The probability density function of an F-distribution can be expressed in terms of the gamma function as follows:

$$\mathrm{F}_{k,m}(\mathrm{t}) = \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\,\Gamma\left(\frac{k}{2}\right)} k^{\frac{k}{2}}\, m^{\frac{m}{2}}\, t^{\frac{k}{2}-1}\,(m+kt)^{\frac{-(k+m)}{2}}$$

where the Γ-function is defined, for parameter α > 0 as follows:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$$

For a given k and m degrees of freedom, we can estimate the position of the peak in the $\mathrm{F}_{k,m}(\mathrm{t})$ function by calculating the first derivative $\mathrm{F}_{k,m}{}'(\mathrm{t})$. This yields:

$$\mathrm{F}_{k,m}{}'(\mathrm{t}) = \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{k}{2}\right)}\left(k^{\frac{k}{2}}\, m^{\frac{m}{2}}\right)\left\{t^{\frac{k}{2}-1}\,\frac{-(k+m)}{2}\,(m+kt)^{\frac{-(k+m+2)}{2}}k + (m+kt)^{\frac{-(k+m)}{2}}\left(\frac{k}{2}-1\right)\,t^{\frac{k}{2}-2}\right\}$$

Equating this expression to zero yields:

$$t^{\frac{k}{2}-1}\left(\frac{k+m}{-2}\right)(m+kt)^{\frac{k+m+2}{-2}}.k = -\,(m+kt)^{\frac{k+m}{-2}}\left(\frac{k}{2}-1\right)\,t^{\frac{k}{2}-2}$$

For $t > 0$ we may simplify this to:

$$\left(\frac{k+m}{-2}\right)(m+kt)^{\frac{k+m+2}{-2}}.k = -\,(m+kt)^{\frac{k+m}{-2}}\left(\frac{k}{2}-1\right)\frac{1}{t}$$

Also, since $m + kt > 0$ we may write this as:

$$\left(\frac{k+m}{-2}\right)(m+kt)^{-1}k = \left(1-\frac{k}{2}\right)\frac{1}{t}$$

Solving this expression for t yields the result

$$t_{max} = \frac{km-2m}{km+2k}$$

That is, for a given pair of degrees of freedom k and m, the peak in the probability density function of F should occur at a value given by a relatively simple combination of k and m. To find the maximum of this expression for k, m > 0 (assuming that negative or zero degrees of freedom are not permissible) involves partially maximising the numerator while simultaneously minimising the denominator (so, using partial differentials). However, we merely wish to demonstrate here that this expression does not exceed 1. We can easily do this by assuming the converse to be true:

Suppose

$$t_{max} = \frac{km - 2m}{km + 2k} > 1$$

Therefore, since $km + 2k > 0$ we may multiply both sides by it to obtain

$$km - 2m > km + 2k$$

$$\Rightarrow -2m > 2k$$

$$\Rightarrow -m > k$$

But k, $m > 0$, so $k < -m$ clearly cannot hold. Therefore $t_{max} \ngtr 1$ and therefore we must have $t_{max} \leq 1$ as desired.

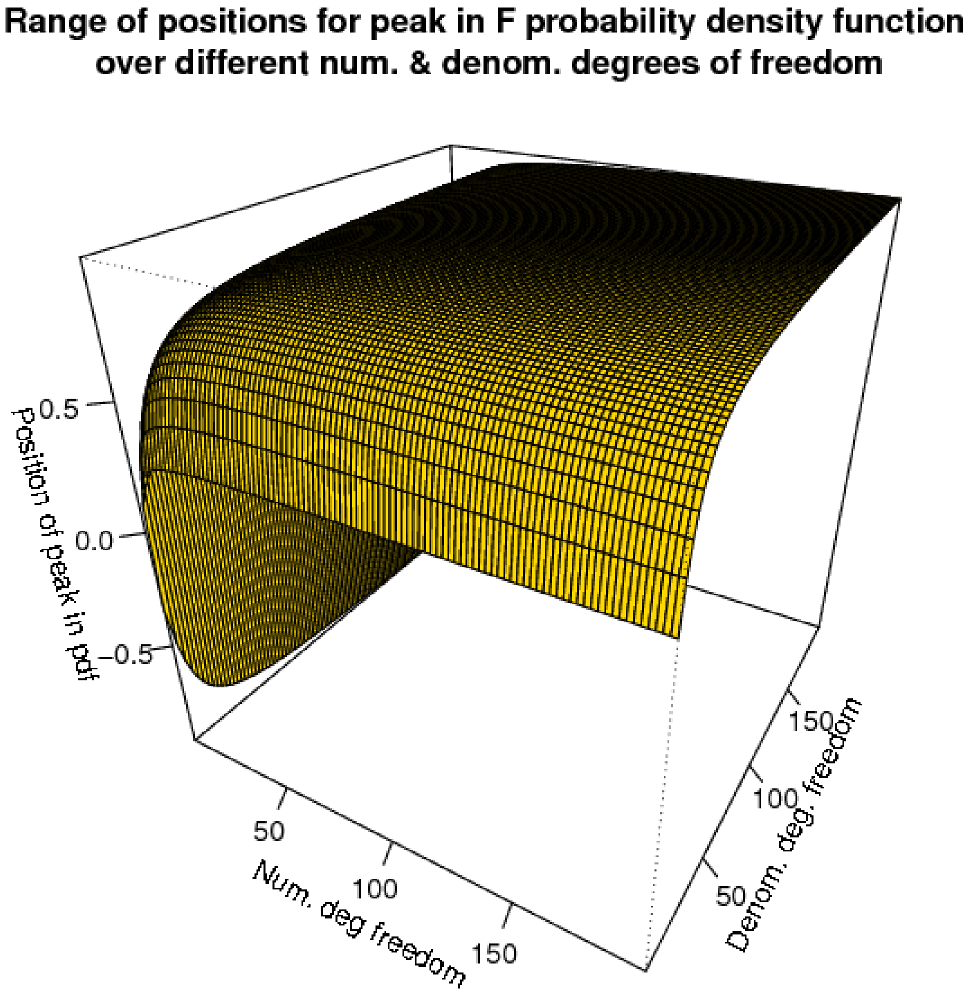


Figure IX. Range of positions for the peak in the probability density function of an F-distribution (z-axis) over different numerator and denominator degrees of freedom. Note that z=1 is not exceeded.

Appendix X: Example R-script used in the simulation process

Basic function to generate phenotypes based on user-inputted effect sizes and heritability:

```
###########################################################################
QTLer <- function (flname) {
 markernames <- rownames(read.genotypes(flname)$geno)
 genotypenames <- read.genotypes(flname)$individuals
 nrQTLs <- as.integer(readline("How many QTLs do you want to simulate?  "))
 QTLmat <- matrix(, nrow = nrQTLs, ncol=length(genotypenames), byrow=TRUE)
 colnames(QTLmat) <- genotypenames

 for (i in 1:nrQTLs){

  QTLlocation <- as.character(readline("Input marker name:   "))
  while(QTLlocation %in% markernames == FALSE) {
   QTLlocation <- as.character(readline("Invalid marker name. Please enter
again:   "))}

  max <- as.numeric(readline("Effect of QTL to simulate?   "))
  dominant <- readline("Is this a dominant-effect QTL (y/n)?   ")
  tempvals <- c(max,0)
  names(tempvals) <- c("1","0")

  if(dominant == "y") {
   tempvalue <- locus.genovalue(read.genotypes(flname), QTLlocation,
tempvals, dominant.allele="1")}
  else {
   tempvalue <- locus.genovalue(read.genotypes(flname), QTLlocation,
tempvals)}

  QTLmat[i,] <- tempvalue
  }

 QTLsum <- colSums(QTLmat)
 genvar <- var(QTLsum)

 herit <- as.numeric(readline("What heritability is required? (0 - 1)"))
 while (herit == 0 | herit == 1) {
  herit <- as.numeric(readline("Error: please choose another heritability
within (0.0 - 1.0):   "))}

 envar <- genvar*(1-herit)/herit
 envnoise <- round(rnorm(length(genotypenames), mean = 0, sd =
sqrt(envar)),digits = 2)
 QTLandNoise <- rbind(QTLsum,envnoise)

 mu <- as.numeric(readline("Select a basic mean for the trait (mu):   "))
 fullvalues <- rbind(QTLandNoise,mu)

 Phenotype <- colSums(fullvalues)
 return(Phenotype)
 } #QTLer()
###########################################################################
```