A statistical framework for the detection of quantitative trait loci in plant multi-parent populations composed of crosses

Vincent Garin

Thesis committee

Promotor
Prof. Dr F.A. van Eeuwijk
Professor of Applied Statistics
Wageningen University & Research

Co-promotor
Dr M. Malosetti
Researcher, Applied Statistics
Wageningen University & Research

Other members
Prof. Dr I.J. Mackay, IMplant consultancy, UK
Dr C.C.A. Maliepaard, Wageningen University & Research
Prof. Dr L. Moreau, French National Institute for Agricultural Research, France
Prof. Dr B.J. Zwaan, Wageningen University & Research

This research was conducted under the auspices of the C.T. de Wit Graduate School
of Production Ecology & Resource Conservation (PE&RC)

A statistical framework for the detection of quantitative trait loci in plant multi-parent populations composed of crosses

Vincent Garin

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 13 September 2019
at 4 p.m. in the Aula.

Vincent Garin
A statistical framework for the detection of quantitative trait loci in plant multi-parent populations composed of crosses,
203 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2019)
With references, with summary in English

ISBN: 978-94-6395-011-4
DOI: https://doi.org/10.18174/494464

Contents

Chapter 1

General Introduction - Mendel's experiments: An epistemological framework for plant genetics

"Few scientists think of agriculture as the chief, or the model science. Many, indeed, do not consider it a science at all. Yet it was the first science - the mother of sciences; it remains the science which make human life possible; and it may well be that, before the century is over the success or failure of Science as a whole will be judged by the success or failure of agriculture. [...] When human beings first learned the cycle of plants and seeds, they were scientists. [...] No scientist performs a greater act of faith in the predictability of the operation of natural laws than the farmer who plows a part of this year's harvest back into the earth. [...] The description and domestication of tens thousands of varieties of the main cereals was started at least six thousand years before Darwin and Mendel, and the slow work of improving yield through genetic selection begun."

André Mayer et Jean Mayer (Agriculture, The Island Empire, 1974)

1.1 Introduction

Mendel's experiments and their conclusions (Mendel, 1866) are often presented as the origin of (plant) genetics as a science. Mendel's experiments constitute a rare complete illustration of the scientific process (Mendel et al., 1993). Indeed, Mendel was one of the few scientists that personally carried out the whole research process. He defined a precise investigation question based on previous researches. He designed an experiment developing a population where crosses were controlled. He collected a large amount of data. He analysed them using some mathematical modelling. And finally, he made a synthesis of his observations by proposing a theory to explain the inheritance of characteristics in hybrids.

According to Popper (1959), the central problem of epistemology (philosophy of science) is the growth of knowledge. The research for method to improve knowledge and to increase trust in the developed conclusions is the centre of the epistemological thinking from Aristotle (Aristotle et al., 2008) to Descartes (Descartes, 1637), from Pascal (Pascal, 1651) to Popper (Popper, 1959). To some extent, Mendel's experiments can be considered as an epistemological framework for plant genetics. Mendel combined many novelties coming from a wide range of scientific domains (biology, mathematics, and physics) to reach his conclusions Mendel et al. (1993). For example, he was one of the first to perform controlled crosses to develop his population, or to use mathematical modelling to analyse his data. The approach he developed can truly be seen as a method allowing to produce knowledge in plant science. A methodology that is still central in actual research, including this thesis, which explains the revolutionary nature of Mendel's works. Therefore, in this introduction, we present the main methodological points developed by Mendel. In parallel, we introduce the questions, the background elements, and the chapters of this thesis showing their relationship to a more general methodology allowing the development of the scientific knowledge in plant genetics.

1.2 Define a research question - the detection of QTLs

Any science starts with an interrogation. Aristotle considered that Humans have a natural desire to know stimulated by astonishment (Aristotle et al., 2008). To paraphrase Popper (1959), all science is a cosmology, a questioning about the world. The fundamental question of Mendel was to determine general laws that govern the transmission of traits in hybrids (Mendel, 1866). At the end of his work, he proposed a theory termed particulate inheritance summarized in what we now call Mendel's laws.

1.2.1 The concept of QTL

Mendel was able to explain the transmission of binary or discrete characteristics, like flower colour, as the results of a two allele (factor) combination, each one being randomly sampled from one of the parents. However, as Darwin (1859) emphasized it, continuity is a constant in the development of natural organisms (*"Natura non facit saltum"*). Therefore, the huge majority of natural variation are continuous or quantitative. Fisher (1918) was the first that proposed a novel theory to conciliate the particulate inheritance principle and the continuity observed in most of the traits. He made the hypothesis that the continuity in phenotypic traits was due to the joint action of several Mendelian factors. Fisher was the first to propose a theoretical notion of what is now called a quantitative trait locus (QTL). In his seminal paper on the correlation between relatives, Fisher (1918) proposed a model where an infinite number of QTLs influence independently the variance of quantitative traits. He also hypothesised that the effect of a QTL could depend on another QTL (epistasis) and that a QTL could interact with the environment. This explanatory model proposed by Fisher can be seen as the starting point of quantitative genetics theory.

1.2.2 The detection of QTLs

The detection of QTLs is the central question of this thesis. According to Doerge (2002, p. 44), "A QTL is a region of any genome that is responsible for variation in the quantitative trait of interest. Originally, the notions of gene and QTL were theoretical concepts. The researches of Sax (1923) can be seen as a first attempt to map a QTL using Mendelian markers. With the discovery of the DNA (Watson et al., 1953) and the development of the sequencing technologies (Sanger et al., 1977), the researchers progressively had a more direct access to the genome where the QTLs ultimately reside. Since then, the detection of QTLs has been developed. Generally, a QTL detection experiment requires the four following steps: a) to develop a population (of plants) that is genetically variable for the studied trait; b) a DNA marker system to genetically characterize the population; c) a phenotypic characterization of the population; and d) statistical methods to evaluate the association between the genotypic and the phenotypic information. These different elements will be further developed in the next sections.

1.2.3 The notion of genetic architecture

The combination of quantitative genetics theory to empirical research allowed to refine the understanding about the underlying mechanisms that determine the phenotypic variation. The concept of genetic architecture is often used to summarize the factors that determine the variation of quantitative traits (Mackay, 1996) and (Wu et al., 2007, p. 14):

- Quantitative traits are due to the simultaneous actions of multiple QTLs.
- QTLs are assumed to act in ways that can be: additive, dominant, epistatic with other QTLs, and/or interacting with environmental factors.
- The size of the individual QTL effects varies from very large to small.
- The QTLs affecting a trait may be distributed along the genome at random or in a certain pattern.

1.3 Experimentation and data collection - the development of controlled populations as a major tool for plant genetics

The development of controlled plant populations is certainly the most critical element of Mendel's experiments. The model plant organism used (peas) and the studied characteristics are the fruit of important thinking and of two years of preliminary experiments (Mendel et al., 1993). Mendel selected peas because it is a natural self-pollinated species that helps to avoid accidental cross-pollination. Mendel also had the good intuition to restrict his research to discrete traits, which facilitated the interpretation of the results. We can also emphasize that Mendel is the inventor of the back-crossing scheme, which helped him to establish the independence of trait segregation (Mendel et al., 1993). Since Mendel's experiments, the use of experimental populations is a central tool in plant science, which allows the scientists to investigate new hypotheses and to express their creativity. The type of plant population used as genetic resource is also a central element of this thesis. As emphasized by Cavanagh et al. (2008), the plant population is the starting point of any QTL detection experiment.

1.3.1 Bi-parental crosses

Historically, QTL detection has been performed in bi-parental crosses between two parents with contrasting phenotypes like the back-cross (BC), the F2 or the recombinant inbred line (RIL) populations (Rakshit et al., 2012). The bi-parental crosses exploit principally recent recombination taking place during the experiment (Cavanagh et al., 2008). Because there are not enough recombination events to shuffle the genome into small fragments, the QTLs are located on wide chromosomal regions (Rakshit et al., 2012). This keeps a large extension of the linkage disequilibrium (LD) and helps to maintain some detection power when marker density is low. The main limitation of the bi-parental crosses is the reduced amount of genetic diversity they address. The allelic genetic space of the bi-parental crosses is limited to the alleles segregating in the two parents.

1.3.2 Association panels

An alternative to the bi-parental crosses is the use of an association panel. Association panels combine a set of unrelated or distantly-related individuals sampled from a hypothetical population in order to increase the addressed genetic diversity (Myles et al., 2009). Association panels exploit the large number of historical recombinations the populations went through (Cavanagh et al., 2008). Therefore, the genome is fragmented into small pieces, which can improve the QTL detection resolution if the marker coverage is dense enough (Kessner & Novembre, 2015; Yu et al., 2008). Association panels also suffer from weaknesses like the lack of a consensus map (Cavanagh et al., 2008) and mostly from unknown population structure and cryptic relatedness (Astle & Balding, 2009). Indeed, genetic drift, selection, populations admixture, sub-population structure, and/or mutation can create LD between loci that are distant from each others (e.g. on different chromosomes) (Flint-Garcia et al., 2003; Patterson et al., 2006). This distant association between different regions of the genome can cause false positive QTL detection.

1.3.3 Multi-parent populations

The multi-parent populations (MPPs) can be seen as a solution to address the limitation of both bi-parental crosses and association panels. There are different types of MPPs. A first category of MPPs is composed of individuals that come from several crosses (at least two) with shared parents (at least one). The nested association mapping (NAM) population (McMullen et al., 2009), the diallel, and the factorial designs (Cockerham, 1963) are example of MPPs composed of crosses. The populations used in breeding programs can also be defined as MPPs (Würschum, 2012). In breeding programs, we typically use incomplete factorial designs where selected parents are crossed with reference or elite lines. There exist more complex MPPs like the multi-parent advanced generation inter-cross (MAGIC) or the Arabidopsis multiparental RIL (AMPRIL) populations (Cavanagh et al., 2008; Huang et al., 2011). In those populations, the DNA of the final lines is a mixture of the founders that were intercrossed for several generations. In the present thesis, we focused on the MPPs composed of crosses.

With respect to the bi-parental crosses, the MPPs address a larger genetic diversity (Blanc et al., 2006). The joint analysis of several crosses allows to increase the degree of polymorphism and to reduce the chances of fixation at the QTL position (Muranty, 1996; Xu, 1998). The use of MPPs also increases the statistical inference space of the QTL effect by testing the QTL in different genetic backgrounds (Xie et al., 1998). Therefore, using MPPs allows to detect QTLs with consistent effects in several genetic backgrounds, which helps to find QTL effects that are more generalizable (Jansen et al., 2003; Verhoeven et al., 2006). This is a major advantage compared to the QTLs detected in bi-parental crosses that often have an effect specific to a particular genetic background (Bernardo, 2016).

Many simulations studies have shown that MPPs allowed to increase the QTL detection power with respect to the bi-parental crosses (Rebaï & Goffinet, 1993, 2000; Muranty, 1996; Xie et al., 1998; Xu, 1998; Jansen et al., 2003; Wu & Jannink, 2004; Verhoeven et al., 2006). Compared to analysing each cross separately, a joint MPP analysis allows to increase the total population size and the sample size to measure the QTL effect when it has a consistent effect over crosses (Rebaï & Goffinet, 1993; Li et al., 2005; Verhoeven et al., 2006). Using MPP reduces the chance of fixation at the QTL position (Xu, 1998). Finally, The use of more than two parents allows to get a sample that is more representative of the QTL variation present in the hypothetical reference population, especially for multi-allelic QTLs (Muranty, 1996; Wu & Jannink, 2004). Real data analyses performed in different species supported the idea that the quality of QTL detection can be improved by the use of MPPs (Schwegler et al., 2013; Li et al., 2005; Blanc et al., 2006; Steinhoff et al., 2011).

With respect to association panels, MPPs offer more control of the false positive rate due to better information about the pedigree and the population structure (Myles et al., 2009). The shuffling of the parental genome happening during the MPP development allows to break the cryptic relatedness existing between non-neighbouring positions (Yu et al., 2008). Therefore, it reduces the chances to detect spurious marker-trait associations.

To some extent, it is possible to draw a continuum between bi-parental crosses on the one side, and association panels on the other (Figure 1.1). In the bi-parental crosses, all individuals come from a single cross or family. In association panels we can imagine that each individual represents a different family. The MPPs are an intermediate situation with several medium-sized or small families.

1.3.4 Genotyping and phenotyping

The genotyping and the phenotyping of the studied population is the way to characterize it and to collect data. The genotypic information is represented by DNA contained in a set of molecular markers spread along the genome. The marker scores of the genotypes composing the studied population are measured during the population genotyping procedure. The single nucleotide polymorphism (SNP) is the privileged type of marker polymorphisms used in QTL detection. The SNPs are characterized by two alleles varying by a single base pair (Wu et al., 2007). At the diploid level, the possible combinations of these two alleles give three possible genotypes: homozygous AA, homozygous BB or

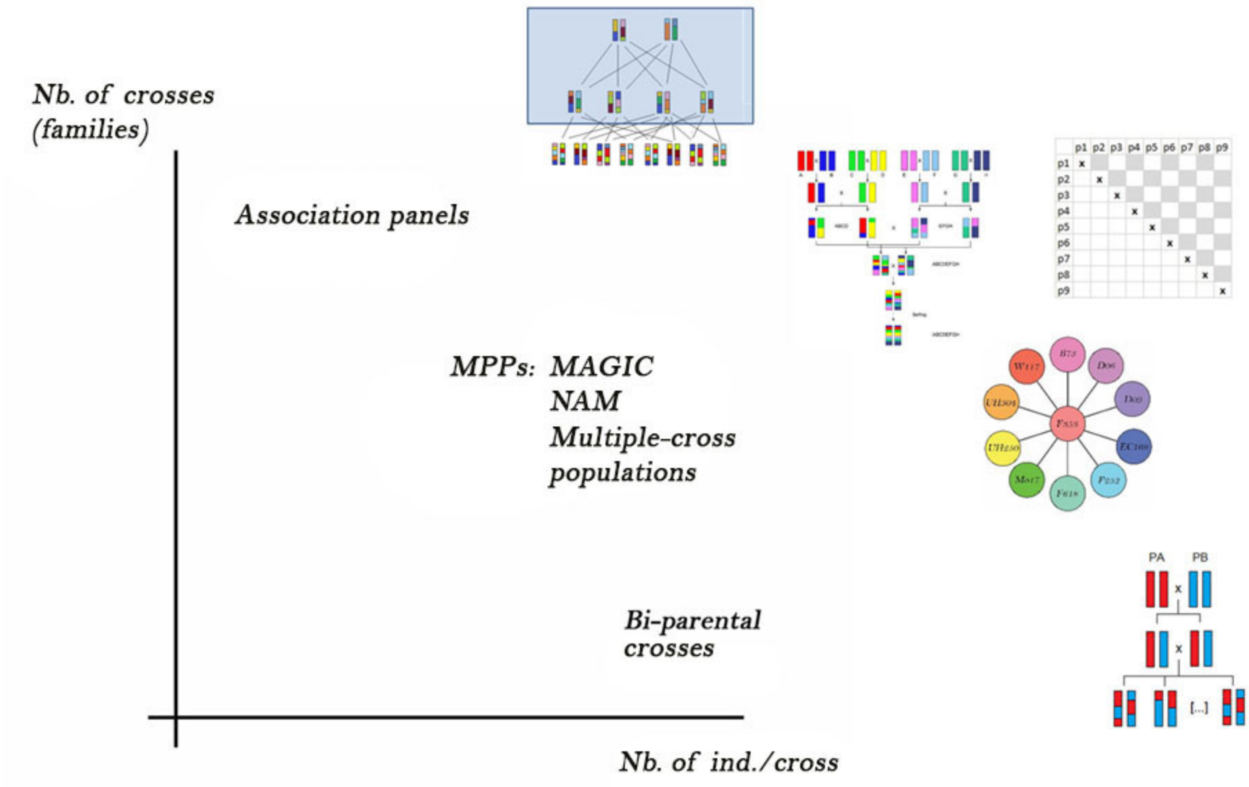


Figure 1.1: Classification of the different artificial plant populations on a two axes: a) Number of crosses (families), and b) Number of individual per cross.

heterozygous AB. The phenotypic characteristics of interest are measured when the plant population is grown during controlled experiments.

At that stage, it is again interesting to draw a parallel with Mendel. Mendel was aware of the probabilistic law of large numbers (Mendel et al., 1993). This could explain why he produced such an enormous amount of data. Indeed, he raised around 24,000 pea plants and examined 11,000 of them critically up to six generations (Orel & Wood, 2000). Once again, we can see that modern concepts like 'big data' were already present in Mendel's works.

1.4 Data analysis - the definition of statistical models for QTL detection in MPPs

Once the plant population has been genetically and phenotypically characterized, it is possible to search for QTLs seeking for statistical association between the genotypic and the phenotypic information. Mendel was one of first scientist who used mathematical models to describe his data and to propose a relationship between the genotype and the phenotype (Mendel et al., 1993). The degree of formalism and the mathematics used by Mendel explain, at least partially, why his works remained ignored by many of his contemporaries. His revolutionary approach was rediscovered later by people like Fisher who extended and brought to a very high level the application of mathematics and statistics to describe and analyse biological data (Fisher, 1958). For example, Fisher (1918) proposed to partition the phenotypic variance into a genetic, an epistatic, and an environmental component. Such a model can be seen as the origin of statistical genetics models used for QTL detection. The core of this thesis is the development of a statistical methodology to detect QTLs in MPPs. The development of QTL detection models is closely related to the biological properties of the population used. In the next sections, we will briefly describe the statistical methodologies that have been developed to detect QTLs in bi-parental crosses, in association panels, and in MPPs. We conclude by presenting an overview of the thesis chapters.

1.4.1 QTL detection models in bi-parental crosses

Basically, QTL detection is the research of statistical association between the DNA polymorphisms contained in markers and the phenotypic variation. The simplest solution is to compare the phenotype means of the genotypes grouped by marker scores using ANOVA assuming that the marker loci coincide with the QTL (Soller et al., 1976).

The chance that the marker and the QTLs have the same position is extremely small. The QTLs are generally situated between markers. Therefore, Lander & Botstein (1989) developed a method called interval mapping testing for QTL association between markers using a probabilistic estimation of the QTL genotype score given flanking markers and the genetic distance to the tested position. Interval mapping models can be formulated in the context of maximum likelihood theory. Under the alternative hypothesis, we assume that the phenotypic data are distributed according to a mixture of normal distributions representing the different genotype classes (Wu et al., 2007). Alternatively, interval mapping models can be estimated using a linear regression of the phenotypic scores on the QTL genotype probabilities conditioned on the flanking markers (Haley & Knott, 1992).

The computation of interval mapping models along the genome is called simple interval mapping (SIM). It allows to obtain a QTL detection profile by plotting the p-values of the tested QTLs against their genetic positions along the genome. The presence of multiple linked QTLs makes the QTL detection more difficult by increasing the size of the residual error. Therefore, authors like Zeng (1993, 1994) and Jansen (1993), proposed composite interval mapping (CIM) models using cofactors representing other

QTL positions present in the genome. The computation of such a model reduces the within marker class phenotypic variance and increases the QTL detection power.

1.4.2 QTL detection in association panels - Linkage vs Linkage disequilibrium analysis

A central difference between controlled crosses like the F2 or the RIL populations and association panels is the possibility to model the pattern of DNA transmission to a reference genotype. In bi-parental crosses and in MAGIC populations, algorithms based on hidden Markov models allow to model the DNA transmission pattern between the founders and the last generation (Zheng et al., 2014; Broman et al., 2003). Such a technique allows to obtain identity by descent (IBD) probabilities of the parental DNA origin. The research of association between the pattern of DNA transmission and the phenotype is generally called linkage analysis (Astle & Balding, 2009).

In association panels however, the population structure and the genetic relationship between lines is mostly unknown (Astle & Balding, 2009). This absence of knowledge about the pedigree prevents from modelling the DNA transmission pattern. Therefore, association panel analysis is still seeking for association between the phenotype and the genotype but without modelling the DNA transmission. QTL detection in association panels generally uses raw SNP markers scores, which represent identical by state (IBS) information. Such a method only relies on the LD that can exist between the markers and the QTL positions. Therefore, QTL detection in association panel is generally called linkage disequilibrium analysis. The detection of QTLs in association panels requires a dense marker coverage to ensure that at least one marker is in LD with each QTL.

The reduced information about the population structure and the genetic relationships in association panels is a major hindrance, which can cause false positive QTL detections (Flint-Garcia et al., 2003; Patterson et al., 2006). Different strategies have been proposed to account for population structure in association panel QTL analysis. The use of a random polygenic effect with a variance covariance (VCOV) structure modelled by a kinship matrix calculated using marker scores is a common strategy (Malosetti et al., 2007). Various strategies have been proposed to calculate such a kinship matrix (VanRaden, 2008; Astle & Balding, 2009). Rincent et al. (2014) and Yang et al. (2014) showed that removing the part representing the tested QTL from the kinship matrix allowed to improve the QTL detection. Therefore, they proposed to exclude the markers of the scanned chromosome from the kinship matrix computation. Other authors proposed to include only markers associated with the trait (Lippert et al., 2013) or to account for the local LD pattern by assigning weights to the markers used (Speed et al., 2012; Rincent et al., 2014).

1.4.3 The QTL detection models in MPPs

The main task of this thesis is to define statistical models to detect QTLs in MPPs composed of crosses. Different approaches have already been proposed to detect QTLs using MPP data. These methods are based on linear models (Jourjon et al., 2005; Rebaï & Goffinet, 2000), linear mixed models (Xu & Atchley, 1995; Crepieux et al., 2004), or models with parameters determined using a Bayesian modelling approach (Yi & Xu, 2001). A fundamental difference between QTL detection in MPPs and QTL detection in bi-parental crosses or association panels is the possibility to test for multi-allelic QTL effects. QTL detection in bi-parental crosses always assumes two parental alleles at the QTL position. In association panels generally, we also assume bi-allelic QTLs corresponding to the SNP marker scores. In MPPs however, it is interesting to consider multi-allelic QTL models with QTL alleles coming from the parents or some ancestral lines above the parents. Therefore in MPPs, the definition of the genetic relatedness between the genotypes and the origin of the QTL alleles are more central questions.

Linear models

The use of a linear model to detect QTLs in MPPs is a first method that relies on a large number of researches. Solutions have been proposed to fit linear QTL detection models using least squares (Rebaï & Goffinet, 1993; Jourjon et al., 2005), iterative reweighted least squares (Rebaï & Goffinet, 2000), and maximum likelihood (Li et al., 2005). An important feature of the proposed models is the assumption of multi-allelic QTL models, of different QTL allele origins, and of different QTL modes of action. Blanc et al. (2006) proposed a disconnected model where the QTL allelic effects are assumed to be different in every cross due to the interaction between a parental allele and the cross genetic background. Another possibility is a connected model where each parent carries a different QTL allele with a consistent effect across crosses (Li et al., 2005; Blanc et al., 2006). The connected model can be seen as a diallel model where parent specific combining abilities are expressed in term of QTLs (Rebaï & Goffinet, 1993, 2000; Verhoeven et al., 2006). When the number of parents is lower than the number of crosses (e.g. in a half diallel), the parental model allows to reduce the number of parameters to estimate.

In the same vein, Jansen et al. (2003) and Leroux et al. (2014) proposed to further reduce the number of QTL alleles by assuming that parents who are genetically similar receive their alleles from a common ancestor. They proposed strategies to cluster parents along the genome in ancestral groups using local haplotype similarities. This method has been described as a linkage/linkage disequilibrium technique (Meuwissen et al., 2002) because it uses both the IBD information to link the offspring to the cross parents and the IBS marker scores to cluster the parents. The maximal reduction in the number of QTL alleles can be achieved if we assume a bi-allelic model where the allele classes correspond to the SNP marker scores, like in association panel analysis (Liu et al., 2012; Würschum et al., 2012).

From a theoretical point of view, the reduction of the number of assumed QTL alleles allows to increase the power of the statistical test and the sample size to measure the QTL allelic effects (Rebaï & Goffinet, 1993; Li et al., 2005; Jansen et al., 2003). The analysis of real data showed however, that more parsimonious models are not necessarily the best option (Steinhoff et al., 2011; Bardol et al., 2013). The reduction of the number of assumed QTL alleles could lead to too simplistic models that fail to model complex QTL allelic series (Giraud et al., 2014).

Mixed model with random QTL effects

A second class of methods for QTL detection in MPPs is based on the linear mixed models. An interesting property of those models is the possibility to fit the QTL terms as random and to estimate a single variance component associated with the QTL effect. When the population design is clearly identified, and the number of crosses and/or parents is not too large, it is possible to make assumptions about the number of alleles, their origin and to properly estimate the genetic effect of the founder lines (Würschum, 2012). In many MPPs however, the combination of material makes the use of fixed QTL effects difficult because the number of parameters to estimate increases with the number of crosses or parents. The information to estimate these parameters can also be reduced when the crosses are too small (Xie et al., 1998; Xu, 1998). This situation can be encountered in outbred populations (Wu et al., 2007) or in MPPs coming from breeding programs (Crepieux et al., 2005). In such a situation, the difficulty to specify a discrete set of alleles makes the use of a single random term to model the QTL effect attractive.

The use of random QTL effects has been proposed already long time ago to overcome the difficulty to define a finite set of alleles in populations with a complex pedigree (Xu & Atchley, 1995). Haseman & Elston (1972) showed that individuals with similar phenotypes are more likely to share alleles IBD. This led researchers to propose random QTL term models where the VCOV structure associated to the QTL term represents the IBD relationship existing between the analysed genotypes (Xu, 1996). With respect to a fixed QTL effect, the interpretation of a random QTL effect can be interesting. Indeed, we can see the set of individual crosses, parental or ancestral QTL effects as a sample from a larger population of all potential QTL effects. Therefore, we are not any more interested in estimating individual QTL effects but the variability that can be inferred at the whole population level (Xu, 1998).

Different forms of VCOV can be associated to the QTL term which offers some flexibility to model the QTL effects by taking into consideration different sources of genetic relatedness (Meuwissen et al., 2002; van Eeuwijk et al., 2010b). For example, Xie et al. (1998) proposed a VCOV that only integrated the relatedness due to full-sibs relationships. Crepieux et al. (2004) extended the QTL VCOV formulation integrating the half-sibs relatedness and the relatedness due to shared ancestors above the parental lines. This idea to have different models reflecting different sources of genetic relatedness is similar to the approach developed in the linear model framework.

The Bayesian approach

The last category of methods to detect QTLs in MPPs uses the Bayesian approach to fit the models. The complexity of the QTL detection models can rapidly grow including parameters like: the number of QTLs, their positions, the type of effects (additive, dominance), the genotypes at the QTL positions, and environmental factors. In such a situation, the specification of a closed mathematical expression for the underlying stochastic process becomes impossible and the Bayesian approach becomes a solution (Yi & Xu, 2001). The Bayesian approach takes simultaneously into consideration all sources of uncertainty, which can be extremely helpful when the number of parameters and the randomness associated to them increase (Bink et al., 2008).

The use of the Bayesian approach to detect QTLs is particularly interesting in populations with a complex pedigree (Yi & Xu, 2001; Bink et al., 2002). As in the linear model and mixed model approaches, researchers have proposed solutions to infer ancestral relatedness at the QTL position using marker and/or pedigree information. For example, Ter Braak et al. (2010) developed an algorithm to determine, in a probabilistic way, latent ancestral classes. Such an information was successfully integrated in Bayesian QTL detection models (van Eeuwijk et al., 2010b; Bink et al., 2012). The flexibility offered by the Bayesian approach allowed Jannink & Wu (2003) to treat the number of alleles at the QTL position as a random parameter. This method is particularly elegant. Therefore, the Bayesian approach seems to offer a large flexibility to model complex biological processes. However, it can be computationally intensive (van Eeuwijk et al., 2010b).

As we could see in the three different approaches (linear models, mixed models and the Bayesian approach), the development of methods to define correctly the genetic relatedness between lines allows to improve the detection of QTLs in MPPs. From a general point of view, the definition of the genetic relatedness between lines is a central task in QTL detection. Therefore, QTL detection should be seen as the research of statistical association between a measurement of genetic relatedness and the phenotypic variation.

1.4.4 Overview of the thesis

The objective of this thesis is the development of a statistical methodology for QTL detection in MPPs. In chapter two, we presented linear models already proposed in the literature in a consistent framework. The main feature of these models was the form of the QTL effect following different assumptions about the number of QTL alleles, their origins, and their modes of action. We proposed a collection of four QTL models that could be classified from the least to the most parsimonious. First, the cross-specific model assumed that the QTL allelic effects were different in each cross. Second, the parental model assumed that each parent contributed to the population by a different QTL allele with a consistent effect across crosses. Third, the ancestral model assumed that the parents with genetic similarities beyond a threshold value received their alleles from a common ancestor. The last model was a bi-allelic model with QTL allele classes corresponding to the SNP marker scores.

We extended the proposed QTL detection framework by developing a procedure determining a multi-QTL effect model where the effects of the QTL positions could be characterized by different assumptions. Indeed, so far, the proposed QTL detection methods restrict the model to a single type of QTL effect, keeping the same type of incidence matrix for all tested loci. Since the allelic effect present in MPPs may vary across loci, we proposed a procedure to build a model where different loci could be modelled by different types of QTL effects.

We also extended the proposed QTL detection framework by allowing the variance of the model error term to be different in each cross. Such an error term allowed to better reflect the potential heterogeneity of error that is present between the crosses than a single error term that is generally used. We also proposed a model where the residual polygenic variance was represented by a random term with a VCOV modelled by a genotype relationship matrix based on pedigree records. We moved to the mixed model framework to implement these models allowing to handle more complex dependency structures between genotypes due to the crossing scheme and/or shared pedigree.

To facilitate the use of our QTL models, we wrote an R package called mppR available on GitHub and CRAN. This package assists the user in all the QTL detection steps from raw data processing to results visualisation. We defined a unique and portable class of object to gather all sources of data (genotype marker matrix, phenotype records, crossing scheme, etc.). We programmed a sequence of functions to perform several operations of data processing like the data quality control, the IBD probabilities computation, and the clustering of the parental lines. The central task of the package was the QTL detection using the defined models. We implemented a procedure composed of a SIM followed by a CIM to determine a multi-QTL model. We wrote a cross-validation function to evaluate how the detected QTLs and their effects would generalize to a pseudo-independent dataset. Finally, the package provided a visualisation of the results and an estimation of the QTL effects.

In the third chapter, we tested the QTL detection framework proposed in chapter two using NAM data. We proposed some hypotheses on how our models should react in populations characterised by different levels of genetic relatedness. More specifically, we tested if populations with a narrower genetic basis would be better described by QTL terms with a reduced number of (ancestral) QTL alleles. On the other hand, we hypothesized that MPPs covering a larger genetic diversity would be better described by QTL terms with a larger number of alleles reflecting this diversity. We also tested the usefulness of the cross-specific variances of the error term in NAMs with various level of genetic distance between the central and the peripheral parents. The data we used did not allow us to support our hypotheses nor the usefulness of models with cross-specific error variances. However, we could show that in some cases, the use of the multi-QTL effect model gave a better description of the phenotypic data.

In the fourth chapter, we evaluated our QTL detection framework by simulation. We simulated phenotypic values controlled by different genotypic models. The genotypic models we used corresponded to the QTL models described in chapter two (cross-specific, parental, ancestral, and bi-allelic). We determined what should be an optimal strategy to design MPPs evaluating the QTL detection power in different MPP designs. These designs were characterised by a different form (diallel, chessboard, factorial, and NAM), and by a different number of parents used (five or nine). In general, we concluded that using larger individual cross sizes allowed to detect more QTLs than using a larger number of parents to cover a wider genetic diversity. However, the sampling of a larger number of parents could be useful to detect rare QTLs with a large effect.

In the fifth chapter, we extended our framework to perform QTL detection in MPP data characterized in multiple environments. The method generally used to analyse MPPs characterized in multiple environments is to perform a QTL detection on genotype BLUEs representing a main effect across environments. Such a strategy does not allow to estimate the QTL by environment (QTLxE) effects. Therefore, we proposed methods to estimate the QTLxE effect either by analysing separately the data measured in each environment or by a joint analysis taking into consideration the covariance due to the same genotype measured in different environments. We completed our methodology by a one-stage analysis on plot data integrating also the effects of experimental design factors. We illustrated our MPP GxE methodology using two different NAM populations. We showed that compared to separate within environment analyses, a joint analysis allowed to get a more complete picture of the trait genetic architecture. We also illustrated the possibility to extend our models to characterize the effect of environmental covariates like temperature, water precipitation, etc. on the QTL effects. Such an eco-physiological model can be helpful to investigate the biological mechanisms behind the QTLxE effects.

In the sixth chapter, we discussed the main results of this thesis and confronted them to the existing scientific corpus. We emphasized how our methods improve the knowledge on QTL detection in MPPs, its realisation, and the use of the results for plant breeding. We finish the thesis by discussing the implication of this thesis on the use of MPPs for marker assisted selection.

To conclude this overview, we would like to emphasize the capacity of our methodology to model the phenotypic variation due to a wide range of elements. We first developed models to characterize the genetic effect and determine if the QTL effects are consistent across different genetic backgrounds. Then we extended our methodology to estimate the consistency of the QTL effects across environments. We also proposed one-stage analyses to model simultaneously non-genetic variation due to experimental design elements and the QTL effects. The methodology developed in this thesis should therefore allow to detect QTL effects with consistency in different genetic background and across environments that should be useful for marker assisted selection.

1.5 Discussion - Contribute to the theoretical development

The last part of a scientific work is the discussion that allows to have a global consideration on the results of the data analysis. The discussion is also the possibility to confront any elements of the scientific process used to the existing scientific corpus on the topic. Such a comparison allows: to answer pending questions, to propose solutions for unsolved problems, and/or raise new questions/issues that should be addressed in further researches. Ultimately, the discussion is the place where new theoretical development should be proposed. These developments will generally take the form of hypotheses trying to explain the studied phenomenon. The theoretical hypotheses developed in the discussion will stimulate new questions that will be the object of new scientific research. This lead us back to the first step of the research process. Science can therefore be seen as a continuous loop where the outcomes of a research stimulate the questioning of new scientific hypotheses (Figure 1.2).

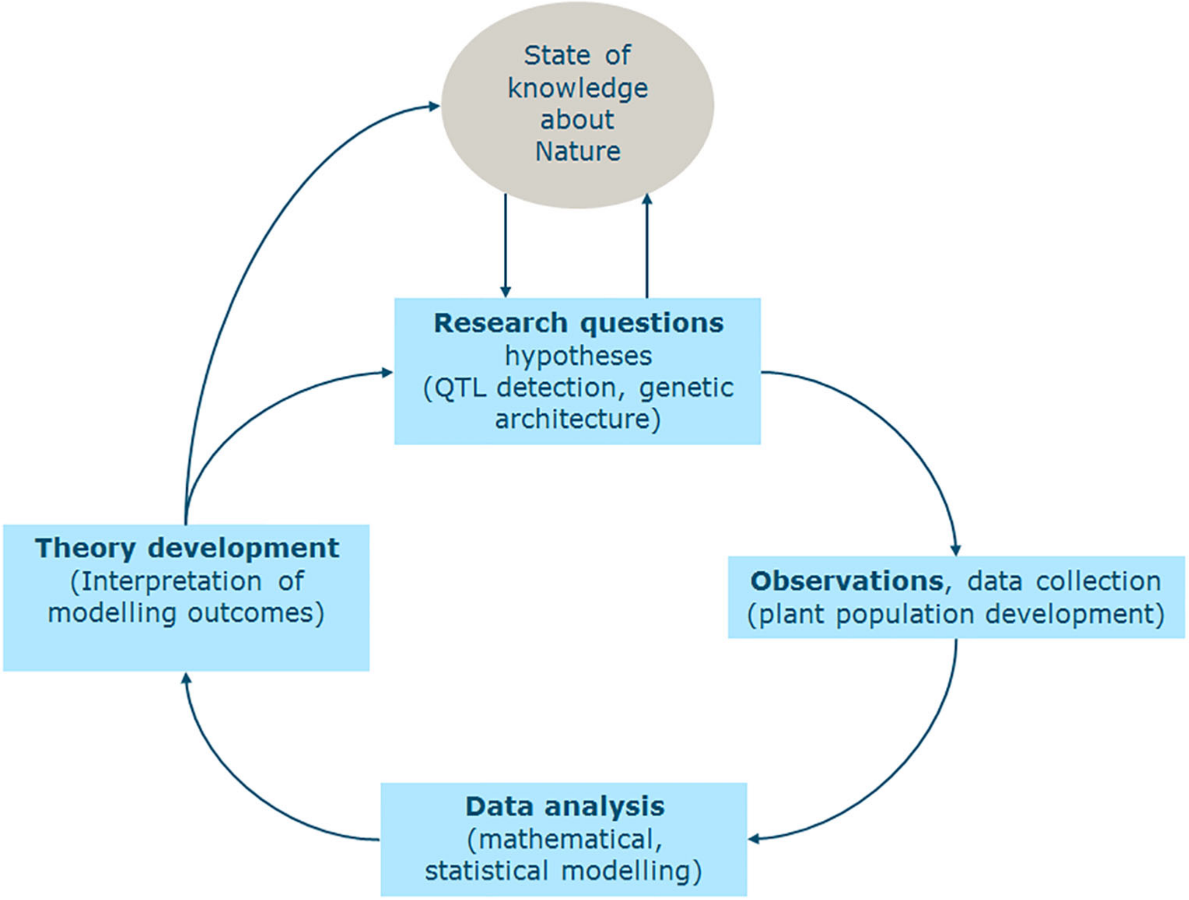

State of knowledge about Nature
Research questions hypotheses (QTL detection, genetic architecture)
Theory development (Interpretation of modelling outcomes)
Observations, data collection (plant population development)
Data analysis (mathematical, statistical modelling)

Figure 1.2: Possible illustration of the scientific process in plant science.

One more time Mendel was decisive at this level because the theoretical developments he proposed became the first scientific knowledge on genetics later coined as Mendel's laws (Mendel et al., 1993). Scientific knowledge is however never fixed and will be modified, completed or updated by further scientific developments, especially in a world in evolution. For example, Mendel continued his researches trying to explain quantitative variation, but he faced some limitations (Mendel et al., 1993). We had to wait for Fisher (1918) and new theoretical developments to better explain quantitative variation.

1.6 Conclusions

The description of the research process proposed in this introduction illustrates the inter-disciplinary nature of plant genetics. Indeed, the key of a successful research is certainly the combination of many disciplines (biology, statistics, mathematics, information technologies) to realize a number of connected operations. Each step of the scientific process influences the actions and decisions that can be taken in the next steps. Concerning plant genetics, the research question influences the type of population that should be developed. Then, the properties of the population should be integrated in the statistical model to describe the data, etc. Looking back, it is extraordinary to think that a single person like Mendel could carry out the whole research process. Mendel was an exception. Modern science, as most of the human activity, has been characterised by a phenomenon of specialisation and became progressively articulated between several people and/or even between several institutions (Durkheim, 1894; Ellul, 1954).

The success of science and technology depends therefore on the capacity to coordinate a wide range of operations without losing the sense and the logic of the global process. Increasing the consciousness of the whole process is certainly a key element if we want to make scientific progresses. This is a critical challenge, especially in a world characterized by rapid changes. Time and patience are necessary. Let us remember that Mendel took 10 years to run his experiments and to produce a unique paper. A single piece of work that opened a new branch of science.

Statue of Gregor Mendel in the garden of old Brno convent by Coeli - own work, Public Domain

Chapter 2

mppR: An R Package for QTL Analysis in Multi-parent Populations using Linear Mixed Models

Vincent Garin[1, 2], Valentin Wimmer[3], Dietrich Borchardt[3], Fred van Eeuwijk[1], Marcos Malosetti[1]

1. Biometris, Wageningen University & Research Center

2. C.T. de Wit Graduate School for Production Ecology & Resource Conservation (PE & RC)

3. KWS SAAT SE

The complete version of the software presented in this chapter can be downloaded from GitHub:

`https://github.com/vincentgarin/mppR`

A reduced version of the software was published on CRAN as:

Garin V., Wimmer V., Borchardt D., van Eeuwijk F., Malosetti M. (2018). mppR: Multi-Parent Population QTL Analysis. R package version 1.2.0. `https://CRAN.R-project.org/package=mppR`

Abstract

mppR is an add-on package for the statistical software **R** for QTL analyses in multi-parent populations composed of genotypes from more than one cross like NAM populations, diallels or factorial designs. **mppR** contains functions to assist the user in a range of activities of QTL analysis such as: data processing, QTL detection, visualisation of results, and estimation of QTL effects. **mppR** workflow is structured along main functions allowing to: 1) perform preliminary data quality control; 2) organize data into a single data object; 3) cluster parental lines based on ancestry; 4) perform QTL detection; 5) evaluate QTL discoveries by cross-validation; and 6) determine multi-QTL effect models. The search of QTLs can be done by 16 different models that vary with respect to two main aspects: 1) how the QTL effects are modelled (cross-specific, parental, ancestral or bi-allelic), and 2) the form of the variance covariance structure. The package is illustrated with a NAM maize population.

2.1 Introduction

Quantitative trait locus (QTL) analysis essentially consists in finding a relationship between DNA polymorphisms (e.g., SNPs) and phenotypic variation (Doerge, 2002). QTL detection methods greatly depends on the genetic properties of the population that is used. Historically, QTL detection has been performed in designed experimental populations involving two parental lines (bi-parental crosses). Several methods and software packages have been developed for QTL analysis for such populations, for a review see Varshney et al. (2016). Multi-parent populations (MPPs) are an alternative type of population that can improve the chances of QTL detection while broadening the range of research questions that can be answered (Cavanagh et al., 2008). MPPs can be seen as a compromise between bi-parental crosses and association panels (Myles et al., 2009). Different types of MPPs have been developed including nested association mapping (NAM) populations (McMullen et al., 2009), diallels (Blanc et al., 2006) and factorial designs (Bardol et al., 2013). More complicated MPPs can be created by intercrossing multiple founders followed by inbreeding, like in multi-parent advanced generation inter-cross (MAGIC) populations (Cavanagh et al., 2008). Here, we consider MPPs as a collection of genotypes that are derived by crosses between at least three different parents. In this paper, a MPP QTL analysis is akin to the joint analysis of such a population using a common marker map.

The development of an appropriate statistical methodology taking MPP properties into consideration is a *sine qua non* condition to fully exploit the potential of MPP genetic resources. The most critical question is how to account for genetic relatedness between the genotypes and how to integrate this information into the statistical model. A first simple option is to treat MPPs as an association mapping panel and to apply genome-wide association study (GWAS) QTL detection methods.

The use of a GWAS type of method presents several advantages. First, GWAS methods apply to almost any type of MPP design because the knowledge of population structure is not a prerequisite. The GWAS QTL detection is marker-based, that is, uses identity by state (IBS) information, and so generally only allows for two alleles at each tested position. A second advantage is the existence of powerful algorithms (e.g. EMMA - Kang et al. (2008)) that allow to scan in a reasonable amount of time large marker datasets. Finally, GWAS analyses are also based on sets of well-developed mixed models (Yu et al., 2006). These models allow to account for the population structure and the polygenic effect using a kinship matrix covering the relations between all genotypes from all populations (van Eeuwijk et al., 2010a; Rincent et al., 2014). MPP GWAS type of QTL detection can be performed using packages like **TASSEL** (Bradbury et al., 2007) the **R** library **GAPIT** (Lipka et al., 2012) or the **R** packages **GenABEL** (Aulchenko et al., 2007) and **Sommer** (Covarrubias-Pazaran, 2016).

However, a major limitation of GWAS methods is that they generally use bi-allelic marker models assuming two classes of effects at the QTL position. The bi-allelic assumption represents therefore a risk of failing to reflect the allele diversity potentially present at the QTLs within MPPs (Garin et al., 2017). Indeed, several factors like multiple alleles, cross-specific linkage phase between marker and QTL alleles, or interaction effects between the QTL and genetic background may cause complex allelic series.

Other approaches use available pedigree information to model or infer DNA transmission to the final lines starting from a set of parents or ancestors. This strategy uses identity by descent (IBD) information and gives model with more than two alleles, which can be more appropriate to model complex allelic series (Blanc et al., 2006; Xavier et al., 2015). For example, the **R** package **NAM** (Xavier et al., 2015) proposes to take into consideration factors which can lead to complex allelic series like the difference of linkage phase association between marker and QTL in different crosses. **NAM** uses incidence matrices containing the number of alleles received per parent to estimate random QTL effects and to control for the polygenic background in the rest of the genome. The software package **MCQTL** (Jourjon et al., 2005) functioning in a Linux environment is also an option. **MCQTL** combined with the **R** package **clusthaplo** (Leroux et al., 2014) computes linear models with various assumptions about the origin and number of QTL alleles (cross-specific, parental, or ancestral). We generalize the MCQTL approach to a mixed model context and add the bi-allelic model. In a mixed model context, we can allow for heterogeneity of variance present in MPPs, and other sources of random variation and/or dependence existing between genotypes. In addition, we incorporate a cross-validation strategy to evaluate the QTL detection performance of the different models. Finally, we developed a method to build multi-QTL (MQE) models that allow QTLs with different types of effects at different loci in contrast to the more rigid approach in **MCQTL** (Jourjon et al., 2005) that assumes the same type of effect across the genome. In a nutshell, **mppR** fits a wide range of models with different assumptions about the QTL effects and the variance covariance structure against which those QTLs are tested.

mppR contains functions to support data analysis from data processing to visualisation, staying within the **R** environment (Team, 2016). The use of **mppR** within the linear model framework is free, but some of the components that are based on mixed model technology depend on the **ASReml-R** package (Butler et al., 2009) and require a license. **mppR** is available from the following GitHub repository `https://github.com/vincentgarin/mppR`. A reduced version of **mppR** can also be downloaded from CRAN `https://CRAN.R-project.org/package=mppR`.

This paper is organized as follows. Section 2.2 describes the statistical methodology for the proposed MPP QTL detection procedures. Section 2.3 illustrates in detail the QTL detection procedure describing the different functions using a subset of the maize US-NAM population as example (Yu et al., 2008; McMullen et al., 2009).

2.2 Statistical methodology

2.2.1 Connectivity

mppR allows to analyse any type of MPP design with minimally two crosses between at least three different parents. In such designs, the possibility to estimate QTL parameters (identifiability) is linked to the notion of connectivity of the design. It is always possible to estimate one effect per cross. Therefore, the number crosses (n_c) constitutes the largest number of QTL effects that can be estimated. The estimation of parental and ancestral effects is linked to the connectivity of the MPP design (Rebaï & Goffinet, 2000). Taking for example the parental alleles, it is possible to estimate $n_p - 1 \leq n_c$ parental alleles per connected part of the design. Design connectivity can be defined using graph theory (Weeks & Williams, 1964). Following graph theory, an MPP design can be represented by a graph where parents (alleles) are vertices or nodes and crosses are edges or lines (Figure 2.1). A connected graph, is a graph where there exists a walk from any node i to any other node j. For example, the MPP design in Figure (2.1) is composed of two connected parts. Ideally, MPP QTL analysis should be run using only connected populations. The joint analysis of an MPP composed of several disjunct but internally connected parts is still possible. In that case, connectedness could follow from the sharing of a common ancestor by two parents of the design. For example if parents P_B and P_E of the MPP design of Figure (2.1) would receive their allele from the same ancestor, the MPP design would consist of a single connected part. For a bi-allelic model we assume that the design is fully connected. In any case, even if disconnected parts are analysed jointly, a minimal level of connection will be assumed by assuming that cofactors are shared in the whole population.

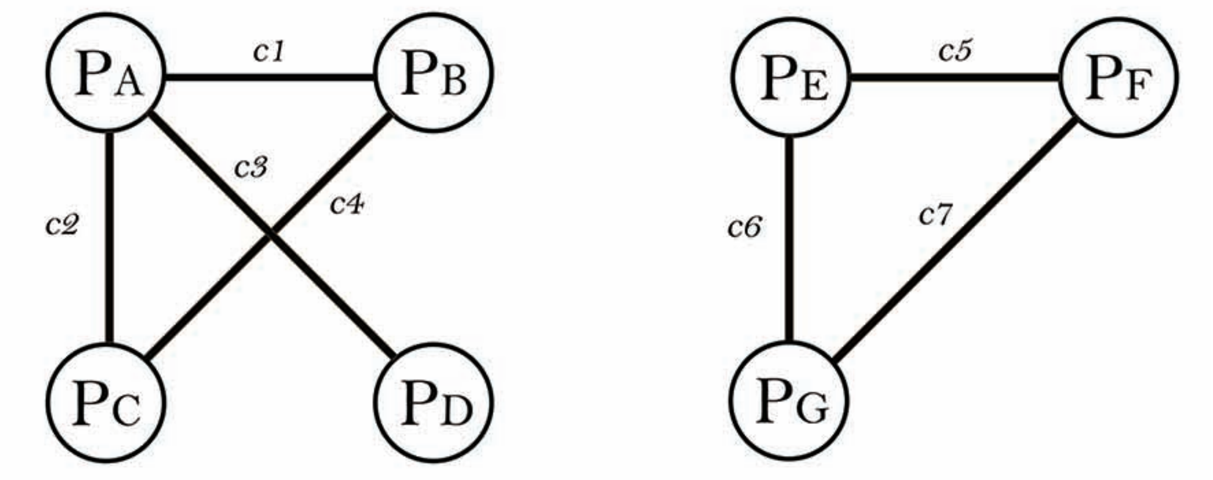


Figure 2.1: Example of a MPP design represented as a graph

2.2.2 General model

We propose to describe the QTL detection model with a focus on the following two main components: a) the assumption made on the form of the QTL effect and allele origin, and b) the assumption about the variance covariance structure of the residual. Let us start by defining the following underlying single locus QTL detection model describing the relationship between the phenotypic values and genotypes coming from several crosses (Rebaï & Goffinet, 1993):

$$y_{ijk} = \mu_{ij} + \alpha_i + \alpha_j + g_{ij} + e_{ijk} \tag{2.1}$$

where y_{ijk} represents the phenotypic value for the k^{th} individual from the cross between parents i and j. μ_{ij} is the cross mean and α_i and α_j represent the effects associated with the QTL alleles coming from parent i and j respectively. The QTL effects are assumed to be strictly additive (no dominance, no epistasis). g_{ij} is the random polygenic effect due to QTLs elsewhere in the genome with distribution $N(0, \sigma_g^2)$. Finally, e_{ijk} represents the random micro-environmental effect (plot error) having distribution $N(0, \sigma_e^2)$. In this model, σ_g^2 and σ_e^2 are unique meaning that the level of polygenic effect and environmental error is considered to be the same in each cross.

Model (2.1) can be rewritten in matrix notation:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{r} \tag{2.2}$$

where, $\boldsymbol{y}$ is the $[N \times 1]$ vector of phenotypic values. $N = \sum_{c=1}^{n_c} N_c$ where N_c is the number of genotypes coming from cross i. $\boldsymbol{X} = [\boldsymbol{X_c}|\boldsymbol{X_Q}]$ is the fixed effect incidence matrix and $\boldsymbol{\beta'} = [\boldsymbol{\beta'_c}|\boldsymbol{\beta'_Q}]$ the vectors of cross intercepts and QTL effects. $\boldsymbol{X}$ is composed of a part that links observations to the particular cross it belongs to ($\boldsymbol{X_c}$ an $[N \times n_c]$ matrix with n_c representing the number of crosses) and $\boldsymbol{X_Q}$ the part related with the QTL effect attached to the particular observation. $\boldsymbol{X_Q}$ is a matrix of dimensions $[N \times n_{al}]$ with n_{al} the number of QTL alleles that are assumed to segregate for the particular QTL locus. Several assumptions are possible concerning n_{al}. They correspond to different statistical models presented in the next section. The form of $\boldsymbol{X_Q}$ varies according to the type of QTL effect assumed. Finally, $\boldsymbol{r}$ represents the vector of random residual terms with distribution $N(0, \boldsymbol{R})$. We propose different models for the residual terms in section (2.2.4).

2.2.3 QTL effects

The QTL effect incidence matrix $\boldsymbol{X_Q}$ is the central term of the model. Assuming a diploid organism, the individual elements of $\boldsymbol{X_Q}$, x_{nl} take values between 0 and 2 and represent the expected number of copies of allele l with $l = 1, ..., n_{al}$ received by genotype n at the QTL position. The column number of $\boldsymbol{X_Q}$ (n_{al}) varies with the number of alleles assumed at the QTL position. We propose four models: cross-specific, parental, ancestral, and bi-allelic. These models correspond to different assumptions concerning the type of QTL effects and the allele origin. They are characterized by different ways to model genetic relatedness between genotypes using either IBD estimates, IBS information, or a combination of both.

Cross-specific model

The first model assumes that the QTL alleles that segregate within a particular cross are different from those that segregate in another cross. Cross-specific QTL effects can be seen as parental alleles interacting with the cross genetic background. Under this assumption, QTL alleles are nested within crosses and so QTL effects are estimated per cross. In the cross-specific model, $x_{nl} \in [0, 2]$ represents the expected number of allele copies received from one of the cross parents given the flanking markers. The expected number of parental allele copies is estimated using IBD probabilities computed by the package **R/qtl** (Broman et al., 2003). These probabilities are estimated with respect to the parents of each cross. For illustration purpose, let us take the following example of a MPP analysis combining material coming from two crosses: cross 1 ($P_A \times P_B$) and cross 2 ($P_A \times P_C$). In that case, we ignore the fact that the two crosses are connected since they share a common parent P_A. Therefore $\boldsymbol{X_Q}$ is a diagonal block structure with diagonal elements specifying the within cross allele origin. Model (2.2) can be re-written like that

$$\boldsymbol{y} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \boldsymbol{\beta_c} + \begin{array}{c} \begin{array}{cccc} P_A & P_B & P_A & P_C \end{array} \\ \left(\begin{array}{cccc} 2 & 0 & & \\ 1 & 1 & \multicolumn{2}{c}{0} \\ 0 & 2 & & \\ & & 2 & 0 \\ \multicolumn{2}{c}{0} & 1 & 1 \\ & & 0 & 2 \end{array}\right) \end{array} \boldsymbol{\beta_Q} + \boldsymbol{r} \tag{2.3}$$

It is not possible to estimate two effects per cross since the parental scores are linearly dependent. The design matrix for the QTL effect in a cross-specific model is therefore constrained by redefining the parental information of a cross as half the difference between parent i and parent j. Therefore, for the cross-specific model, $\boldsymbol{X_Q}$ is of dimension $[N \times n_c]$ where n_c is the number of crosses and the vector $\boldsymbol{\beta_Q}$ is of dimension $[n_c \times 1]$. The cross-specific model contains the upper limit for the number of QTL effects that can be estimated. Indeed, in connected MPPs, the maximum number of effects that can be estimated is $n_c \geq n_p - 1$ where n_p is the number of parents (Rebaï & Goffinet, 2000; Jansen et al., 2003). This model corresponds to the disconnected model described in Blanc et al. (2006).

Parental model

In the cross-specific model, all crosses are considered unrelated. A second option is the parental model that adds the connection between crosses via the parents shared between crosses. In that case, the parental QTL incidence matrix is simply obtained by re-arranging the columns of model (2.3) taking into consideration the connections created by the use of common parents. This model estimates one allele effect per parental line, which is considered to be independent of the genetic background. The QTL effect of parent p is assumed to be constant in all crosses where this parent has been used (Blanc et al., 2006).

In a connected MPP, if $(n_p - 1) < n_c$, one expects the parental model to be more powerful than the cross-specific model because the number of parameters to estimate is reduced (Blanc et al., 2006). The reduction in the number of parameters to estimate should also help to get better estimates of the QTL effects because the sample size used to estimate these effect increases (Li et al., 2005). Full half diallels, with at least four parents, represent the most connected system where the number of crosses $n_c = (n_p * (1 - n_p))/2$ is maximised with respect to the number of parents (Jansen et al., 2003). Coming back to the previous example (2.3), we integrate in the QTL incidence matrix the fact that the two crosses are connected via the common parent P_A.

$$\boldsymbol{y} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \boldsymbol{\beta_c} + \begin{array}{c} \begin{array}{ccc} P_A & P_B & P_C \end{array} \\ \begin{pmatrix} 2 & 0 & \\ 1 & 1 & 0 \\ 0 & 2 & \\ 2 & & 0 \\ 1 & 0 & 1 \\ 0 & & 2 \end{pmatrix} \end{array} \boldsymbol{\beta_Q} + \boldsymbol{r} \tag{2.4}$$

In the parental effect model, the matrix $\boldsymbol{X_Q}$ is of dimension $[N \times n_p]$ and $\boldsymbol{\beta_Q}$ is of dimension $[n_p \times 1]$. The parental model corresponds to the connected model described in Blanc et al. (2006).

Ancestral model

The third option, called ancestral model, goes one level up in the pedigree and uses relatedness between parents to cluster them into a reduced number of ancestral groups. We assume that parents belonging to the same cluster transmit the same allele (Jansen et al., 2003; Leroux et al., 2014). Different options can be used to cluster parental lines. One of them is the **R** package **clusthaplo** (Leroux et al., 2014). **clusthaplo** is an algorithm to cluster parental lines along the genome based on genetic similarity. **clusthaplo** uses a sliding window to define ancestral classes at each marker position based on local genetic similarities using marker scores within the window. If the local marker density is not large enough, **clusthaplo** uses the global genetic similarity defined by a kinship coefficient. **mppR** contains a function to call **clusthaplo**.

The ancestral QTL incidence matrix $\boldsymbol{X}^*_{\boldsymbol{Q}}$ can be obtained by modifying the parental IBD QTL incidence matrix $\boldsymbol{X}_{\boldsymbol{Q}}$ using **clusthaplo** results. The ancestral model uses therefore both IBD and parental relatedness IBS information. Continuing our example (2.4), let us assume that at the considered QTL position parents A and C belong to the same ancestral group A_1, and parent B falls apart in group A_2.

$$\boldsymbol{X}^*_{\boldsymbol{Q}} = \boldsymbol{X}_{\boldsymbol{Q}} \times \boldsymbol{A} = \begin{array}{c} \begin{array}{ccc} P_A & P_B & P_C \end{array} \\ \begin{pmatrix} 2 & 0 & \\ 1 & 1 & 0 \\ 0 & 2 & \\ 2 & & 0 \\ 1 & 0 & 1 \\ 0 & & 2 \end{pmatrix} \end{array} \times \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{array}{c} \begin{array}{cc} A_1 & A_2 \end{array} \\ \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 2 \\ 2 & \\ 2 & 0 \\ 2 & \end{pmatrix} \end{array} \tag{2.5}$$

The matrix $\boldsymbol{X}^*_{\boldsymbol{Q}}$ of the ancestral model is of dimension $[N \times n_a]$ where n_a is the number of ancestral alleles. The corresponding vector $\boldsymbol{\beta}^*_{\boldsymbol{Q}}$ is of dimension $[n_a \times 1]$. The elements of $\boldsymbol{\beta}^*_{\boldsymbol{Q}}$ represent the estimates of the ancestral additive effects. The ancestral-effect model corresponds to the LDLA models used by Bardol et al. (2013) and Giraud et al. (2014).

Parental and ancestral model constraints

The estimation of parental (ancestral) QTL effect also requires the application of a constraint to the QTL incidence matrix. From a theoretical point of view, it is possible to estimate maximally $n_p - 1$ $(n_a - 1)$ QTL effects per connected part of the design (Rebaï & Goffinet, 2000; Weeks & Williams, 1964). Therefore, the QTL effects are estimated setting to zero the most frequent parental (ancestral) allele within each connected part. For example in the example of Figure (2.1), we could have P_A set as reference of the first connected part and P_E set as reference of the second connected part. An alternative is to force the QTL effects to sum to zero. The sum to zero constraint will also take place within each connected part.

Bi-allelic model

The last possibility is the bi-allelic model. If the marker is at the QTL position, the bi-allelic model assumes that genotypes with the same SNP score transmit the same allele. Therefore, we assume that the same allele segregates in the whole population which connects all parts of the design that were not connected before.. Genetic relatedness is therefore defined based on marker IBS information only. In this model, using the most frequent allele set as reference, $\boldsymbol{X_Q}$ become a vector $[N \times 1]$ with values 0, 1 or 2 corresponding to the number of copies of the minor allele.

$$\boldsymbol{y} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \boldsymbol{\beta_c} + \begin{array}{c} SNP_1 \\ \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \\ 2 \\ 0 \end{pmatrix} \end{array} \boldsymbol{\beta_Q} + \boldsymbol{r} \tag{2.6}$$

This model corresponds to association mapping models (e.g., model B in Würschum et al. (2012)). In a connected MPP, if $(n_p - 1) < n_c$, the models can be ordered from less to more parsimonious models (cross-specific, parental, ancestral, bi-allelic).

2.2.4 Variance covariance structure

A second important part of the QTL detection problem is the data variance covariance structure (VCOV). Several assumptions are possible concerning the VCOV of model (2.2) $\boldsymbol{V} = Var(\boldsymbol{y}) = Var(\boldsymbol{r}) = \boldsymbol{R}$

Homogeneous residual term

The simplest form that can be assumed for $\boldsymbol{V}$ is a homogeneous residual term (HRT) variance. In this case, $\boldsymbol{R} = \boldsymbol{I_N}\sigma_r^2$. This corresponds to a linear model where residual terms are considered to be independent and to belong to the same distribution. In the HRT model, the variance of the polygenic term (σ_g^2) and the error variance (σ_e^2) of model (2.1) are both pooled in the unique variance residual term σ_r^2.

Cross-specific residual terms

Another possibility is to consider cross-specific variance residual terms (CSRT). The homogeneous random term assumption could not reflect the potential heterogeneity of variance between the different crosses. Indeed, genetic heterogeneity of the crosses and/or various levels of genetic distances between the parents could result in different levels of variance for the polygenic effects ($\exists \quad \sigma_{gc.i}^2 \neq \sigma_{gc.j}^2 \quad \forall i \neq j \quad i, j = 1, ..., n_c$). In such a situation, the heterogeneity of variance existing between crosses will require cross-specific residual terms ($\boldsymbol{R} = \bigoplus_{c=1}^{n_c} \sigma_{r_c}^2$) and $\boldsymbol{V}$ will take the form:

$$\boldsymbol{V} = \bigoplus_{c=1}^{n_c} \sigma^2_{r_c} = \begin{pmatrix} \boldsymbol{I_{N_1}}\sigma^2_{r_1} & & & \boldsymbol{0} \\ & \boldsymbol{I_{N_2}}\sigma^2_{r_2} & & \\ & & \ddots & \\ \boldsymbol{0} & & & \boldsymbol{I_{N_{n_c}}}\sigma^2_{r_{n_c}} \end{pmatrix} \tag{2.7}$$

Pedigree and HRT

A third option models directly the polygenic effect g_{ij} of model (2.1) by splitting the residual term into a part that follows a relationship structure determined by the kinship among individuals, and a nugget or independent residual, each of them with a separate variance component. Therefore, model (2.2) becomes

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{g} + \boldsymbol{r} \quad with \quad \boldsymbol{g} \sim N(0, \boldsymbol{G}\sigma^2_g) \tag{2.8}$$

The polygenic term $\boldsymbol{g}$ and the residual term $\boldsymbol{r}$ are assumed to be independent ($cov(\boldsymbol{g}, \boldsymbol{r}) = 0$). $\boldsymbol{G}$ is a genetic relationship matrix based on pedigree information computed with the method of Luo et al. (1992). The elements of $\boldsymbol{G}$ represent the inbreeding coefficient which for example takes values $0, 0.25$ or 0.5 for unrelated individual, half-sibs and full sibs respectively. In such case, $\boldsymbol{V} = \boldsymbol{G}\sigma^2_g + \boldsymbol{R}$

Once again we can have two assumptions for $\boldsymbol{R}$. In the first case, $\boldsymbol{R} = \boldsymbol{I}\sigma^2_r$ and $\boldsymbol{V}$ is defined as

$$\boldsymbol{V} = \boldsymbol{G}\sigma^2_g + \boldsymbol{I_N}\sigma^2_r \tag{2.9}$$

Pedigree and CSRT

In the second case, $\boldsymbol{R} = \bigoplus_{c=1}^{n_c} \sigma^2_{r_c}$ which gives

$$\boldsymbol{V} = \boldsymbol{G}\sigma^2_g + \bigoplus_{c=1}^{n_c} \sigma^2_{r_c} \tag{2.10}$$

The computation of pedigree models can sometime be problematic. The correlation between the genetic background matrix ($\boldsymbol{G}$) and the QTL term could be a possible cause. The combination of four types of QTL effects (cross-specific, parental, ancestral, bi-allelic) and four VCOV (HRT, CSRT, pedigree + HRT, pedigree + CSRT) gives a grid of 16 different QTL detection models.

2.2.5 Test statistics

The significance of the QTL effects $\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}$ can be estimated using the Wald test (Wald, 1943). In the case of a HRT or a CSRT model, after simplification, the Wald test can be rewritten like that:

$$W(\hat{\boldsymbol{\beta}}) = \boldsymbol{y}'\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{y}} \tag{2.11}$$

$W(\hat{\boldsymbol{\beta}})$ is distributed as a chi-squared distribution with the degrees of freedom being equal to the number of QTL alleles. Expression (2.11) shows that the test statistic depends on the correlation between the observed phenotypic values ($\boldsymbol{y}$) and the model fitted values ($\hat{\boldsymbol{y}}$) and on the estimated VCOV ($\hat{\boldsymbol{V}}$). For that reason, the choice of the QTL incidence matrix ($\boldsymbol{X_Q}$) should be such that the phenotypic variations is captured as accurately as possible. If these variations are due to parental or cross-specific effects, corresponding QTL effects should perform better at the price of a larger number of parameters to estimate. On the other hand, if the effects are similar through the MPP, a reduced number of parameters will capture this variability and allows gains in power. The VCOV structure should also be selected to reflect local patterns of variability. If heterogeneity is present between crosses, the CSRT model will give test statistics considering this heterogeneity.

2.2.6 QTL detection procedure

The QTL detection procedure proposed in **mppR** is based on the following steps: a) Optional significance threshold determination by permutation test (Churchill & Doerge, 1994); b) cofactors selection by simple interval mapping (SIM); c) multi-QTL model search using composite interval mapping (CIM) (Zeng, 1993, 1994); d) simultaneous evaluation of the selected candidate QTL positions after backward elimination.

Significance threshold determination

The QTL significance threshold can be determined by permutation. The use of permutation aims at reproducing the conditions of the null hypothesis (no QTL present or no association between the marker and the QTL) by breaking the link between the phenotype and the genetic markers (Churchill & Doerge, 1994). Permutations allow to build a null hypothesis for the test statistic that reflects the characteristics of the experiment and should be valid for any distribution of the quantitative trait (Churchill & Doerge, 1994). The number of permutations should be at least 1000. Alternatively, the user can also specify the significance threshold value.

Multi-QTL model determination

The determination of a multi-QTL model is done using CIM. Such a strategy is based on fitting the model using cofactors representing other QTLs than the tested QTL (Zeng, 1993, 1994). The selection of cofactors is based on a SIM scan using the following model where X_c is a cross-specific intercept and X_Q model the QTL effect.

$$\boldsymbol{y} = \boldsymbol{X_c \beta_c} + \boldsymbol{X_Q \beta_Q} + \boldsymbol{r} \tag{2.12}$$

Cofactors can be selected with minimum distance in between based on the -log10(p) value SIM profile. CIM profile is computed based on the following model

$$\boldsymbol{y} = \boldsymbol{X_c \beta_c} + \boldsymbol{X_q \beta_q} + \boldsymbol{X_Q \beta_Q} + \boldsymbol{r} \tag{2.13}$$

where $\boldsymbol{X_q}$ represents the selected cofactors. During the CIM scan, an exclusion window can be set around the tested QTL position to remove cofactors and avoid too strong collinearity between the cofactors and the tested position. Finally, the selected candidate QTLs can be simultaneously tested after a backward elimination. An optional confidence interval for each QTL position can be obtained using a -log10(p) value drop-off interval from a CIM profile where cofactors on the scanned chromosome were excluded.

2.2.7 Multi-QTL effect (MQE) model

A variation on the common QTL model with a single type of effect is the multi-QTL effect (MQE) model. In the MQE model, the QTLs present in the final model can have different types of effects (cross-specific, parental, ancestral or bi-allelic). To build an MQE model we use a forward selection procedure. For each QTL to be added, genome wide profiles with a consistent QTL effect are calculated for each types of QTL effects that have been specified by the user. The model that is fitted at each position within a profile is:

$$\boldsymbol{y} = \boldsymbol{X_c \beta_c} + \boldsymbol{X_{Q1} \beta_{Q1}} + \boldsymbol{r} \tag{2.14}$$

Where the (first) QTL ($\boldsymbol{X_{Q1} \beta_Q}$) has an effect that is either cross-specific, parental, ancestral, or bi-allelic along the genome. From each of these profiles, the most significant position based on the -log10(p) value statistic is selected (e.g., $\boldsymbol{X_{Q1.cr}}$, $\boldsymbol{X_{Q1.par}}$, $\boldsymbol{X_{Q1.anc}}$, $\boldsymbol{X_{Q1.biall}}$). Note that the selected QTL positions for the different types of effects can be different. The QTL that increases most the R^2_{adj} (2.18) is selected, with it type of effect, and added to the model as a cofactor for the next set of genome wide scans. If at step 1 we selected a bi-allelic QTL, then at step 2 the QTL profiles will be based on the following models:

$$\boldsymbol{y} = \boldsymbol{X_c \beta_c} + \boldsymbol{X_{q1.biall} \beta_{q1}} + \boldsymbol{X_{Q2} \beta_{Q2}} + \boldsymbol{r} \tag{2.15}$$

For the set of QTL effects specified by the user, genome wide scans are performed via a test for the QTL effect in the term $\boldsymbol{X_{Q2} \beta_{Q2}}$. The forward selection process stops when no further significant QTLs can be identified. At this point, a final list of QTLs is compiled by a backward elimination. A final model with t QTLs could look like:

$$\boldsymbol{y} = \boldsymbol{X_c}\boldsymbol{\beta_c} + \boldsymbol{X_{Q1.biall}}\boldsymbol{\beta_{Q1}} + ... + \boldsymbol{X_{Q(t-1).par}}\boldsymbol{\beta_{Q(t-1)}} + \boldsymbol{X_{Qt.anc}}\boldsymbol{\beta_{Qt}} + \boldsymbol{r} \tag{2.16}$$

2.2.8 QTL effect estimation

Once a final list of QTLs is determined, the estimates for the regression coefficients in the corresponding multi-QTL model provide the QTL effects. For this model a global goodness of fit can also be calculated using R^2. Partial R^2 statistics are indicators of the contributions of each individual QTLs.

Genetic effect estimation and interpretation

In the cross-specific model (2.2.3), the genetic predictors for the additive effects represent half the difference between the second parent and the first parent for their conditional QTL genotype probabilities. The elements of $\boldsymbol{\beta_Q}$ represent the within cross allele substitution effect.

For the parental (2.2.3) and the ancestral (2.2.3) models, from a theoretical point of view, it is possible to estimate a maximum of n_p-1 (n_a-1) QTL effects per connected part of the design (Rebaï & Goffinet, 2000). Within each connected part, the most frequent parental (ancestral) allele is used as reference. The estimated parental (ancestral) QTL effects must be interpreted as a deviation with respect to the connected part reference. Referring again to the example in Figure (2.1), if P_A is set as reference of the first connected part, its value will be zero and the other parent alleles effects (P_B, P_C and P_D) will represent deviations with respect to P_A allele. If the second part (P_E, P_F, and P_G) was analysed jointly with the first part, the effect of alleles P_F and P_G will represent deviation with respect to P_E and will be independent of the first part parental effects.

An alternative is to use a sum to zero constraint. In that case the parental (ancestral) QTL effects are forced to sum to zero. The individual QTL effects represent a deviation with respect to the central tendency. Here also, the constraints are defined within each connected part. For the bi-allelic model (2.2.3), the estimated genetic effect represents the additive effect of one allele copy of the minor allele with respect to the most frequent allele set as reference.

Global R^2

To compute the R^2 statistics, we compare the residual sums of squares of a full model with QTL(s), to the one of a reduced model without QTLs.

$$R^2 = 1 - \frac{\sum_{n=1}^{N} r^2_{n(full)}}{\sum_{n=1}^{N} r^2_{n(red)}} \tag{2.17}$$

Independent of the choice for the VCOV, R^2 is always computed from a linear model (HRT). The R^2 measurement can be adjusted to take into consideration the number of parameters used.

$$R^2_{adj} = 1 - \frac{\sum_{n=1}^{N} r^2_{n(full)}/df_{full}}{\sum_{n=1}^{N} r^2_{n(red)}/df_{red}} \tag{2.18}$$

Where df_{full} and df_{red} represent the degrees of freedom of the full and reduced model, respectively.

Partial R^2

For each single QTL, a partial R^2 can be calculated by the difference between the R^2 of the full model (all QTL positions) and the R^2 of the model that drops that particular QTL (difference R^2). **mppR** also calculates partial R^2 by comparing a model without QTLs with single locus QTL models (single R^2). These estimates can also be adjusted using formula (2.18). The partial R^2 is an estimation of the individual QTL contribution to the phenotypic variation. The difference and single R^2 give estimates of the lower and upper bound explained variance by individual QTLs.

2.2.9 Cross-validation

A cross-validation (CV) approach can be used to evaluate the performance of the QTL detection process and to assess the QTL effect in a pseudo-independent population (Utz et al., 2000). CV allows to assess predictability of QTL effects in data not used for model training. The proposed CV procedure is an adaptation of Utz et al. (2000) procedure to the MPP context. A single run of CV is composed of the following steps:

1. **Partitioning of the dataset**. The full dataset $(\boldsymbol{y_{DS}}, \boldsymbol{X_{DS}})$ is partitioned *within cross* into k subsets. Then for the k repetitions each k subset is successively used as validation set (VS) $(\boldsymbol{y_{VS}}, \boldsymbol{X_{VS}})$, the other $(k-1)$ subsets go into the training set (TS) $(\boldsymbol{y_{TS}}, \boldsymbol{X_{TS}})$.

2. **Explained genetic variance in the TS**. The training set is used to detect QTLs. These QTLs allow to evaluate the proportion of explained genetic variance in the TS using $\hat{p}_{TS} = \frac{R^2_{adj.TS}}{h^2}$ where $R^2_{adj.TS}$ is the adjusted R^2 (2.18) and h^2 the heritability to be specified by the user.

3. **Predicted genetic variance in the VS**. We can now use the estimated QTL effects in the TS ($\hat{\boldsymbol{\beta}}_{\boldsymbol{TS}}$) to predict phenotypic values in the VS ($\hat{\boldsymbol{y}}_{\boldsymbol{VS}} = \boldsymbol{X}_{\boldsymbol{VS}}\hat{\boldsymbol{\beta}}_{\boldsymbol{TS}}$). The proportion of predicted genetic variance in the VS is $\hat{p}_{VS} = \frac{R^2_{VS}}{h^2}$, where R^2_{VS} is the squared Pearson correlation between the observed and predicted values. $\hat{p}_{VS}$ is computed within cross. A measurement at the whole MPP level is obtained by calculating the weighted average of the within cross values ($\bar{p}_{VS}$) accounting for the cross sizes. The relative bias between $\hat{p}_{TS}$ and $\bar{p}_{VS}$ is $1 - (\frac{\bar{p}_{VS}}{\hat{p}_{TS}})$.

2.3 Illustration: US-NAM population QTL analysis

2.3.1 Overview

mppR contains functions to perform all steps of an MPP QTL analysis, starting from the data processing to the visualisation of results. Figure 2.2 shows a schematic representation of the **mppR** workflow. The first part concerns the raw data processing to gather all required data in a single `mppData` object. The `mppData` object can be used in the functions: `mpp_proc()`, `mpp_CV()` and `MQE_proc()` which are generic functions to perform QTL analyses, cross-validation, and multi-QTL effect model computations, respectively.

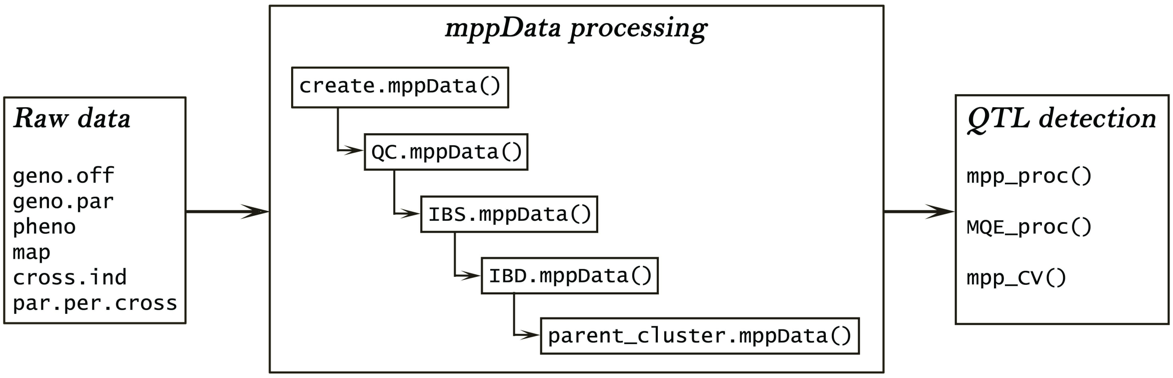


Figure 2.2: mppR workflow.

2.3.2 US-NAM data and raw data format

We included a subset of the maize US-NAM population (McMullen et al., 2009) within the **mppR** package as example dataset to illustrate the different functions. We introduce these data and the required format to be used in the workflow presented in Figure 2.2.

```
> library(mppR)
```

The data consist of three parts obtained from `www.panzea.org`. `USNAM_geno` is a random sample of the US-NAM population including the marker information of 506 genotypes and 102 markers. The entries include the parents and 500 recombinant inbred lines (RIL) coming from 5 crosses between the central line B73 and the peripheral parents CML103, CML322, CML52, Hp301, and M37W.

```
> data("USNAM_geno", package = "mppR")
> dim(USNAM_geno)

[1] 506 102

> rownames(USNAM_geno)[1:6]

[1] "B73"    "CML103" "CML322" "CML52"  "Hp301"  "M37W"

> table(substr(rownames(USNAM_geno)[-c(1:6)], 1, 4))

Z002 Z006 Z008 Z010 Z016
100  100  100  100  100
```

Genotypic data

Raw genotypic data used in **mppR** must be bi-allelic markers. The genotypic data are expected to be formatted as a `character matrix`, with one letter for each allele. So, when using the ACTG coding all possible genotypes are AA, CC, AC, etc. Missing values must always be coded `NA`. We impose a strict data format to make mppR functioning smoothly. Any deviation from this format will produce an informative error message.

To start the data processing, the genotypic data must be split into offspring and parent genotype scores. These matrices represent the `geno.off` and `geno.par` arguments in the `create.mppData()` function. For these two arguments, the genotypes define the rows with genotype identifiers as row names and the markers are in columns with the marker identifiers as column names. The order of the markers must be the same as the one in the `map` argument. The `geno.off` genotypes list must also be in the same order as the one of the `pheno` argument.

```
> geno.off <- USNAM_geno[7:506, ]
> geno.par <- USNAM_geno[1:6, ]
```

Map data

`USNAM_map` is a three columns genetic map with marker indicator, chromosomes and map positions given in centi-Morgan (cM). It has the required format for the argument `map` in function `create.mppData()`. The marker identifiers must be `character`. The chromosomes and genetic positions must be `numeric`. The list of markers must be column names of the `geno.off` and `geno.par` arguments.

```
> data("USNAM_map", package = "mppR")
> head(USNAM_map)

mk.names chr pos.cM
1   L00411   1    0.0
2   L00569   1    3.7
3   L00068   1    9.7
4   L01003   1   13.4
5   L00196   1   15.6
6   L00609   1   17.9

> map <- USNAM_map
```

Phenotypic data

The file `USNAM_pheno` is a `numeric matrix` containing the phenotypic measurements of 500 offspring genotypes for the trait upper leaf angle (ULA). The row names represent the genotype identifiers. They must be identical to the row names of `geno.off`. `USNAM_pheno` can be used as `pheno` argument for the `create.mppData()` function. The `pheno` argument can contain several traits. A cross indicator `character vector` can be formed by subsetting the genotype names. It specifies to which cross each genotype belongs and can be used for the `cross.ind` argument in `create.mppData()`.

```
> data("USNAM_pheno", package = "mppR")
> head(USNAM_pheno)

ULA
Z002E0001  75
Z002E0002  55
Z002E0005  60
Z002E0010  70
Z002E0011  75
Z002E0012  70

> pheno <-  USNAM_pheno
> cross.ind <- substr(rownames(pheno), 1, 4)
```

Cross parents information

The last raw data source is provided via the `par.per.cross` argument. It is a three column `character matrix` with one row per cross specifying: 1) the cross indicator that must be identical and appear in the same order with the one used in `cross.ind` ; 2-3) the parents 1 and 2 of the crosses. The parents' identifiers must be identical to the row names of `geno.par`.

```
> par.per.cross <- cbind(unique(cross.ind), rep("B73", 5),
+                        rownames(geno.par)[2:6])
```

The `par.per.cross` matrix can be used in the `design_connectivity()` function to obtain and visualize the connected parts. For example, using the illustration of Figure (2.1), we have

```
> ppc_ex <- cbind(paste0("c", 1:7),
+                 c("PA", "PA", "PB", "PA", "PE", "PE", "PG"),
+                 c("PB", "PC", "PC", "PD", "PF", "PG", "PF"))

> design_connectivity(ppc_ex)

$`1`
[1] "PA" "PB" "PC" "PD"

$`2`
[1] "PE" "PF" "PG"
```

For the next part of the illustration, we assume that the `geno.off`, `geno.par`, `map`, `pheno`, `cross.ind`, and `par.per.cross` objects are loaded in the global environment.

2.3.3 Data processing

The initial step is the processing of the raw data to gather all required data in a single `mppData` object. The functions `create.mppData()`, `QC.mppData()`, `IBS.mppData()`, `IBD.mppData()`, and `parent_cluster.mppData()` must be called in the defined sequence to form a complete `mppData` object. Any deviation from this sequence will be signalled by an error message.

create a `mppData` object

The function `create.mppData()` creates a unified `mppData` object containing all raw data sources.

```
> mppData <- create.mppData(geno.off = geno.off, geno.par = geno.par,
+                           map = map, pheno = pheno, cross.ind = cross.ind,
+                           par.per.cross = par.per.cross)
```

```
mppData object created!

500 genotypes
5 crosses
6 parents
1 phenotype(s)
1 connected part
```

Marker quality control - QC

Before QTL analysis, the raw data should go through a quality control (QC). This procedure will ensure that marker format is correct and that markers are informative, meaning that their segregation rate is sufficient to provide a reliable basis to investigate and estimate the QTL effects. The function `QC.mppData()` performs a default QC.

The user should be aware that it is difficult to propose general settings for QC that will be suitable for all MPPs. Moreover, the type of model fitted should also guide the QC. For example, if the user wants to give more emphasis to the cross-specific model, the QC should ensure that there is enough within cross segregation. On the other hand, the bi-allelic model assumes that the QTLs segregate in the whole population. Therefore for this model, minimum segregation can be evaluated at the whole population level. The procedure implemented in `QC.mppData()` is the following:

1. Remove markers with more than two alleles.
2. Remove markers that are monomorphic or fully missing in the parents.
3. Remove markers with a missing rate higher than `mk.miss` across the entire MPP.
4. Remove genotypes with more missing markers than `gen.miss`.
5. Remove crosses with less than `n.lim` genotypes.
6. Keep only the most polymorphic marker when multiple markers map at the same position.
7. Filter markers based on minor allele frequency (MAF). Different options are possible.

   A) The first one filter marker based on MAF at the whole population level, and/or on MAF within crosses. The markers with a MAF below a threshold given by `MAF.pop.lim` at the whole population level will be discarded.

   The user can specify the critical values for MAF within cross using `MAF.cr.lim`. By default, the within cross MAF values are defined by the following function of the cross-size N_c: $MAF(N_c) = 0.5$ if $N_c \in [0, 10]$ and $MAF(N_c) = (4.5/N_c) + 0.05$ if $N_c > 10$. This means that up to 10 genotypes, the critical within cross MAF is set to 50%. Then it decreases when the number of genotype increases until 5% set as a lower bound.

If the within cross MAF is below the limit in at least one cross, then marker scores of the problematic cross are either put as missing (`MAF.cr.miss = TRUE`) or the whole marker is discarded (`MAF.cr.miss = FALSE`). By default, `MAF.cr.miss = TRUE`, which allows to include a larger number of markers and to cover a wider genetic diversity.

B) An alternative is to select only markers that segregate in at least one cross at the `MAF.cr.lim2` rate.

```
> mppData <- QC.mppData(mppData = mppData, n.lim = 15, MAF.pop.lim = 0.05,
+                        MAF.cr.miss = TRUE, mk.miss = 0.1,
+                        gen.miss = 0.25, verbose = FALSE)
```

IBS processing

To compute a bi-allelic model, the user needs to convert genotype data into IBS format using `IBS.mppData()`. This function transforms genotype scores into 0, 1, 2 coding where the score represents the number of minor allele copies. The user can also perform imputation of the missing values (`impute = TRUE`). Different options are available, some of them rely on the `codeGeno()` function from the package **synbreed** (Wimmer et al., 2012).

```
mppData <- IBS.mppData(mppData = mppData, impute = TRUE,
impute.type = 'random')

Summary of imputation
total number of missing values                : 1737
number of random imputations                  : 1737
```

IBD processing

The other models (cross-specific, parental, and ancestral) use IBD probabilities. The function `IBD.mppData()` estimates IBD probabilities after converting the marker genotype data into within cross ABH format. For each cross, the maker scores of the two cross parents are used as reference. Homozygous offspring genotype scores similar to parent 1 get score `"A"` (`"B"` for parent 2). Heterozygous genotypes are scored `"H"`. If at least one of the parents is missing or the parents are monomorphic, the offspring will receive `NA`. The regular ABH conversion assumes that the reference parent scores are fully homozygous or missing. However, if some parent marker scores are heterozygous, the ABH conversion can still be performed setting argument `het.miss.par = TRUE`. In that case, when a parent score is heterozygous or missing and the other parent is homozygous, the function will try to infer the score of the allele that was transmitted by the heterozygous or the missing parent looking at the segregation pattern. Then the computation of the IBD probabilities is done by calling the function `calc.genoprob()` of the **R/qtl** package (Broman et al., 2003).

The type of population must be specified in argument `type`. Different population types are possible: F-type ("F"), back-cross ("BC"), backcross followed by selfing ("BCsFt"), double haploid ("DH"), and recombinant imbred lines ("RIL"). The number of F and BC generations can be specified using `F.gen` and `BC.gen`. The agument `type.mating` specifies if F and RIL populations were obtained by selfing or by sibling mating. DH and RIL populations are read as back-cross by R/qtl. For these two population types, heterozygous scores will be treated as missing values.

```
> mppData <- IBD.mppData(mppData = mppData, het.miss.par = TRUE, type = 'RIL',
+                         type.mating = 'selfing')

--Read the following data:
498  individuals
98  markers
1  phenotypes
--Cross type: bc
```

Parent clustering

The final step of data processing is to group parental lines for the ancestral model via the function `parent_cluster.mppData()`. The clustering of the parental lines is necessary to calculate the ancestral model. If the parent clustering is skipped, the other models (cross-specific, parental, bi-allelic) can still be computed.

The parental clustering can be realized calling the **R** package **clusthaplo** (`method = "clusthaplo"`). **clusthaplo** can be downloaded there: `https://cran.r-project.org/src/contrib/Archive/clusthaplo/`. In that case, `parent_cluster.mppData()` performs the clustering using the `"threshold"` method. `window` specifies the size of the window around the marker. The marker falling into this window will be used for the local clustering. The value specified in argument `K` is the minimum number of markers that should be present in the computation window. Below `K`, **clusthaplo** starts to use the general kinship coefficient. A visualisation of the clustering results can be obtained using `plot = TRUE`. In that case, the plots will be saved in a folder at the location specified in `plot.loc`. The function stores the average number of ancestral group along the genome in `mppData$n.anc`.

```
> set.seed(68769)
> mppData <- parent_cluster.mppData(mppData = mppData, method = "clusthaplo",
+                                   K = 10, window = 25, plot = FALSE)

> mppData$n.anc
[1] 5.112245
```

An alternative (`method = "given"`), is to provide your own parent clustering information via the argument `par.clu`. `par.clu` must be an `integer matrix` where the columns represent the parental lines and the rows the markers. The columns names must be the same as the parents list of the `mppData` object. The row names must be the same as the map marker list of the `mppData` object. At a particular position, parents with the same value are assumed to inherit from the same ancestor.

```
> data("par_clu", package = "mppR")

> mppData <- parent_cluster.mppData(mppData = mppData, method = "given",
                                    par.clu  = par_clu)
```

A summary of the `mppData` objects can be obtained calling the generic function `summary()`.

```
> summary(mppData)

object of class 'mppData'

Type of population:  Recombinant inbred line by selfing

No. Genotypes:  498

Crosses  Z002   Z006   Z008  Z010  Z016
Parent 1 B73    B73    B73   B73   B73
Parent 2 CML103 CML322 CML52 Hp301 M37W
N        100    100    98    100   100

Phenotype(s):  ULA
Percent phenotyped:  100

Total marker:  98
No. markers: 56 42
```

2.3.4 mppData manipulation

Three functions allow to manipulate and modify the mppData objects: subset.mppData(), pedigree_update.mppData(), and par_clu_chg.mppData().

mppData subset

Subsets from mppData objects can be obtained using the generic function subset(). The mppData objects can be subsetted by markers (mk.list) and/or by genotypes (gen.list).

```
> mppData_sub <- subset(x = mppData, mk.list = mppData$map[, 2] == 1,
+                       gen.list = sample(mppData$geno.id, 200))
```

Update mppData pedigree

The function pedigree_update.mppData() can be used to complete the pedigree relationship. By default, QC.mppData() only uses the parental relationships specified in the par.per.cross argument. However, if pedigree relationship among parents are known, a more complete pedigree information can be provided to the mppData object. This completed pedigree information can be given via the pedigree argument. pedigree is a data.frame containing one row for each genotype and four columns specifying: 1) if the genotype is an "offspring" (last generation) or a "founder" (all other generations); 2) the genotype identifier; and 3-4) the parents 1 and 2. The row giving the pedigree of an individual must appear before any row where that individual appears as a parent. Let us assume for the example that parent CML103 and CML322 share a common parent (A1).

```
> ped.old <- mppData[[4]] # old pedigree information
> founder.info <- data.frame(rep("founder", 2), c("CML103", "CML322"),
                             rep("A1", 2), c("A2", "A3"))
> colnames(founder.info) <- colnames(ped.old)
> ped.new <- rbind(founder.info, ped.old)
> mppData <- pedigree_update.mppData(mppData = mppData, pedigree = ped.new)
```

Change mppData parent clustering

The user can change the parent clustering information of a mppData object using the function par_clu_chg.mppData(). For example, if he wants to cluster the parents using another software than **clusthaplo**. The new parent cluster information must be provided in argument par.clu.

```
> data(mppData)
> data("par_clu", package = "mppR")
> mppData <- par_clu_chg.mppData(mppData = mppData, par.clu = par_clu)
```

2.3.5 QTL analysis

When all data elements are ready, the user can start the QTL analysis. In all functions involving the computation of a QTL model (`mpp_CIM()`, `mpp_CV()`, `mpp_perm()`, `mpp_proc()`, `mpp_SIM()`, `MQE_proc()`, `QTL_gen_effects()`, etc.), the arguments `Q.eff` and `VCOV` describe the type of QTL effect and the form of the VCOV. `Q.eff` takes the values `"cr"`, `"par"`, `"anc"`, `"biall"` for the cross-specific, parental, ancestral, and bi-allelic model, respectively. `VCOV` takes the values `"h.err"` or `"h.err.as"` for the HRT model, `"cr.err"` for the CSRT model, `"pedigree"` for the pedigree model with HRT, and `"ped_cr.err"` for the pedigree model with CSRT. A complete QTL analysis can be performed using the generic function `mpp_proc()`. This is a wrapping function for individual functions, indicated in parenthesis, performing each a part of the following QTL detection procedure:

1. SIM scan to select cofactors (`mpp_SIM()`). In that case, we fit an ancestral model (`Q.eff = "anc"`) with HRT residual term (`VCOV = "h.err"`).

2. Cofactor selection with SIM -log10(p) value above the `threshold` value with a minimum distance (`win.cof`) between selected positions (`QTL_select()`). The cofactor selection procedure is done per chromosome. It first selects the most significant position and then removes the positions in the neighbourhood of the selected position from the candidate list of QTLs/cofactors. The process continues until no position is significant any more. The threshold (`thre.cof` or `thre.QTL`) values can be determined by permutation using `mpp_perm()`.

3. CIM scan (`mpp_CIM()`) using the selected cofactors except within an exclusion window (`window`) around the selected cofactors where they are excluded from the model. It is possible to perform several consecutive run of CIM using `N.cim` .

4. Selection of QTL candidates with CIM -log10(p) value above `thre.QTL` and a minimum distance (`win.QTL`) between the selected positions (`QTL_select()`). The selection procedure is the same as for the cofactors.

5. If `backward = TRUE`, backward elimination on the list of selected QTL positions (`mpp_back_elim()`).

6. Estimation of the QTLs genetic effects (`QTL_gen_effects()`), the global and partial QTLs R^2 (`QTL_R2()`).

7. If `CI = TRUE`, computation of QTL confidence intervals from a CIM- profile (excluding cofactors on the scanned chromosome). The confidence interval is based on

a -log10(p) value drop-off value (drop).

8. plot of the -log10(p) value CIM QTL profile (plot.QTLprof()) and if plot.gen.eff = TRUE, visualisation of the genome-wide significance of the QTL effect per cross or per parent.

```
> QTL_proc <- mpp_proc(pop.name = "USNAM", trait.name = "ULA", trait = "ULA",
+                      mppData = mppData, Q.eff = "anc",
+                      plot.gen.eff = TRUE, N.cim = 1, thre.cof = 3,
+                      win.cof = 20, window = 20, thre.QTL = 3,
+                      win.QTL = 20, CI = TRUE, drop = 1.5,
+                      verbose = FALSE, output.loc = tempdir())
```

The results of mpp_proc() are returned as a list of **R** objects. These results are also saved in different files at the location specified in argument output.loc. The created folder contains a report (QTL_REPORT.txt) with a summary of results such as the number of detected QTL, the global R^2, and for each QTL the estimated genetic effects per cross or parent.

2.3.6 Multi QTL Effect (MQE) model

A multi-QTL effects model 2.2.7 can be determined using MQE_proc(). The user has to specify the types of tested QTL effects in the argument Q.eff. The window argument specifies the distance on both sides of an already detected QTL position where the search will be forbidden. A backward elimination on the final list of detected QTLs can be performed using (backward = TRUE). The results of the last MQE CIM run can be plotted using the function MQE_plot(). This MQE_plot() will colour the QTL positions corresponding to the type of QTL effect assumed at the position. This will be automatically done by MQE_proc() if plot.MQE = TRUE. The plot (plot_MQE.pdf) will be saved with the other results at output.loc.

```
> MQE <- MQE_proc(pop.name = "USNAM", trait.name = "ULA", mppData = mppData,
+                 Q.eff = c("par", "anc", "biall"), window = 20,
+                 verbose = FALSE, output.loc = tempdir())
```

2.3.7 QTL effects estimation

Once a list of QTL candidates has been determined, it is possible to estimate the QTL effect per cross or per parents (parental, ancestral, and bi-allelic model) using the function QTL_gen_effects(). For the cross-specific model (Q.eff = "cr") (2.2.3), the effects are given in absolute value and represent the substitution effect of one allele copy from the parent increasing the trait.

For the parental and ancestral models (Q.eff = "par" (2.2.3) or Q.eff = "anc" (2.2.3)), the QTL effects are given per parents and must be interpreted as deviation with respect to most frequent allele within the connected part set as reference. For the parental and ancestral models, the parental alleles are listed from the most (top) to the least frequent (bottom). For the ancestral model, parents with the same score correspond to the same ancestral allele according to **clusthaplo** results. In a NAM population, all crosses are connected via the central parent (B73), which is the most frequent allele, and was therefore set as reference.

```
> SIM <- mpp_SIM(mppData = mppData, Q.eff = "anc")
> cofactors <- QTL_select(Qprof = SIM)
> CIM <- mpp_CIM(mppData = mppData, Q.eff = "anc", cofactors = cofactors,
+                plot.gen.eff = TRUE)
> QTL <- QTL_select(Qprof = CIM)
> gen.eff <- QTL_gen_effects(mppData = mppData, QTL = QTL, Q.eff = "anc")
>
> summary(gen.eff, QTL = 1)

QTL effects
***********

Number of QTL(s): 1



QTL 1
-----

mk.names chr pos.cM
L00929   2      1



QTL effect per cross or parent:

        Effect   Std.Err    t-test   p-value      Sign Con.part Par.all
B73      0.000000 0.0000000  0.000000 1.000000e+00           c1      AA
CML103 -1.636412 0.7505413 -2.180310 2.971851e-02    *       c1      CC
CML52  -1.636412 0.7505413 -2.180310 2.971851e-02    *       c1      CC
Hp301  -2.688274 0.7321971 -3.671517 2.681344e-04  ***       c1      CC
M37W   -2.688274 0.7321971 -3.671517 2.681344e-04  ***       c1      CC
CML322 -4.864590 1.0477228 -4.643012 4.435899e-06  ***       c1      CC
```

For the bi-allelic model (2.2.3), the estimated genetic effect represents the additive effect of the minor allele with respect to the most frequent one, the latter set as reference. When parental genotype information is given, the results are given for each parent by multiplying the allele additive effect by the number of parent allele copies.

Results visualisation

A QTL profile can be obtained by passing the `mpp_SIM()` or `mpp_CIM()` results to the `x` argument of the function `plot.QTLprof()` (Figure 2.3). QTL or cofactors positions can also be plotted on the graph (dashed lines) using the argument `QTL`.

```
> plot(x = CIM, QTL = QTL, type = "l")
```

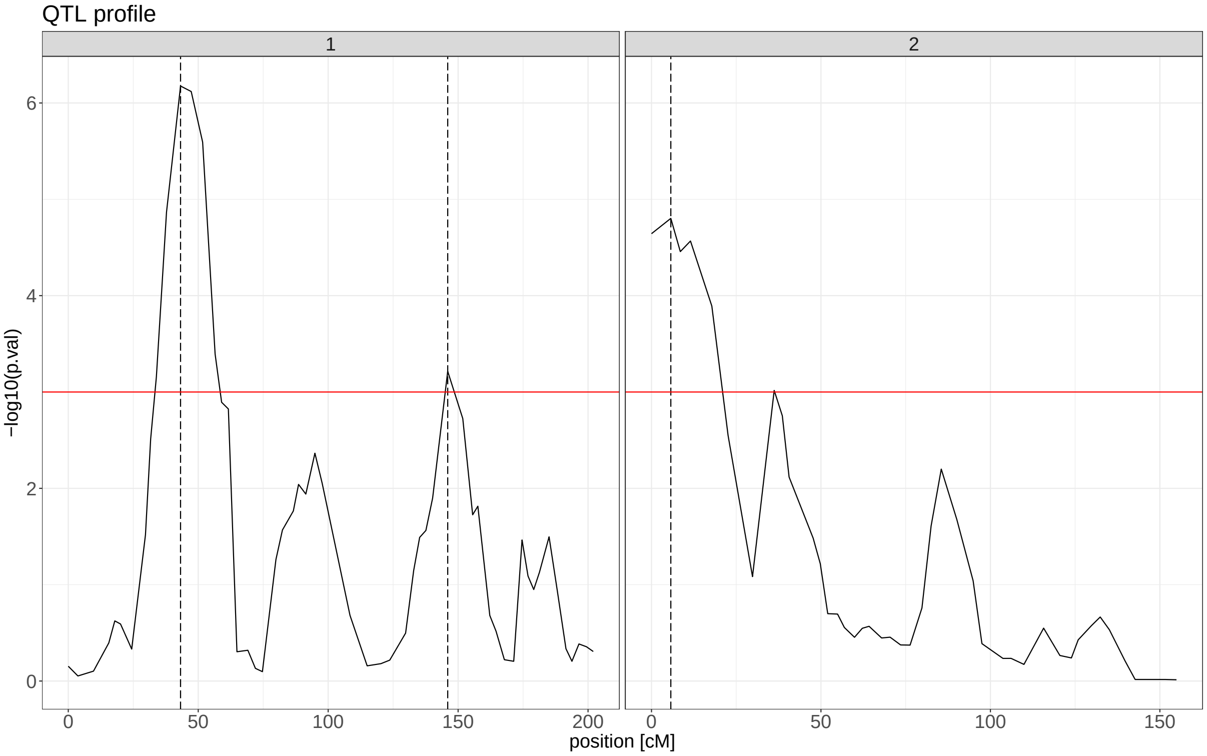


Figure 2.3: CIM QTL profile from an ancestral model using HRT. The cofactors positions are represented by vertical dashed lines.

`Mpp_SIM()` or `mpp_CIM()` results obtained with `plot.gen.eff = TRUE` can also be passed to the function `plot.QTLprof()` with argument `gen.eff = TRUE` to obtain a visualisation of the genetic effect distribution along the genome (Figure 2.4). Once again, the positions passed to the `QTL` argument will be drawn on the graph.

```
> plot(x = CIM, gen.eff = TRUE, mppData = mppData, QTL = QTL, Q.eff = "anc")
```

The interpretation of the genetic effect plot depends on the type of QTL effects. For a cross-specific model, red colour means that the allele coming from parent 1(A) increases the trait value, blue means that allele coming from parent 2(B) increases the trait.

For the parental and ancestral models, the effect must be interpreted as deviation with respect to the reference within a connected part. The reference allele is always defined

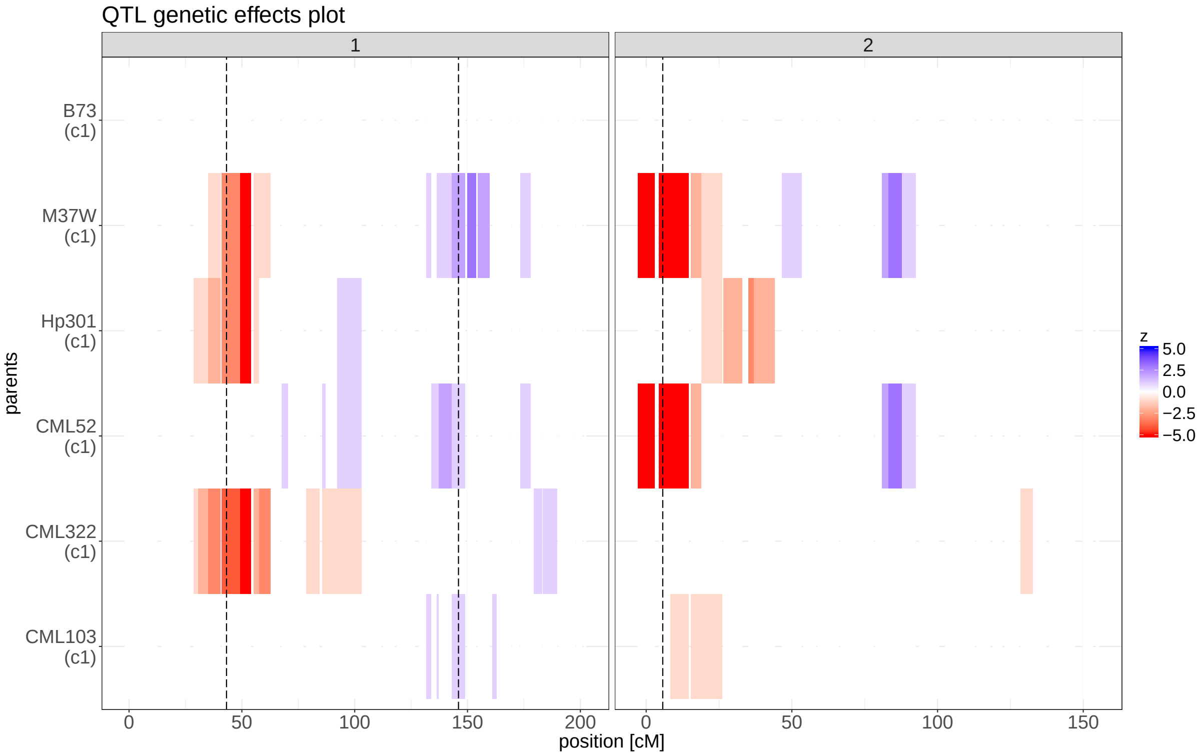


Figure 2.4: Visualisation of the genome-wide genetic effect significance from a CIM ancestral model using HRT. The vertical dashed lines represent the detected QTL positions.

as the most frequent allele. It can change at different positions. Therefore, it is not possible to establish a unique reference allele for the whole genome. The parental alleles significances are assessed per connected part. The red (blue) colour means that the parental allele decrease (increase) the trait value. Parents are ordered from the top to the bottom given the number of times their allele was set as reference in the whole genome. For example the upper parent was the one for which its allele was the highest number of times set as reference in the whole genome. These plots should be interpreted as a rough indication of signal distribution.

2.3.8 Cross-validation

The cross-validation procedure (2.2.9) can be performed by the function `mpp_CV()`. The arguments `Rep` and `k` represent the number of repetitions of the k-fold procedure. The heritability, allowing to express R^2 in terms of percentage of the explained genotypic variance, can be specified in `her`. By default `her = 1`, therefore the results are expressed in terms of phenotypic variation. The results of the CV procedure will be saved in a folder at `output.loc`. The different CV QTL profiles can be visualized using transparency plots with the function `plot_CV()` (Figure 2.5).

```
> set.seed(89341)
> CV <- mpp_CV(pop.name = "USNAM", trait.name = "ULA", mppData = mppData,
+              Q.eff = "cr", her = 0.4, Rep = 1, k = 3, verbose = FALSE,
+              output.loc = tempdir())
```

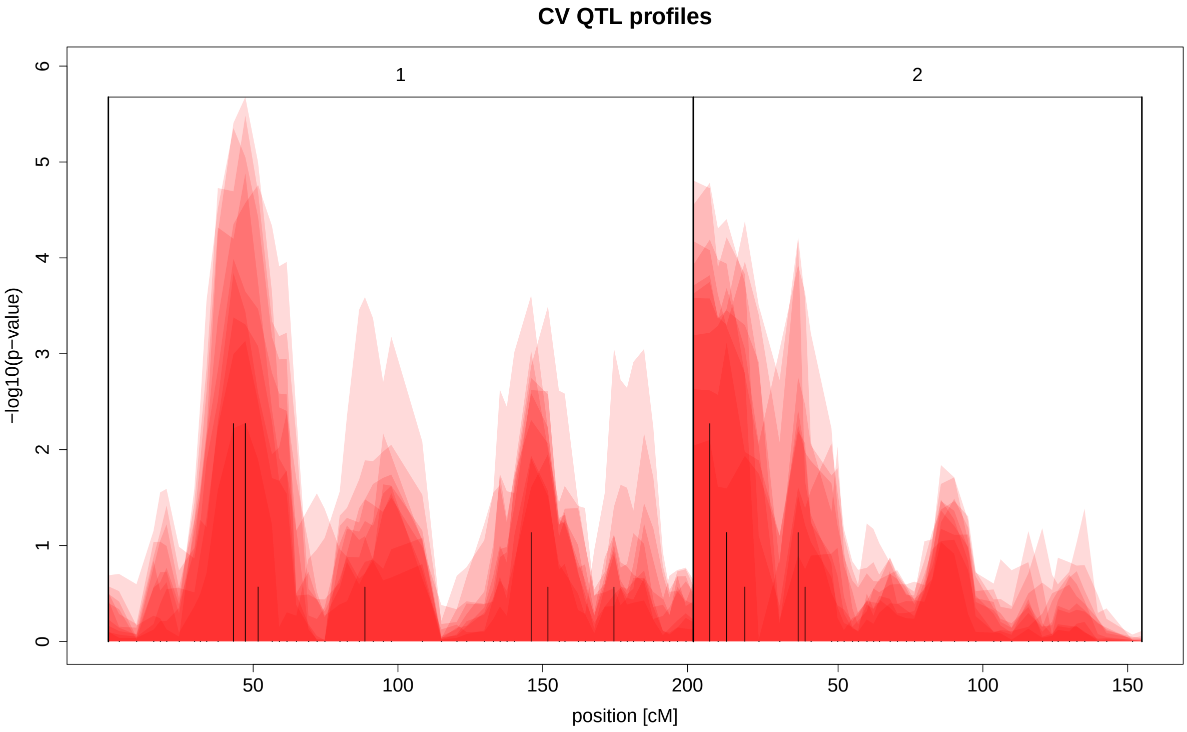


Figure 2.5: Transparency plot of CV results representing an overlay of QTL profiles. The black bars are proportional to the number of times a position was detected as QTL.

2.3.9 Parallelization

All functions involving genome scan(s) (`mpp_perm()`, `mpp_SIM()`, `mpp_CIM()`, `mpp_proc()`, `mpp_CV()` or `MQE_proc()`) can be executed in parallel for the HRT model only (`VCOV = h.err`). The number of cores can be specified using `n.cores`. Parallelization is done using functions from the **parallel** library.

2.3.10 Limitations

The QTL analyses performed in **mppR** are all based on "exact" (complete REML computation) mixed model computation at the tested position. We do not use approximation. This choice makes operations requiring the computation of multiple genome scans by mixed models very long. The realisation of permutation tests (`mpp_perm()`), cross-validation (`mpp_CV()`) or the determination of an MQE model (`MQE_proc()`) by mixed models (all models with `VCOV` different than `"h.err"`) are technically possible but can become cumbersone for modestly sized datasets (e.g., 500 genotypes and 1000 markers).

From our estimation, mixed model calculations can take 20 to 50 times longer than those for comparable linear models.

2.4 Conclusions

mppR is a package for QTL analysis in multi-parent populations working in the **R** environment. It can analyse any type of MPP design composed of more than one cross between at least three parents like NAM populations, diallels or factorial designs, where individual crosses are of the Fx, BCx, BCsFt, RIL or DH type. It contains functions to perform all the steps of the QTL detection from the data processing to the results visualisation. The QTL detection process is based on a grid of 16 possible models varying according to the type of QTL effects and the assumption about the residual variance covariance structure.

Chapter 3

How do the type of QTL effect and the form of the residual term influence QTL detection in multi-parent populations? A case study in the maize EU-NAM population

Vincent Garin[1, 2], Valentin Wimmer[3], Sofiane Mezmouk[3], Dietrich Borchardt[3], Marcos Malosetti[1], Fred van Eeuwijk[1]

1. Biometris, Wageningen University & Research Center

2. C.T. de Wit Graduate School for Production Ecology & Resource Conservation (PE & RC)

3. KWS SAAT SE

This chapter is published as:

Garin, V., Wimmer, V., Mezmouk, S., Malosetti, M., & van Eeuwijk, F. (2017). How do the type of QTL effect and the form of the residual term influence QTL detection in multi-parent populations? A case study in the maize EU-NAM population. Theoretical and Applied Genetics, 1-12.

Abstract

For the type of QTL effect in QTL models for multi-parent populations (MPPs) various options exist to define them with respect to their origin. They can be modelled as referring to close parental lines or to further away ancestral founder lines. QTL models for MPPs can also be characterized by the homo- or heterogeneity of variance for polygenic effects. The most suitable model for the origin of the QTL effect and the homo- or heterogeneity of polygenic effects may be a function of the genetic distance distribution between the parents of MPPs. We investigated the statistical properties of various QTL detection models for MPPs taking into account the genetic distances between the parents of the MPP. We evaluated models with different assumptions about the QTL effect and the form of the residual term using cross-validation. For the EU-NAM data we showed that it can be useful to mix in the same model QTLs with different types of effects (parental, ancestral or bi-allelic). The benefit of using cross-specific residual terms to handle the heterogeneity of variance was less obvious for this particular data set.

3.1 Introduction

The papers by Rebaï & Goffinet (1993) and Muranty (1996) are early examples of quantitative trait locus (QTL) detection with populations derived from more than two parents. More recently, QTL mapping using multi-parent populations (MPPs) have increased in popularity, where these MPPs include nested association mapping populations (NAM) (McMullen et al., 2009), diallels (Blanc et al., 2006) and factorial designs (Bardol et al., 2013), as well as more complicated MPPs created by intercrossing multiple founders followed by inbreeding, like in multi-parent advanced generation inter-cross (MAGIC) populations (Cavanagh et al., 2008). Here, we consider MPPs as a collection of crosses between at least three different parents and focus on a NAM population, which involves crosses between a central parent and a set of peripheral ones. An MPP QTL analysis would therefore be the joint analysis of such a population using a common marker map. Other authors have sometimes called it family mapping (Würschum, 2012), combined cross analysis (Li et al., 2005) or multiple-cross analysis (Jourjon et al., 2005).

To structure our reasoning we present some assumptions on the genetic properties of specific MPPs and make plausible how these properties can affect the choice of a statistical model for QTL mapping. In MPPs, the use of more than two parents potentially increases the allelic diversity in the MPP as a whole and so, increases the chance of segregation at any particular genomic position (Xu, 1998). MPPs allow to test genetic effects within different backgrounds (Blanc et al., 2006) and so extend the statistical inference space for the QTL effects (Xie et al., 1998). If the addition of parental genotypes does not increase the allelic diversity at a particular locus, the use of an MPP can still be advantageous with respect to a bi-parental cross because QTLs representing the same ancestral locus will benefit from an increased sample size to estimate their effects (Li et al., 2005).

The allelic diversity of the population may be related to the genetic distance among the parents of the population. When parents are genetically distant, the expected number of segregating alleles at a particular locus increases with the probability that these alleles are unique to a parental line. On the other hand when the parents are genetically closer, one expects a reduced number of alleles segregating at a particular locus and that the alleles are shared throughout the population.

It seems beneficial to QTL detection and QTL effect estimation that the statistical model for the phenotypic variation in an MPP takes into account the genetic properties (number of alleles and diversity) of the MPP. The genetic diversity contained in the MPP can be translated to properties of the statistical model via the form of the QTL effects and the structure of the polygenic variation. This latter variation is the natural variation against which to test QTL effects and determines statistical quantities like power and false positive rate.

Models for the QTL effect If the number of segregating alleles at a particular QTL increases (e.g. in a diverse MPP), the statistical model can capture that diversity by allowing more parameters for the QTL effect. In crossing schemes starting from pure lines, it implies estimating a maximum of one effect per parental line (parental model). On the other hand, when MPP genetic diversity is lower, parental relatedness can be used to infer a reduced number of ancestral segregating alleles that needs to be estimated, thereby increasing model parsimony and probably also QTL detection power (ancestral model). The lower bound will be reached when only two alleles are segregating in the totality of the MPP (bi-allelic model). Within fixed QTL effect models, the reduction of QTL parameters to be estimated to improve QTL detection power has been a central objective (Rebaï & Goffinet, 1993; Jansen et al., 2003; Blanc et al., 2006; Leroux et al., 2014).

The assumption concerning the number of alleles at a QTL position can also be seen from a pedigree or historical perspective. Indeed, when parental QTL effects are appropriate it implies that the allele origin is more recent than that of ancestral QTLs common to several parents. One can also argue that bi-allelic QTL effects are closer to the original mutation (Powell et al., 2010). These different assumptions about the number of alleles correspond to different ways of modelling genetic relatedness between lines at the QTL position. So far, QTL studies used models that assumed a single type of QTL effect or allele origin (e.g. Blanc et al. (2006) or Würschum et al. (2012)). We can however imagine that allele origin and the number of alleles segregating at a QTL position can vary along the genome. Therefore, in the present article, we will compare QTL models assuming a single type of genetic relatedness along the genome with a model that allows different types of allele origin.

Models for the polygenic term The genetic relationship between the parents of an MPP will also have an influence on the magnitude and structure of the polygenic or residual genetic term. The more diverse the population is, the more heterogeneous the residual variance is expected to be. The heterogeneity of the residual genetic variance may depend on the level of genetic relatedness between the parents of an MPP. Differences in genetic relatedness between pairs of parents can induce different levels of polygenic effect variation, inducing heterogeneity in residual genetic variance. Several studies of QTL mapping in MPPs applied linear models assuming a homogeneous variance for the residual genetic term (Li et al., 2005; Blanc et al., 2006; Yu et al., 2008). Depending on the particular MPP, this assumption might be unrealistic affecting the statistical test used to detect QTLs. To handle heterogeneous variances, some authors used transformed phenotypic data (Walling et al., 2000; Li et al., 2005; Guo et al., 2006). Others like Xu (1998) proposed to fit the QTL model by iteratively re-weighted least-squares. Polygenic effects can also be directly modelled for heterogeneity of variance in mixed models (Xu & Atchley, 1995; Yu et al., 2006; Wei & Xu, 2016). In our study we alleviated the restriction of classical linear models of homogeneous polygenic variance by using models with cross-specific variances for the residual genetic term.

We summarize our expectations for QTL detection in MPPs by the following propositions. We will refer to them to guide the discussion of the results.

Proposition 1 Models assuming common effects across the population, like the ancestral or the bi-allelic models, should perform relatively better in MPPs with a narrower genetic basis than in genetically diverse MPPs, since the probability of shared polymorphism is higher in the former than in the latter. In diverse populations however, the opposite is expected, requiring models with more QTL effect terms to capture the allelic diversity.

Proposition 2 The use of different types of QTL effects corresponding to different origins of the QTL allele at different positions along the genome should give a more adequate description of the phenotypic variation with increased QTL detection power.

Proposition 3 MPP genetic diversity should be reflected in the genetic variance of the crosses composing the population. Diverse MPPs present potentially more heterogeneity of the within cross genetic variance than less diverse populations. In diverse populations the use of cross-specific residual terms should give a better description of the data than a homogeneous residual term model. In more homogeneous populations the difference between cross-specific residual terms and homogeneous residual term should be minor.

In our paper, we evaluate various models for QTL detection in MPPs with different types of QTL effects and residual genetic terms. Beyond the currently existing methods, we propose a multi-QTL effect (MQE) model that allows various types of QTL effects at individual loci, where loci can differ in the most suitable type of QTL effect. We also relax the assumption of constant variance for the residual genetic term by using a cross-specific residual term (CSRT) model. We performed QTL detection in three subsets of the EU-NAM Dent population characterized by different degrees of genetic relatedness between the parents. The different models were evaluated using cross-validation (CV).

3.2 Materials and methods

To ensure the transparency and the reproducibility of our research, all data files, scripts and required software can be found in the following repository `https://github.com/vincentgarin/MPP_EUNAM`. This material makes it possible to reproduce all steps of the analysis, tables and figures of the article and of the supplemental material. To test the various models, we used the maize EU-NAM Dent panel (Bauer et al., 2013) and formed subsets.

Table 3.1: EU-NAM population crosses and simple matching coefficient (SM) between the central (F353) and peripheral parents. Average adjusted mean values ($\bar{X}$), genetic variance components (σ_g^2) and within cross heritability (h^2) for dry matter yield (DMY) and plant height (PH), and number of sampled lines per cross in the different subsets (short, heterogeneous, long).

			DMY			PH					
Cross	Parent	SM	$\bar{x}$	σ_g^2	h^2	$\bar{x}$	σ_g^2	h^2	short	het.	long
CFD11	UH304	0.761	192.5	47.8	59.4	288.5	37.4	68.9	60	65	0
CFD06	F252	0.633	178.1	120.9	78.0	285.7	65.9	79.9	76	0	0
CFD04	D09	0.618	187.4	31.3	41.9	284.5	67.3	86.5	78	85	0
CFD07	F618	0.589	194.7	20.2	31.2	291.4	45.0	79.5	79	0	0
CFD03	D06	0.586	189.9	78.2	70.9	292.2	52.8	73.8	68	90	0
CFD10	UH250	0.575	187.5	67.6	61.3	287.8	65.5	85.3	0	0	94
CFD09	Mo17	0.567	184.5	88.9	52.3	292.9	58.5	75.8	0	0	53
CFD12	W117	0.565	176.2	75.5	60.4	273.9	102.5	84.6	0	68	84
CFD05	EC169	0.558	184.2	46.5	56.4	283.5	64.8	84.7	0	0	66
CFD02	B73	0.557	193.2	95.6	64.1	294.8	60.8	81.1	0	53	64
Total									361	361	361

3.2.1 Genotypic data

The Dent panel of the EU-NAM population was composed of double haploid (DH) lines originating from 10 crosses between the central line F353 and 10 peripheral parents. This population was developed to represent the maize diversity in Northern Europe and also included the central parent of the US-NAM population, and it has been described in detail in Bauer et al. (2013) and Lehermeier et al. (2014). The offspring lines and the 11 parental lines were genotyped with the Illumina MaizeSNP50 BeadChip containing 56,110 single nucleotide polymorphisms (SNPs) (Ganal et al., 2011). Raw genotypic data were obtained from: `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50558`. We used the consensus map calculated by Giraud et al. (2014) available at: `http://maizegdb.org/data_center/reference?id=9024747`. From the original list of markers we selected the Panzea markers to avoid ascertainment bias (Bustos-Korts et al., 2016).

3.2.2 Population subsets

In the absence of pedigree information, genetic relatedness between parental lines can be estimated by similarity of molecular marker scores. We used the genetic similarity coefficient defined by Nei & Li (1979) (Supplemental file S1), which in our situation corresponds to the simple matching (SM) coefficient. We composed different subsets of the EU-NAM population (Table 3.1) involving parents with different levels of genetic relatedness between the central and the peripheral parents (see matrix of pairwise SM and PC - supplemental Table S2 and Figure S3). We formed: 1) a "short" subset with the five parents closest to the central parent; 2) a "long" subset with the five most distant parents from the central parent; and 3) a "heterogeneous" subset with a mixture of distant and close parental lines. The average SM between pairs of parents in the subsets decreased from the short to the long subset. $\bar{SM}$ is equal to 0.639, 0.613 and 0.573 for the short, heterogeneous and long subset, respectively.

To assure that all subsets were of equal size, we randomly selected 361 lines from the crosses of the short and heterogeneous subsets to make their size equal to that of the long subset. For QTL analyses, we removed markers that were not segregating in any of the MPPs, and that showed a minor allele frequency < 0.01 or missing values $> 10\%$ across the entire MPP. When multiple SNPs mapped at a single chromosome position, we selected the most polymorphic locus. After pre-processing, 5737, 5934 and 6212 SNPs were used for the short, heterogeneous and long subsets respectively, 3348 of which were common between the MPPs (see genetic maps - supplemental Figures S4).

3.2.3 Phenotypic data

We used the raw phenotypic data provided by Lehermeier et al. (2014) `http://www.genetics.org/content/198/1/3/suppl/DC1` and calculated the adjusted means, variance components, and heritability following their procedure (Table 3.1). Christina Lehermeier kindly communicated to us the list of genotypes used in her study which allowed us to use the same lines as in Giraud et al. (2014) and Lehermeier et al. (2014). We selected the traits with the lowest and highest average heritability over all crosses: biomass dry matter yield (DMY, decitons per hectare, $\frac{dt}{ha}$, $\bar{h^2} = 57\%$) at the whole plant level and plant height (PH, cm, $\bar{h^2} = 81\%$).

3.2.4 Statistical methodology

Let us start with the general single locus model for an MPP following the notation of Rebai and Goffinet (1993):

$$y_{ijk} = \mu_{ij} + \alpha_i + \alpha_j + g_{ij} + e_{ijk} \tag{3.1}$$

where y_{ijk} represents the phenotypic adjusted mean for the k^{th} individual from the cross between parents i and j. μ_{ij} is the cross mean and α_i and α_j represent the additive effects associated with the QTL alleles coming from parent i and j respectively. g_{ij} is the random polygenic effect due to QTLs elsewhere in the genome with distribution $N(0, \sigma_g^2)$. Finally, e_{ijk} represents the random micro-environmental effect (plot error) having distribution $N(0, \sigma_e^2)$.

Model 3.1 can be rewritten in matrix notation:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{r} \tag{3.2}$$

where, $\boldsymbol{y}$ is the $[N \times 1]$ vector of phenotypic values. $\boldsymbol{X} = [\boldsymbol{X_c}|\boldsymbol{X_Q}]$ is the fixed effect incidence matrix and $\boldsymbol{\beta}' = [\boldsymbol{\beta}_{\boldsymbol{c}}'|\boldsymbol{\beta}_{\boldsymbol{Q}}']$ the vector of cross intercepts and QTL effects. $\boldsymbol{X}$ is composed of a part that links observations to the particular cross it belongs to ($\boldsymbol{X_c}$ an $[N \times n_c]$ matrix with n_c representing the number of crosses) and $\boldsymbol{X_Q}$ the part related to the QTL effects. $\boldsymbol{X_Q}$ is a matrix of dimensions $[N \times n_{al}]$ with n_{al} the number of QTL alleles that are assumed to segregate at the particular QTL locus. The individual elements of $\boldsymbol{X_Q}$ take values between 0 and 2 and represent the number of allele copies received by genotype n at locus m. The number of columns n_{al} varies with the number of alleles assumed at the QTL position. We propose three models.

Parental model This first model assumes that each parent contributes a unique allele to the MPP ($n_{al} = n_{par}$). In a NAM population, peripheral parents are used only once so that their QTL effects are nested within the crosses between central and peripheral parents. In the parental model, the individual elements of $\boldsymbol{X_Q}$ are the expected numbers of QTL alleles received from the parents given the genotypes of the flanking markers, which were estimated using identity by descent (IBD) probabilities computed with the calc.genoprob() function from the R package qtl (Broman et al., 2003). The parental model corresponds to the connected model in Blanc et al. (2006).

Ancestral model A second option uses relatedness between parents to cluster them into a reduced number of ancestral groups, so $n_{al} = n_{anc} \leq n_{par}$. Under this model, parents belonging to the same cluster are assumed to transmit the same allele (Jansen et al., 2003; Leroux et al., 2014). For our analyses, the clustering of the parental lines was done at each marker position using a 2 cM window around the position with the R package clusthaplo (Leroux et al., 2014). The grouping is a function of the local similarity score defined by Li & Jiang (2005) and a global similarity defined by kinship coefficients. The results were stored in an ancestral matrix $\boldsymbol{A}$ that allows to modify IBD relationship of the parental model to account for ancestral relatedness (Figure 3.1). The ancestral model uses therefore both IBD and parental marker score information. This model corresponds to linkage disequilibrium linkage analysis (LDLA) models in Bardol et al. (2013) and Giraud et al. (2014).

Bi-allelic model The simplest model assumes that genotypes with the same SNP score transmit the same allele. Genetic relatedness is therefore defined based on marker identity by state (IBS) information. In this model $\boldsymbol{X_Q}$ becomes a vector with values 0, 1 or 2 corresponding to the number of copies of the minor allele ($n_{al} = 2$). For the bi-allelic

$$\boldsymbol{X_Q}* = \boldsymbol{X_Q} \times \boldsymbol{A} = \begin{array}{c} \begin{array}{ccc} P_A & P_B & P_C \end{array} \\ \begin{pmatrix} 2 & 0 & \\ 1 & 1 & \mathbf{0} \\ 0 & 2 & \\ 2 & & 0 \\ 1 & \mathbf{0} & 1 \\ 0 & & 2 \end{pmatrix} \end{array} \times \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{array}{c} \begin{array}{cc} A_1 & A_2 \end{array} \\ \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 2 \\ 2 & \\ 2 & \mathbf{0} \\ 2 & \end{pmatrix} \end{array}$$

Figure 3.1: Example of ancestral QTL incidence matrix formation. Parental matrix $\boldsymbol{X_Q}$ is transformed by ancestral matrix $\boldsymbol{A}$. Let us assume two crosses with a shared central parent: cross 1 ($P_A \times P_B$) and cross 2 ($P_A \times P_C$). Parents A and C are related to the same ancestral source

model missing marker genotypes were imputed by the software package Beagle (Browning & Browning, 2013) via the synbreed R package (Wimmer et al., 2012). This model is used in genome-wide association studies (GWAS), and corresponds to model B in Würschum et al. (2012) and the association mapping model in Liu et al. (2012).

The Wald test was used to test the global null hypothesis of all allele QTL effects equal to 0 (McCulloch & Searle, 2001, 5.39). The choice between the three models can be seen as a search for an optimum between parsimony and goodness of fit. If the allelic series are complex, like in a diverse population, then the parental model will be more suitable. On the other hand, if QTL effects are shared through the population, then the ancestral or the bi-allelic model will allow to gain in power by estimating a reduced number of parameters (for more detail see supplemental file S5).

If cofactors are included when searching for QTLs, $\boldsymbol{X}$ is augmented to $[\boldsymbol{X_c}|\boldsymbol{X_q}|\boldsymbol{X_Q}]$ where $\boldsymbol{X_q}$ is the cofactor incidence matrix and $\boldsymbol{\beta}' = [\boldsymbol{\beta}'_{\boldsymbol{c}}|\boldsymbol{\beta}'_{\boldsymbol{q}}|\boldsymbol{\beta}'_{\boldsymbol{Q}}]$ is the vector of fixed effects, with $\boldsymbol{\beta}'_{\boldsymbol{q}}$ representing the cofactors effects. In all our models, the QTL genetic effects were estimated by setting the most frequent allele as reference. In the parental and ancestral model it corresponded to the central parent F353 or the ancestral group containing F353.

Multi-QTL effect (MQE) model The parental, ancestral and bi-allelic models have already been used in other papers (e.g. Bardol et al. (2013) or Giraud et al. (2014)). These studies restricted the model to a single type of QTL effect, keeping the same type of incidence matrix across all loci. However, allelic effects in an MPP may vary across loci (Bardol et al., 2013), so a more flexible modelling approach would allow the incidence matrix to change from locus to locus. As an alternative, we propose a procedure to build multi-QTL effect models in which different loci can be modelled by different types of QTL effect (parental, ancestral or bi-allelic).

For the residual term $\boldsymbol{r}$ in model 3.2 we propose two models concerning the residual variance $\boldsymbol{R}$. The simplest model assumes constant variance (homogeneous variance residual term, HRT) $\boldsymbol{R} = \boldsymbol{I}\sigma_r^2$. This is the model used for the residual polygenic and environmental variances in the original paper by (Rebaï & Goffinet, 1993). A different model allows cross-specific variance residual terms (CSRT), which is more appropriate when heterogeneous polygenic effects are expected due to heterogeneous genetic distances among parents. In this case, $\boldsymbol{R} = \bigoplus_{c=1}^{n_c} \sigma_{r_c}^2$ where $c = 1, ..., n_c$ is the cross index. This model is similar to the one used by Xu (1998). From a theoretical perspective, the use of a single residual genetic variance component will lead to under and overestimation of this variance, depending on the cross, and therefore to an increase of the number of false positives and false negatives when heterogeneity of the polygenic effect is indeed high (for more detail see supplemental file S5).

3.2.5 Fast CSRT model

The estimation of an exact solution for the the CSRT model at each marker position during the CV procedure is computationally too demanding. Therefore, we propose a fast CSRT (f-CSRT) algorithm to compute approximate solutions. To calculate such an approximation we estimated first the residual term $\boldsymbol{R}$ in a model without QTL term and then used it in the Wald test to estimate the significance of the QTL effects along the genome (for more details see supplemental file S6).

3.2.6 QTL detection procedure

The combination of the four QTL effects (parental, ancestral, bi-allelic and MQE) with the two residual terms (HRT and CSRT) gives eight possible models for QTL detection. For the parental, ancestral and bi-allelic HRT models, the significance thresholds were determined by 1000 genome-wide permutations taking the $-log_{10}(p-value)$ of the upper 95% Wald statistic under the empirical null distribution as the critical value for rejection (Churchill & Doerge, 1994). The determination of significance thresholds For the CSRT models was computationally too demanding. Therefore, we used the same threshold values as the one of the corresponding HRT model. The significance thresholds of the MQE models were obtained by averaging the thresholds of parental, ancestral and bi-allelic models.

For the QTL detection methods based on QTL models with a single effect, a first run of simple interval mapping (SIM) was followed by two runs of composite interval mapping (CIM) by adding markers as cofactors (Zeng, 1993, 1994). We took care that QTLs (and cofactors) were spaced by a minimum distance of 20 cM. A multi-QTL model was created from the full list of QTLs detected after CIM by a backward elimination procedure with confidence level set at $\alpha = 0.01$. We used the same procedure as Han et al. (2016) to compute the proportion of genetic variance explained by the QTLs in the training set

(TS): $pTS = R^2_{adj}/h^2$. For the HRT model we used $R^2_{adj} = 1 - \frac{RSS_{full}/df_{full}}{RSS_{red}/df_{red}}$. RSS_{full} and df_{full} are the residual sum of squares and degree of freedom of a model including QTLs while RSS_{red} and df_{red} come from a model without QTLs. Note that both models contained a cross-specific intercept term that removes the between cross variation.

For the CSRT model, we used the likelihood R^2 defined by Cox & Snell (1989): $R^2_{LR} = 1 - exp(-\frac{2}{n}(logL_{full} - logL_{red}))$ where L_{full} and L_{red} represent the likelihood statistic of the full and reduced model respectively. For the CSRT model, following the recommendations of Sun et al. (2010), we estimated the likelihood of the reduced and full model using maximum likelihood estimation (not REML). We adjusted the likelihood R^2 using formula 2 from Utz et al. (2000): $R^2_{adj} = R^2 - [(\frac{df_{QTL}}{df_{full}}) * (1 - R^2)]$.

To build the multi-locus MQE models, we used a forward selection approach, where at each step, a new QTL was added that was allowed to have either a parental, an ancestral or a bi-allelic effect. To identify a new QTL, we computed three genome-wide profiles using the same type of QTL effect for the tested position (parental, ancestral or bi-allelic). Then we selected in each profile the most significant position based on the $-log_{10}(p - values)$ with its type of QTL effect. From these candidate positions (and effects) we selected the one that increased the most the model R^2_{adj}. The selected position with its type of QTL effect entered the model, and the process was repeated until no more significant positions could be added (for more detail see supplemental file S7).

HRT models were fitted by least-squares (lm() function in R), and CSRT by restricted maximum likelihood (REML) using the asreml-R package (Butler et al., 2009). For the computation of the likelihood R^2 we used the R package nlme (Pinheiro et al., 2017). The results of the f-CSRT models were obtained using the Wald statistics as described in S6. All procedures in this study have been compiled in R packages and are available in the repository (`https://github.com/vincentgarin/MPP_EUNAM/software/mppR_1.0.tar.gz`).

3.2.7 Cross-validation

We adapted the CV procedure described by Utz et al. (2000) to the MPP context. For each of the 48 combinations of QTL model and scenario we performed 100 CV runs by replicating 20 times a five-fold CV procedure. One run of CV was composed of the following steps: 1) The full dataset was partitioned at within-cross level into a training set (TS) and a validation set (VS); 2) QTL detection was performed using the TS and the proportion of genetic variance explained in the TS was computed by $\hat{p}_{TS} = R^2_{adj.TS}/h^2$; and 3) The proportion of genetic variance predicted in the VS was calculated by $\hat{p}_{VS} = cor(\boldsymbol{y_{VS}}, \boldsymbol{\hat{y}_{VS}})/h^2$, representing the Pearson correlation between the observed values ($\boldsymbol{y_{VS}}$) and the predicted values ($\boldsymbol{\hat{y}_{VS}} = \boldsymbol{X_{VS}}\boldsymbol{\hat{\beta}_{TS}}$). The $\hat{p}_{VS}$ were computed within crosses. An estimate at the full MPP level was obtained by taking a weighted average of the within cross values ($\bar{p}_{VS}$) accounting for the cross sizes. We evaluated the relative bias of a model by looking at the difference between ($\hat{p}_{TS}$) and ($\bar{p}_{VS}$).

To reduce the computational time for CV, we thinned the set of markers by selecting the most polymorphic marker at every 1, 1.05, and 1.05 centi-Morgan for the long, heterogeneous and short subset, respectively. For each CV scenario we determined the significance threshold running 1000 permutations on the full dataset. The threshold of the MQE model was again determined by averaging the values obtained for the parental, ancestral and bi-allelic models. For the CV procedure we used the f-CSRT approximation for threshold computation and QTL detection in all scenarios using cross-specific residual terms.

3.3 Results

3.3.1 Subset properties

For both traits, DMY and PH, the estimated genetic variance per cross tended to increase when the genetic relatedness between the peripheral and the central parent decreased (Table 3.1 and supplemental Figures S8). But, as in other studies (e.g. Hung et al. (2012)), this relationship was not significant. The degrees of relatedness based on the allele clustering from clusthaplo resulted in 4.02, 4.09, and 4.56 ancestral alleles on average for the short, heterogeneous, and long subsets respectively. This means that the difference in diversity between subsets was not high. The $-log_{10}(p - value)$ significance thresholds that were computed increased from the parental to the bi-allelic model, probably due to the reduction of auto-correlation between the consecutive tests (Supplemental Tables S9).

3.3.2 Full subsets QTL detection

Table 3.2 presents the results of QTL detection using the full (non-partitioned) subsets for the different combinations of QTL effect and type of residual variance. For the type of QTL effect, we noticed that for DMY, the QTL detection results are similar in the short subset across the different models. In the heterogeneous and long subset however, the more parsimonious models (ancestral and bi-allelic) detected more QTLs and explained a larger percentage of genetic variation. For PH, this tendency was inverted. For example, the parental CSRT model explained 50.4% of the genetic variance while the ancestral and bi-allelic model explained 40.1% and 38.3%, respectively. The MQE models detected more QTLs and explained a larger part of the genetic variance (see also Figure 3.2). This was especially true for the MQE CSRT model because, except for PH in the heterogeneous subset, it explained the largest part of the genetic variance. Concerning the residual term, the results of the HRT and CSRT models were similar for DMY. For PH we could observe, based on the explained genetic variance, that the CSRT models generally outperformed the HRT models.

Table 3.2: QTL detection results of the full subsets analyses (short, heterogeneous, long) per trait (DMY, PH) for the different QTL effects (parental, ancestral, bi-allelic and MQE) and types of residual term (HRT, CSRT).

		DMY				PH			
		parental	ancestral	bi-allelic	MQE	parental	ancestral	bi-allelic	MQE
short	HRT	3^a (20.6)^b	3 (20.6)	3 (18.7)	3 (1/2/−)^c (20.5)	7 (43.6)	6 (40.2)	7 (39.7)	6 (3/1/2) (41.6)
	CSRT	3 (19.4)	4 (21.9)	4 (20.3)	4 (1/2/1) (22.6)	8 (50.4)	6 (40.1)	8 (38.3)	9 (5/1/3) (52)
het.	HRT	-	3 (13.2)	3 (15.3)	3 (1/-/2) (18.5)	7 (46.5)	9 (47.9)	6 (39.1)	10 (4/2/4) (55.3)
	CSRT	1 (8.9)	3 (15.4)	3 (14.5)	4 (1/1/2) (20.8)	10 (57.3)	11 (58)	8 (46.6)	9 (3/3/3) (53.1)
long	HRT	2 (11.3)	1 (5.9)	5 (22.2)	7 (1/-/6) (32)	8 (42.8)	7 (38.7)	8 (38.5)	5 (1/3/1) (35.3)
	CSRT	2 (10.4)	2 (9.1)	5 (22.1)	8 (3/-/5) (37)	8 (43.5)	8 (43)	9 (38.8)	10 (1/4/5) (49.3)

a. Number of detected QTLs. - for no QTL detected.

b. Global adjusted R^2 in %

c. Number of detected QTLs per incidence matrix type (parental/ancestral/bi-allelic)

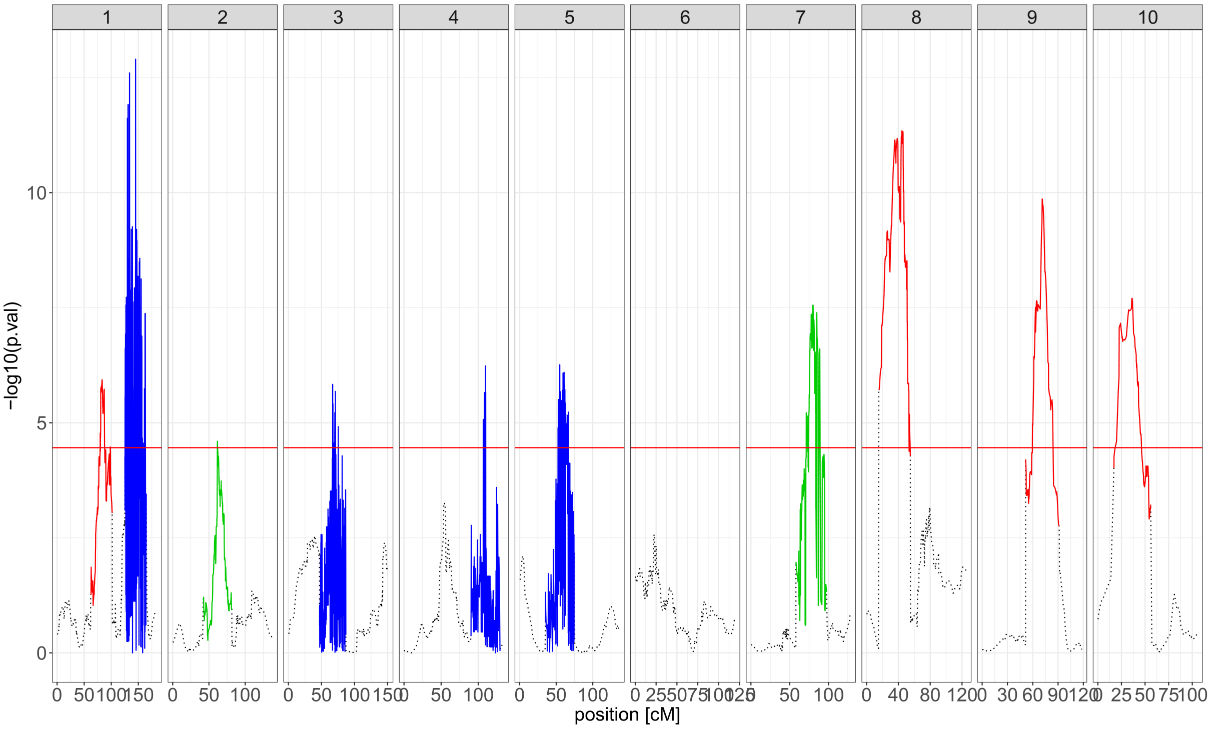


Figure 3.2: Example of MQE QTL profile result for PH in the heterogeneous subset. The colours drawn 40 cM around the detected positions represent the type of QTL effect at that locus (red: parental, green: ancestral, blue: bi-allelic).

3.3.3 Cross-validation

The plots in Figure 3.3 contain the CV results and show the similarity between the HRT and CSRT models. Further, the MQE model had the largest pTS in all configurations. The MQE pVS was also the highest or equal to the highest single type of QTL effect model for DMY but not for PH. For PH, the different types of QTL effect model tend to give similar results in term of pVS.

In the short subset for DMY we observe that pVS increased with the parsimony of the model since the ancestral and the bi-allelic model obtained larger scores. For PH in the same subset, the results were opposite with the parental model having a larger average pVS. In the long subset, for DMY, we could notice no difference in terms of pVS between the parental, ancestral and bi-allelic model. For PH however, the bi-allelic model performed better.

A final noteworthy result is that the difference between pTS and pVS (bias) was often reduced for the more parsimonious models especially for the bi-allelic model. This is for example the case in the short subset for both traits .

3.4 Discussion

3.4.1 Subset properties

The results of parental clustering of the different subsets showed that the genetic diversity was high in the three subsets because the average clustering results showed a number of ancestral alleles that was close to the theoretical maximum number of six, being the number of parental alleles. This is consistent with the fact that this maize population was designed to reflect Northern European genetic diversity. The limited amount of difference between the subsets can be explained by the relatively reduced range of genetic similarities across the crosses as reflected in SM coefficients between the parents of the crosses (Table 3.1).

3.4.2 Model performance given population diversity and type of QTL effect

Our first proposition, implied that the relative performance of QTL models in terms of QTL detection would increase with model parsimony when the MPPs are derived from genetically close parents. The underlying reasoning is that higher genetic proximity allows reducing the number of parameters needed to model the QTL effect, producing a gain in power for QTL detection. We expected that the pVS would increase with model parsimony in the short subset. On the other hand, in the long subset, models with a higher number of parameters like the parental model would help to model the assumed increased diversity and give better results in terms of pVS.

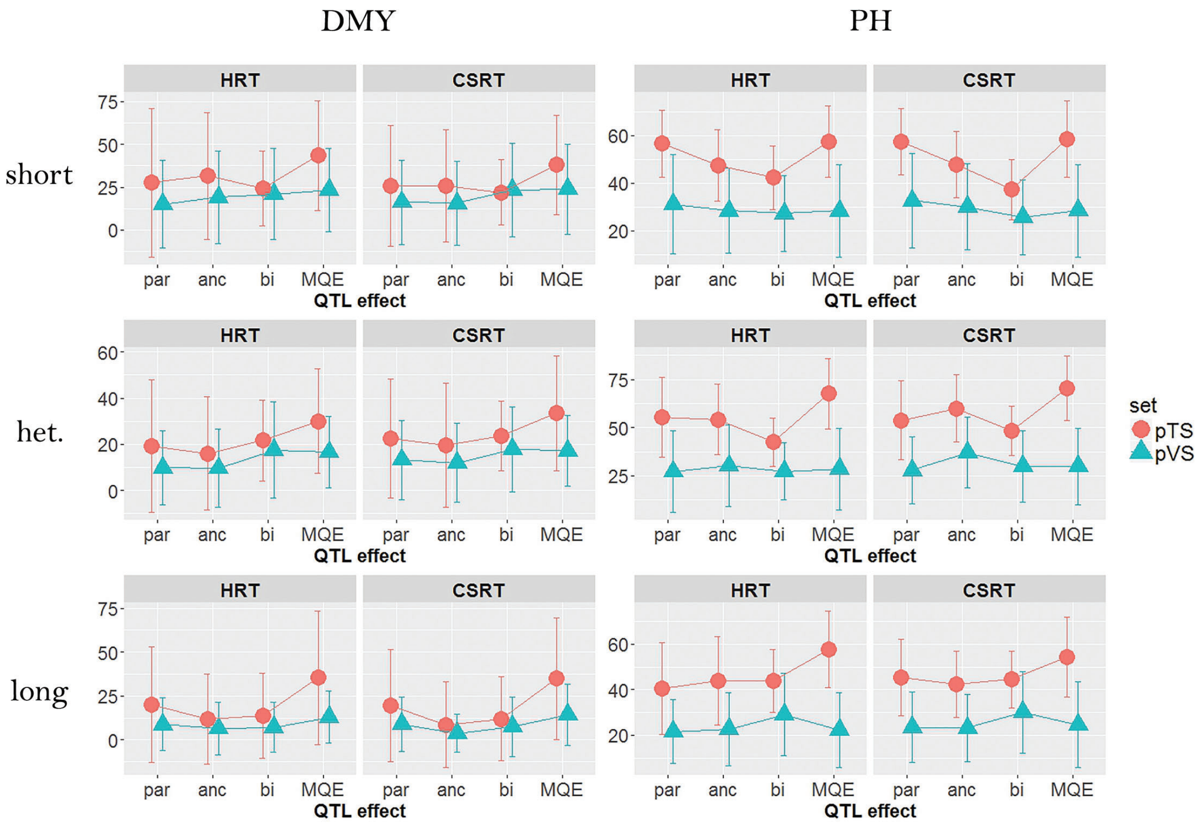


Figure 3.3: Cross-validation results over 100 runs. Average proportion of explained and predicted genetic variance (+/-2*standard deviation) in the training and validation sets for each combination of trait (DMY and PH), subset (short, heterogeneous, and long), type of QTL effect (parental, ancestral, bi-allelic, and MQE), and residual term (HRT and CSRT).

We did not observe these trends in the CV results (Figure 3.3). The increase of pVS from the parental to the bi-allelic model in the short subset for DMY is conform to our expectation. However, for PH in the short subset, we observed the opposite trend with pVS decreasing with model parsimony. In the long subset we expected that the pVS would decrease with more parsimonious models. But for PH, we noticed that the bi-allelic model gave the best result. Various reasons can be mentioned to explain that we did not observe the pattern expected from proposition 1. The first is that the difference in genetic diversity between the short and long subset was not pronounced enough to have different QTL effects models performing differently. A second reason is that increased genetic distance between parents will not automatically translate into increased genetic variance (Hung et al., 2012). Third, the sample sizes may have been too small to show the expected patterns with not enough QTLs being detected and insufficient power to distinguish between QTL effect models and between homogeneous and heterogeneous residual genetic variance.

An important result was that in three scenarios out of six (short-DMY, heterogeneous-DMY, and long-PH) the bi-allelic model gave the largest pVS. From a general point of view, more parsimonious models, especially the bi-allelic model gave results with a reduced bias (difference between pTS and pVS). The improvement of QTL detection in MPP by using more parsimonious models based on shared polymorphism has been an important topic in MPP QTL analysis. While simulation studies have supported this idea (Rebaï & Goffinet, 1993; Jansen et al., 2003; Leroux et al., 2014), only few real data analyses have confirmed it (Blanc et al., 2006; Bardol et al., 2013). Of course, the superiority of the bi-allelic model depends on the sample size. With larger sample sizes the bi-allelic model will be more often inferior to more complex QTL effect models.

Other real data analyses did not support the idea that a power gain could be achieved by integrating relatedness between crosses or parents into the analysis (Li et al., 2005; Coles et al., 2010; Steinhoff et al., 2011, 2012; Liu et al., 2012; Würschum et al., 2012; Giraud et al., 2014). In all these studies, the reference cross or parent specific model yielded the best results. For the full subset analysis of PH (Table 3.2) we also noticed that the parental model explained a higher proportion of the genetic variation. Cross or parent specificity of the QTL effects seems therefore to be important in MPP QTL mapping. Several factors can explain the presence of complex allelic series in an MPP context: 1) multiple alleles; 2) different allele frequencies per cross; 3) cross-specific difference of linkage disequilibrium between markers and QTLs; 4) cross-specific dominance ratio; 5) and/or interaction with the genetic background (Steinhoff et al., 2012; Blanc et al., 2006). For example, the simulation study of Li et al. (2016), demonstrated that the parental model outperformed the bi-allelic model only in cases of strong interaction between the QTL and the genetic background.

Many of these studies did, however, not use CV to validate their results (e.g. Coles et al.

(2010)). In the paper of Han et al. (2016), CV was used to evaluate the model performance and no differences were found in terms of bias between the different tested models. Other authors using CV like Liu et al. (2012) and Würschum et al. (2012) did find a reduced bias for the bi-allelic model with respect to the cross-specific model, as we did. For the reduction of bias by more parsimonious models statistical and biological arguments can be formulated.

From a statistical perspective, the inclusion of multiple QTL effects like in the parental model will make the procedure susceptible to overfitting (Friedman et al., 2001). Therefore, parental effects may model variation that is specific to the TS but that will not necessarily be typical for the VS. Genetically, it has been shown that an important part of the polymorphic variation in maize was cross or even genotype specific (Myles et al., 2009). In contrast, the bi-allelic model contains just two parameters and assumes that these are present across the whole MPP, and so will be shared between TS and VS.

From a biological perspective, we interpret the parental model as being based on more recent sources of relatedness, whereas the bi-allelic model is closer to the original mutation (Powell et al., 2010). Since SNPs represent older polymorphisms, they tend to be better distributed across the whole population (Nicholson et al., 2002; Speed & Balding, 2015). Therefore, QTLs detected with the bi-allelic model may be better spread throughout the whole population and be more easily transmitted to the next generation.

3.4.3 Multi-QTL Effect (MQE) model

As mentioned in the result section, the MQE model performed better in terms of pTS than the models assuming a single type of QTL effect along the genome (Figure 3.3). The larger proportion of genetic variance explained in the TS by the MQE model in comparison to the VS may be explained by the greedy forward regression strategy used to build the MQE model. Indeed, this strategy includes many genome scans, because at each QTL detection step a genome-wide scan is performed for each type of QTL effect conditional on all earlier identified QTLs. The increased number of scans can lead to an overfit of the data present in the TS, while the variation modelled in the TS is not necessarily typical of the VS, as we noticed for the trait PH. However, for DMY, the MQE method did obtain larger pVS, which supports our second proposition that the inclusion of different types of QTL effects in the same model can lead to a better description of the phenotypic variation.

The MQE model seems therefore to be a useful strategy to model phenotypic variations in MPPs. The MQE model aims at finding the most adequate type of QTL effect for each QTL position. In philosophy, it is similar to the Bayesian approach proposed by Jannink & Wu (2003) who treated the number of alleles at a QTL position as a random parameter. The MQE model is probably computationally less demanding. It may require improvements on the correction for multiple testing to avoid overfitting by taking into account the number of scans performed. Alternatively, it may be better to take the maximum observed threshold across the three threshold values corresponding to the three types of QTL effects in place of the mean threshold, as we did now.

3.4.4 Model performance under different assumptions for the residual term

Our proposition 3 implied that heterogeneity of genetic distance between the central and the peripheral parent would require an elaborate model to accommodate the heterogeneity of variance for the polygenic effects. Cross-specific residual terms should give an improved description of the residual genetic variance against which to test for the QTL effects in comparison to a simpler model based on a single residual variance component (supplemental file S5). We expected the difference between the HRT and CSRT model to be largest in the heterogeneous subset. The CV results (Figure 3.3) did not show any difference between the HRT and CSRT model, maybe due to too small population sizes. In the full subset analysis (Table 3.2) we could however notice that in several cases, the CSRT model outperformed the HRT model. For example, the MQE CSRT model explained a larger proportion of genetic variance than the MQE HRT model in five scenarios out of six.

The absence of difference between the HRT and the CSRT model for the CV results can be caused by the use of the f-CSRT approximation . Alternatively, the sample sizes for the crosses may have been too small to allow HRT and CSRT to be tested as being different. The f-CSRT seems to have less QTL detection power than the exact solution because the correction for heterogeneous residual term only reflects the general level of heterogeneity. In the exact solution, however, the residual genetic variance is calculated at each genomic position conditional on the estimated QTL effects. In that case, the Wald statistics can truly benefit from the within cross variance reductions following from the included QTL effects. Therefore, given the full subsets results (Table 3.2), we still consider that the CSRT model can improve QTL detection in MPP.

3.5 Conclusions

The result presented in this study illustrate the potential of using different types of QTL effect models, that also represent different way to model genetic relatedness in an MPP. We demonstrated that it can be interesting to integrate multiple assumptions about the origin of the QTLs in the same multi-locus QTL model. The question of genetic relatedness definition is one of the most important ones in genetics (Fisher, 1918). We know that statistical dependence between haplotypes is the result of a complex evolutionary/selection process where mutation, recombination and coalescence of lineage act jointly (Rosenberg & Nordborg, 2002). We think that the use of relatedness and shared polymorphisms, through a better modelling of genetic relationships, can improve QTL detection in MPPs.

A first way to improve the estimation of genetic relatedness modelling is by using SNP markers in place of pedigree information. According to Powell et al. (2010) this represents a more unified way to measure relatedness and allows to solve the apparent conflict between IBD and IBS methods because when marker density is high the different categories of ancestors merge. Genetic relationship matrices (GRMs) have been widely used in GWAS analysis to control for population structure and/or model polygenic effect within mixed models (Yu et al., 2006; Malosetti et al., 2007). This technique was also employed with success to estimate genetic effects (Yang et al., 2010; Speed et al., 2012). The extension of such a methodology to MPP QTL detection represents a promising option.

A second, more challenging option is to use methods for IBD computation using ancestral lines higher up in the MPP pedigree as a reference like in Zheng et al. (2015). Finally, Bayesian methods have also great potential to deal with complex pedigrees. In this framework efforts have been made to model more appropriately the relationships between lines. For example, Jannink & Wu (2003) proposed to treat the number of ancestral alleles as a random parameter in an attempt to estimate the most probable relatedness scheme between MPP parents. In the same vein, Ter Braak et al. (2010) developed an algorithm to infer latent ancestral class origins of population's founders allowing to sample parent origin in a Bayesian context.

From a general point of view, we would like to emphasize the main motivation for our QTL mapping approach: try to make as explicit as possible the connection between the biological assumptions and the properties of the statistical model that is used. A constant dialogue between these two dimensions is certainly a promising way to make progress in both the understanding of biological processes occurring in MPPs and their statistical modelling.

3.6 Supplementary Material

S1: Nei and Li genotype similarity coefficient

$GS = 2N_{ij}/(N_i + N_j)$ where N_{ij} is the number of common sites between i and j and $N_{i(j)}$ is the total number of observed sites in individual $i(j)$ (Nei & Li, 1979). In our situation, considering the same number of SNPs in parent i and j makes GS equivalent to the simple matching coefficient (SM).

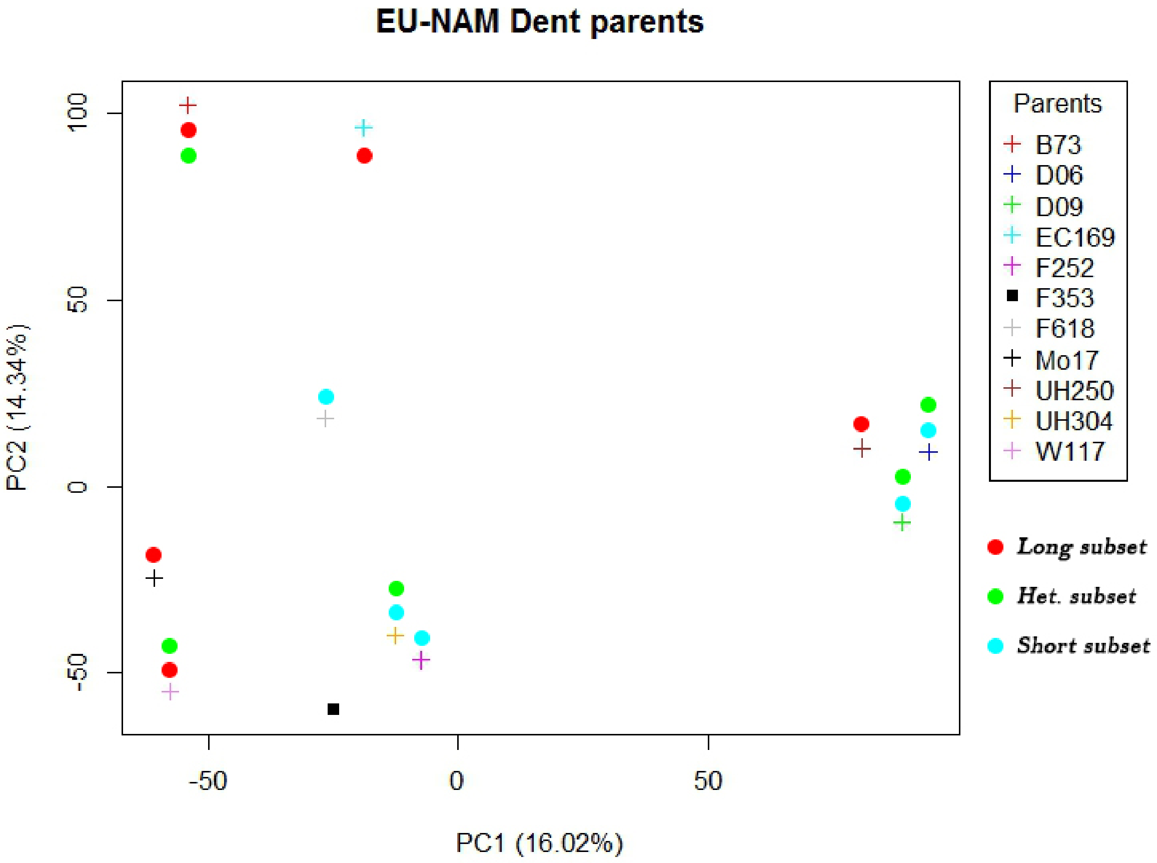
S2: Simple matching coefficients table between parents of the EU-NAM Dent panel

	B73	D06	D09	EC169	F252	F353	F618	Mo17	UH250	UH304	W117
B73	1	0.566	0.556	0.685	0.561	0.557	0.59	0.554	0.579	0.584	0.542
D06		1	0.866	0.643	0.611	0.586	0.599	0.559	0.836	0.624	0.568
D09			1	0.585	0.628	0.618	0.589	0.559	0.762	0.648	0.564
EC169				1	0.556	0.558	0.589	0.558	0.613	0.591	0.538
F252					1	0.633	0.578	0.576	0.609	0.648	0.579
F353						1	0.589	0.567	0.575	0.761	0.565
F618							1	0.552	0.595	0.606	0.559
Mo17								1	0.578	0.586	0.564
UH250									1	0.614	0.563
UH304										1	0.587
W117											1

Average genetic similarity score per subset:

$$\bar{SM}_{short} = 0.639$$
$$\bar{SM}_{het.} = 0.613$$
$$\bar{SM}_{long} = 0.573$$

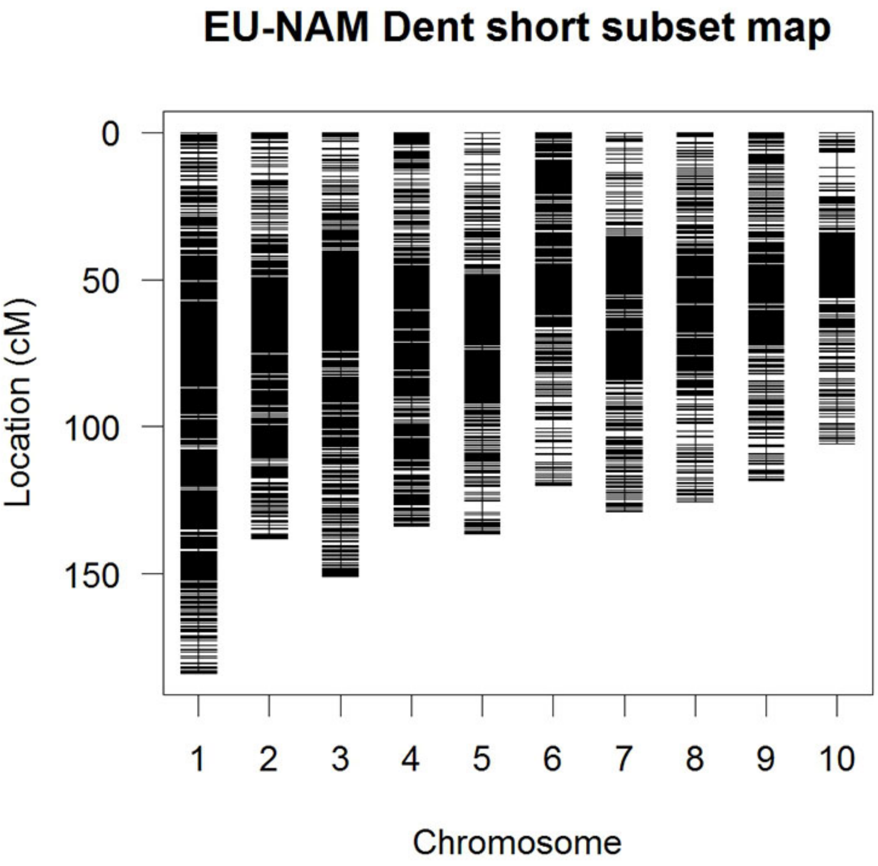
S3: Principal component bi-plot of the EU-NAM Dent parents



S4: Genetic marker map of the different subsets

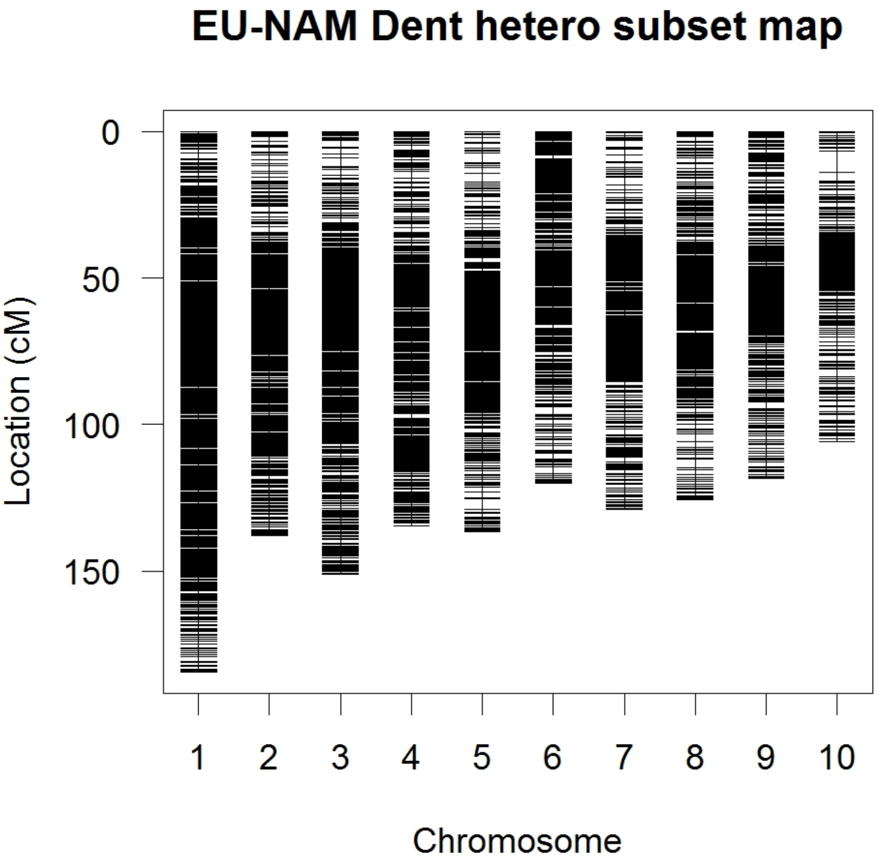
Short subset

Chromosome	N	Length(cM)	Average spacing	maximum spacing
1	985	183.8	0.2	1.8
2	633	137.9	0.2	2.0
3	702	150.9	0.2	2.4
4	617	133.9	0.2	2.2
5	563	136.5	0.2	3.8
6	482	119.9	0.2	3.1
7	470	128.9	0.3	3.3
8	496	125.6	0.3	2.6
9	449	118.4	0.3	2.4
10	340	105.8	0.3	5.3
Overal	5737	1341.6	0.2	5.3



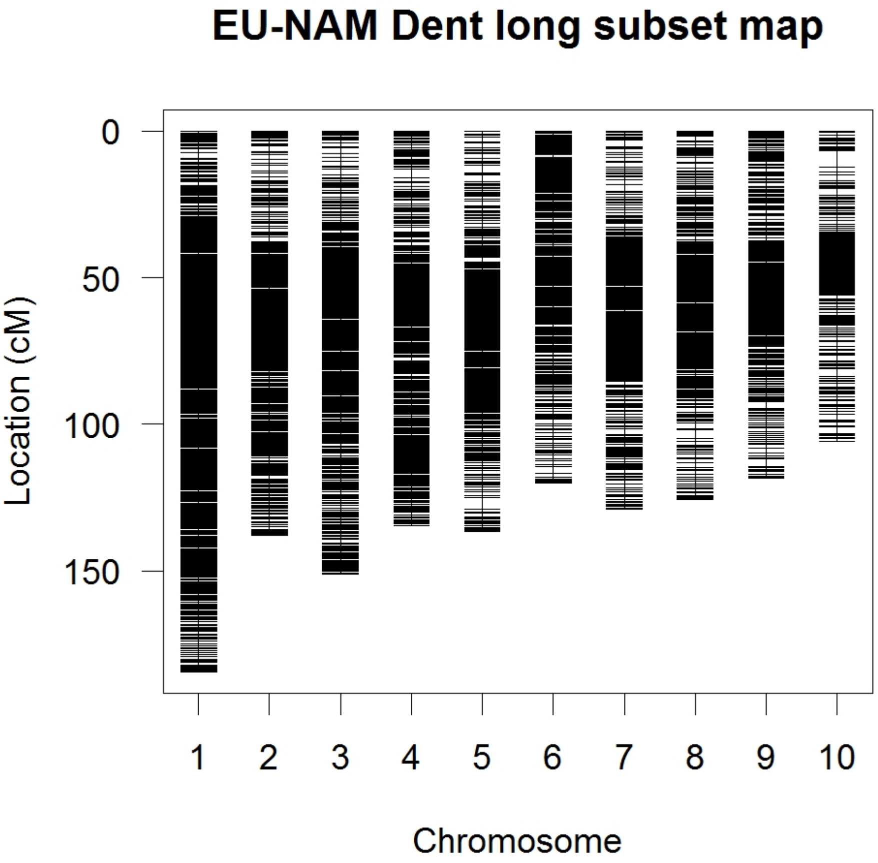
Heterogeneous subset

Chromosome	N	Length(cM)	Average spacing	maximum spacing
1	1055	184.5	0.2	1.9
2	630	137.9	0.2	2.1
3	711	151.0	0.2	2.9
4	620	134.6	0.2	2.2
5	573	136.6	0.2	3.8
6	526	119.9	0.2	2.6
7	493	128.9	0.3	2.6
8	527	125.6	0.2	2.9
9	467	118.4	0.3	2.4
10	332	105.8	0.3	7.3
Overal	5934	1343.2	0.2	7.3



Chromosome	N	Length(cM)	Average spacing	maximum spacing
1	1121	184.5	0.2	1.9
2	641	137.9	0.2	2.0
3	743	151.0	0.2	1.9
4	647	134.6	0.2	3.1
5	598	136.6	0.2	4.1
6	550	119.9	0.2	2.1
7	524	128.9	0.2	2.2
8	560	125.6	0.2	2.9
9	484	118.4	0.2	2.7
10	344	105.8	0.3	5.7
Overal	6212	1343.2	0.2	5.7

Long subset



S5: Test statistic of the QTL effect

Wald test derivation

The significance of the estimated QTL effects $\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}$ can be estimated using the Wald test (Wald, 1943). From model (2), the phenotype values $\boldsymbol{y}$ have an expectation equivalent to $\boldsymbol{X\beta}$ and their variance is $\boldsymbol{R}$ (McCulloch & Searle, 2001, 6.5, 6.6).

In such a situation, we can derive a generalized estimate for $\boldsymbol{\beta}$ and its variance as follow (Rao et al., 2008, 4.65, 4.66):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'R}^{-1}\boldsymbol{y} \tag{3.3}$$

$$V(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X'R}^{-1}\boldsymbol{X})^{-1} \tag{3.4}$$

In its general form, the Wald statistic is equal to (McCulloch & Searle, 2001, 5.39):

$$W = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{0}})'[V(\hat{\boldsymbol{\beta}})]^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{0}}) \tag{3.5}$$

Under the null hypothesis we assume $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{0}} = \begin{bmatrix}0 & 0 & \dots & 0\end{bmatrix}'$ After substituting (3.3) and (3.4) in (3.5), we can rewrite the Wald statistic like that:

$$\begin{aligned} W &= \boldsymbol{y}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{y} \\ &= \boldsymbol{y}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{y}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{H}\boldsymbol{y} \\ &= \boldsymbol{y}'\hat{\boldsymbol{R}}^{-1}\hat{\boldsymbol{y}} \end{aligned} \tag{3.6}$$

where,

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{R}}^{-1}$$

is the generalized hat matrix.

The previous expression W represent a global Wald test coefficient $W(\hat{\boldsymbol{\beta}})$ including both effect of the cross intercepts $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ and the QTL effects $\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}$. The significance of the QTL effects $W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}})$ can be obtained doing the difference between $W(\hat{\boldsymbol{\beta}})$ and $W(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}})$, the Wald statistic of a model including only the cross intercept terms.

$$W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}) = W(\hat{\boldsymbol{\beta}}) - W(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}) \tag{3.7}$$

$W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}})$ is therefore proportional to $W(\hat{\boldsymbol{\beta}})$ described in (3.6). $W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}})$ tests the hypothesis of all QTL effects equal zero versus at least on component of $\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}$ being non null. $W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}})$ follows a χ^2 distribution with degree of freedom (df) equal to the rank of $\boldsymbol{X_Q}$ (the number of estimated QTL effects).

Interpretation

Expression (3.6) allows to summarise the main features of the proposed QTL models: parsimony versus goodness of fit and accurate form of the residual term.

Parsimony vs goodness of fit Since the df of the Wald test depends on the number of estimated parameters, more parsimonious models like the ancestral or the bi-allelic model will automatically increase the significance level of $W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}})$. The use of more parsimonious models is however not always a guaranty of better results (Bardol et al., 2013; Steinhoff et al., 2011). An important criteria to be balanced with parsimony is the necessity to infer allele effects that capture correctly the trait variability. To illustrate this, we can further reduce expression $W(\hat{\boldsymbol{\beta}})$ (3.6) and substitute it in (3.7) to draw the following relationship:

$$W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}) \propto \boldsymbol{y}'\hat{\boldsymbol{y}} = \sum_{n=1}^{N} y_i \hat{y}_n \tag{3.8}$$

From this expression we can see that the more the vector $\boldsymbol{y}$ and the vector $\hat{\boldsymbol{y}}$ vary in the same direction the higher will be $W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}})$. This simply means that the form of the QTL incidence matrix $\boldsymbol{X_Q}$ should be chosen to give genetic estimates $\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}$ allowing a projection $\hat{\boldsymbol{y}} = \boldsymbol{Hy}$ capturing the highest proportion of the trait variability. If these variations are due to parental or cross-specific effects, corresponding genetic effect estimates $\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}$ capturing these local variation should perform better at the price of a higher number of parameter to estimate. On the other hand if the effects are similar through the MPP, a reduced number of parameters will capture this variability and allow gains in power by a lower number of df.

Accurate form of the residual term The test statistic is also influenced by the chosen VCOV. As we can see in expression (3.6), each element composing $W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}})$ is weighted by the estimated $\hat{\boldsymbol{R}}$. The more $\hat{\boldsymbol{R}}$ will reflect the correct form of the residual term the more accurate will be the QTL detection process. In the HRT case, substituting (3.6) in (3.7), we can write the following relationship:

$$W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}) \propto \sum_{n=1}^{N} \frac{y_n \hat{y}_n}{\sigma_r^2} \tag{3.9}$$

In this situation each elements is weighted by an average level of uncertainty σ_r^2 which may not be representative of crosses particularities. In the CSRT situation, we have:

$$W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}) \propto \sum_{c=1}^{n_c} \sum_{n=1}^{N_c} \frac{y_n \hat{y}_n}{\sigma_{r_c}^2} \tag{3.10}$$

Here, the different elements will be weighted by cross-specific variance residual terms $\sigma_{r_c}^2$ that take into account the potential differences of uncertainty between crosses. The more $\sigma_{r_c}^2$ are heterogeneous the more a CSRT model is needed to handle this variability. In such case, a HRT model will tend to be more liberal when uncertainty is in fact larger (cross 1 in Table 3.3) and more conservative when effects are in reality more certain (cross 3 in Table 3.3).

Table 3.3: Illustration of the difference between the HRT and CSRT assumption in an heterogeneous MPP and its effect on the QTL test statistic ($\frac{\beta}{\sigma^2}$).

	$\sigma_{r_c}^2$ ("True")	σ_r^2 ("Average")	Test (HRT)		Test (CSRT)
cross 1	190		$\frac{\beta_1}{100}$	$>$	$\frac{\beta_1}{190}$
cross 2	100	100	$\frac{\beta_2}{100}$	$=$	$\frac{\beta_2}{100}$
cross 3	10		$\frac{\beta_3}{100}$	$<$	$\frac{\beta_3}{10}$

S6: CSRT model approximation

The idea is to first estimate the variance covariance structure $\hat{\boldsymbol{R}}$ and then use it in a generalized estimate of the Wald test (3.6):

$$W(\hat{\boldsymbol{\beta}}) = \boldsymbol{y'}\hat{\boldsymbol{R}}^{-1}\boldsymbol{X}(\boldsymbol{X'}\hat{\boldsymbol{R}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'}\hat{\boldsymbol{R}}^{-1}\boldsymbol{y} \tag{3.11}$$

Since in the CSRT model $\hat{\boldsymbol{R}}$ contains only diagonal elements we can simply invert $\hat{\boldsymbol{R}}$ by doing $\hat{\boldsymbol{R}}^{-1} = 1/\hat{\boldsymbol{R}}$.

SIM

Cr(Int) model: $$\boldsymbol{y} = \boldsymbol{X_c\beta_c} + \boldsymbol{r} \tag{3.12}$$

Cr(Int) + Q model: $$\boldsymbol{y} = \boldsymbol{X_c\beta_c} + \boldsymbol{X_Q\beta_Q} + \boldsymbol{r} \tag{3.13}$$

To estimate the significance of the QTL effect we can use the following incremental Wald statistics

$$W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}) = W([\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}|\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}]) - W(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}) \tag{3.14}$$

The procedure to estimate the significance of the QTL effect genome-wide is the following:

1. Estimate $\hat{\boldsymbol{R}}$ from model (3.12).
2. Compute $W(\hat{\boldsymbol{\beta}}_{\boldsymbol{c}})$ substituting $\hat{\boldsymbol{R}}$ and $\boldsymbol{X} = \boldsymbol{X_c}$ in (3.11).
3. Compute at each position $W([\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}|\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}])$ substituting $\hat{\boldsymbol{R}}$ and $\boldsymbol{X} = [\boldsymbol{X_c}|\boldsymbol{X_Q}]$ in (3.11).
4. Compute the p-value of the QTL effect using (3.14) and $W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}) \sim \chi^2_{df}$ with $df = Rank(\boldsymbol{X_Q})$.

CIM

Cr(Int) + cof model: $$\boldsymbol{y} = \boldsymbol{X_c\beta_c} + \boldsymbol{X_{cof}\beta_{cof}} + \boldsymbol{r} \tag{3.15}$$

Cr(Int) + cof + Q model: $$\boldsymbol{y} = \boldsymbol{X_c\beta_c} + \boldsymbol{X_{cof}\beta_{cof}} + \boldsymbol{X_Q\beta_Q} + \boldsymbol{r} \tag{3.16}$$

To estimate the significance of the QTL effect we can use the following incremental Wald statistics

$$W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}) = W([\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}|\hat{\boldsymbol{\beta}}_{\boldsymbol{cof}}|\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}]) - W([\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}|\hat{\boldsymbol{\beta}}_{\boldsymbol{cof}}]) \tag{3.17}$$

The procedure to estimate the significance of the QTL effect genome-wide is the following:

1. Estimate $\hat{\boldsymbol{R}}_{\boldsymbol{j}}$ for the different combinations of cofactor $\boldsymbol{X}_{\boldsymbol{cof.j}}$ using model (3.15).
2. Compute $W([\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}|\hat{\boldsymbol{\beta}}_{\boldsymbol{cof.j}}])$ for the different combinations of cofactor substituting $\hat{\boldsymbol{R}}_{\boldsymbol{j}}$ and $\boldsymbol{X} = [\boldsymbol{X}_{\boldsymbol{c}}|\boldsymbol{X}_{\boldsymbol{cof.j}}]$ in (3.11).
3. Compute at each position $W([\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}|\hat{\boldsymbol{\beta}}_{\boldsymbol{cof.j}}|\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}])$ substituting $\hat{\boldsymbol{R}}_{\boldsymbol{j}}$ and $\boldsymbol{X} = [\boldsymbol{X}_{\boldsymbol{c}}|\boldsymbol{X}_{\boldsymbol{cof.j}}|\boldsymbol{X}_{\boldsymbol{Q}}]$ in (3.11).
4. Compute the p-value of the QTL effect using (3.17) and $W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}) \sim \chi^2_{df}$ with $df = Rank(\boldsymbol{X}_{\boldsymbol{Q}})$.

S7: Multi QTL effects model

The multi-QTL effect model uses a forward regression to build up a model in which different loci are allowed to have different types of QTL effects. At each step one QTL is added, and each of the different types of effects are compared. To select this position, we compute genome wide profiles using the types of QTL effects given by the user. Each of these profiles uses a single type of effect for all the tested (QTL) position.

$$\boldsymbol{y} = \boldsymbol{X_c\beta_c} + \boldsymbol{X_{Q1}\beta_{Q1}} + \boldsymbol{r} \tag{3.18}$$

Where, the QTL position $\boldsymbol{X_{Q1}\beta_Q}$ is parental, ancestral, or bi-allelic. We calculate therefore as many QTL profile as types of QTL effects chosen by the user.

From each of these profiles, the most significant position based on the -log10(p value) statistic is selected (e.g., $\boldsymbol{X_{Q1.par}}$, $\boldsymbol{X_{Q1.anc}}$, $\boldsymbol{X_{Q1.biall}}$). Note that the selected QTL positions might or not be at the same position. The one that increases the most the R^2_{adj} is selected as QTL. The selected position with its type of QTL effect is added to the cofactors list and the selection process starts again. If at step 1 we selected a bi-allelic QTL, then at step 2 the QTL profiles will be based on the following models:

$$\boldsymbol{y} = \boldsymbol{X_c\beta_c} + \boldsymbol{X_{q1.biall}\beta_{q1}} + \boldsymbol{X_{Q2}\beta_{Q2}} + \boldsymbol{r} \tag{3.19}$$

With again the tested QTL position $\boldsymbol{X_{Q2}\beta_{Q2}}$ taking, in each QTL profile, the form of one of the QTL effect specified by the user.

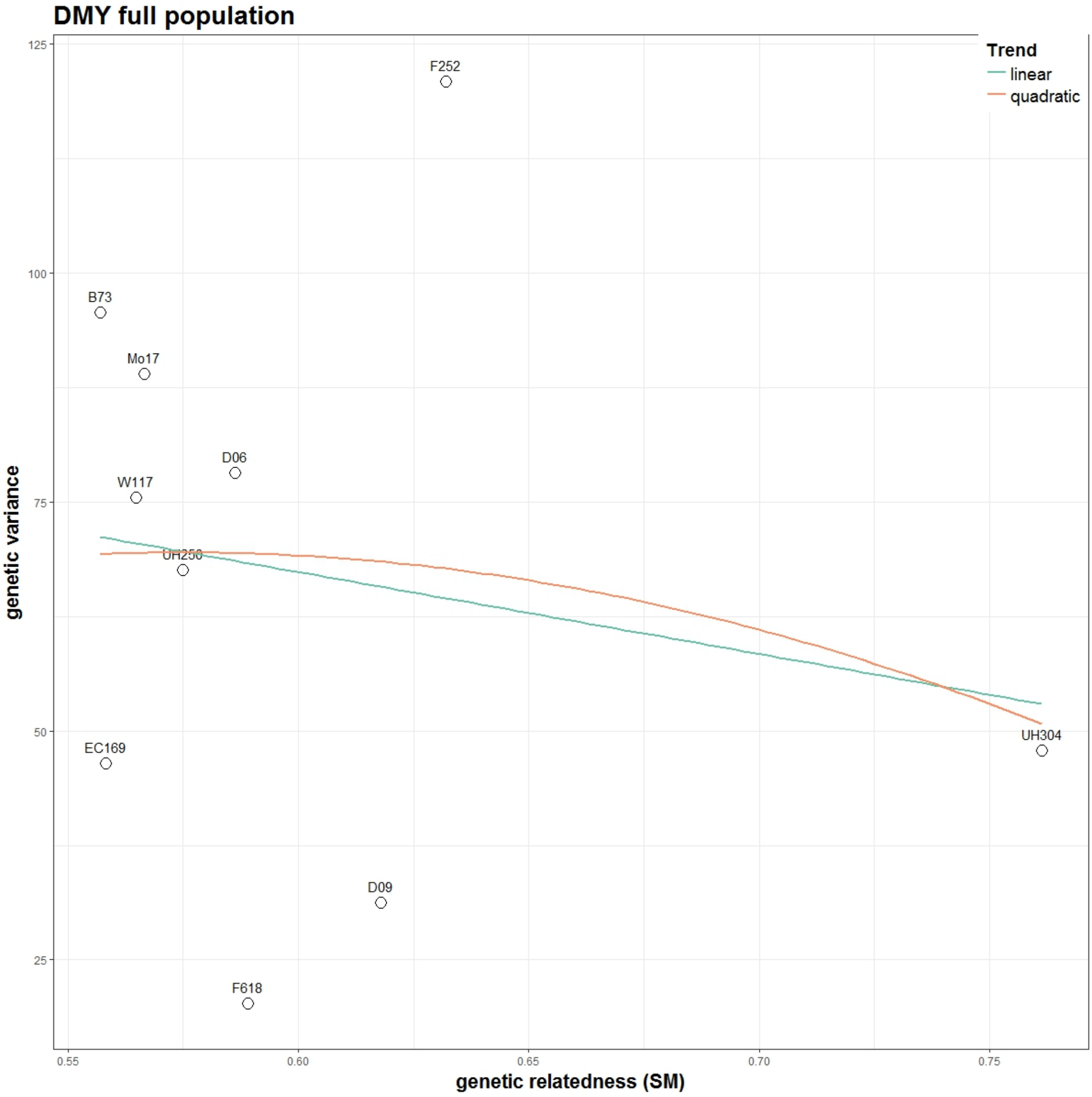
The procedure stops when there is no more significant position. The final list of QTL is tested simultaneously using a backward elimination. The final model could look like that:

$$\boldsymbol{y} = \boldsymbol{X_c\beta_c} + \boldsymbol{X_{q1.biall}\beta_{q1}} + ... + \boldsymbol{X_{q(t-1).par}\beta_{q(t-1)}} + \boldsymbol{X_{Qt.anc}\beta_{Qt}} + \boldsymbol{r} \tag{3.20}$$

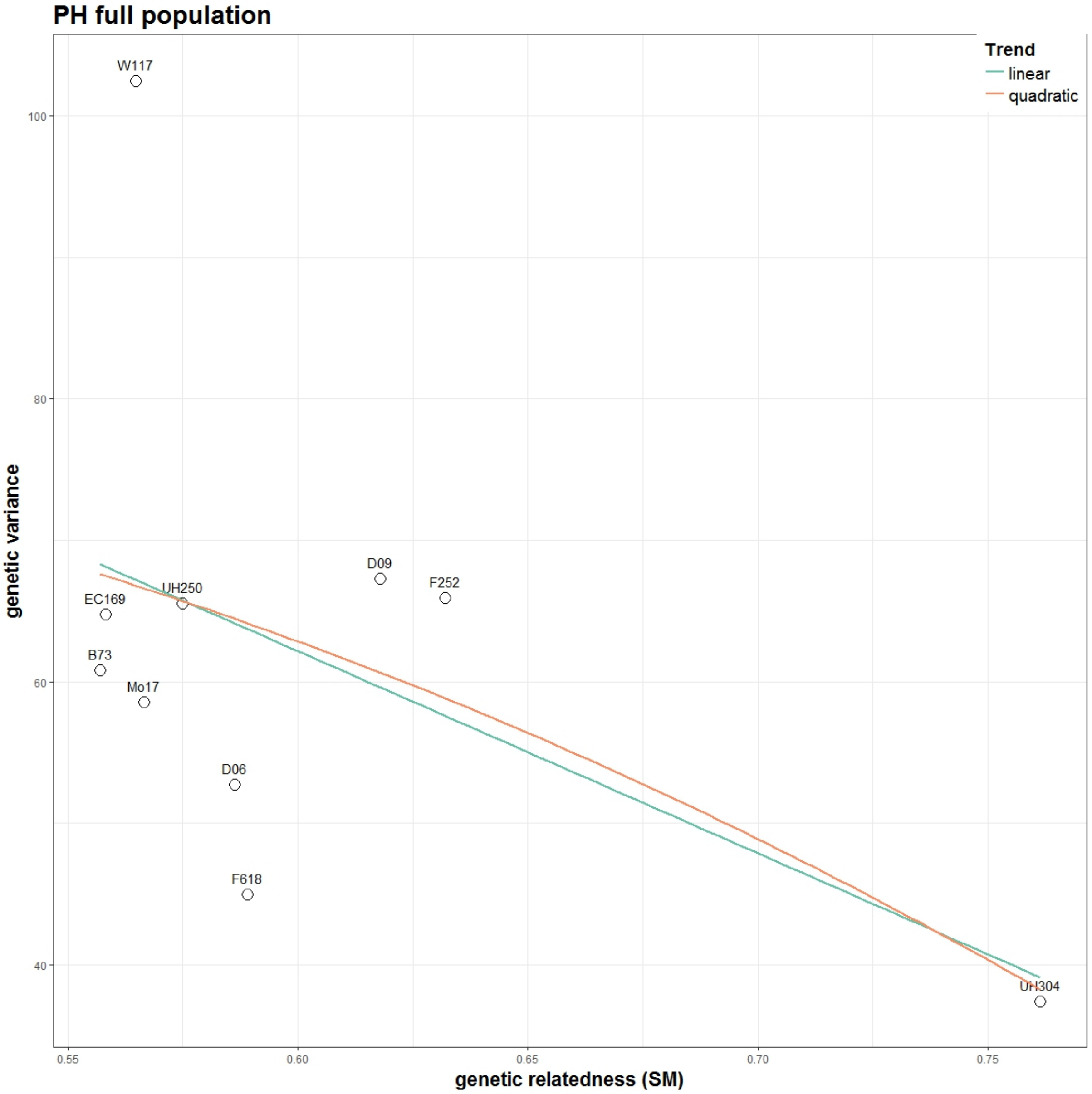
S8: Genetic variance versus parental relatedness plots

The following graphs represent the scatter plots of estimated genetic variance per family on *SM* coefficient between the central parent and the peripheral one (see S1 and S2). The two trends represent the linear and quadratic trend respectively.

Dry matter yield



Plant height



S9: Permutation threshold results

Significance thresholds determined by 1000 genome-wide permutations taking the -log10(p-value) of the upper 95% Wald statistic under the empirical null distribution as the critical value for rejection.

Dry mater yield

	parental	ancestal	bi-allelic	Average - MQE
short	3.89	4.09	4.66	4.21
het.	4.03	4.27	5.01	4.44
long	3.82	3.97	4.56	4.12
Average	3.91	4.11	4.74	

Plant height

	parental	ancestal	bi-allelic	Average - MQE
short	3.86	4.03	4.65	4.18
het.	3.97	4.22	5.18	4.46
long	3.77	4.14	4.85	4.25
Average	3.87	4.13	4.89	

Chapter 4

To improve QTL detection power in multi-parent populations it is better to increase the cross sizes rather than the parent numbers: a simulated study inspired by a sugar beet example

Vincent Garin[1, 2], Valentin Wimmer[3], Dietrich Borchardt[3], Marcos Malosetti[1], Fred van Eeuwijk[1]

1. Biometris, Wageningen University & Research Center

2. C.T. de Wit Graduate School for Production Ecology & Resource Conservation (PE & RC)

3. KWS SAAT SE

Abstract

Background Multi-parent populations (MPPs) are important resources for studying plant genetic architecture and detecting quantitative trait loci (QTLs). In MPPs, the QTL effects can show various levels of allelic diversity, which is an important factor influencing the detection of QTLs. In MPPs, the allelic effects can be more or less specific. They can depend on an ancestor, a parent or the combination of parents in a cross. In this paper, we evaluated the effect of QTL allelic diversity on the ability to detect QTL in MPPs.

Results We simulated: a) cross-specific QTLs; b) parental and ancestral QTLs; and c) bi-allelic QTLs. Inspired by a real application, we tested different MPP designs (diallel, chessboard, factorial, and NAM) derived from five or nine parents to explore the ability to sample genetic diversity and detect QTLs. Using a fixed total population size, the QTL detection power was larger in MPPs with fewer but larger crosses that were derived from a reduced number of parents. The use of a larger set of parents was useful to detect rare alleles with a large effect on the trait. The benefit of using a larger set of parents was however conditioned on an increase of the total population size. We also determined empirical confidence intervals for QTL location to compare the resolution of different designs. For QTLs representing 6% of the phenotypic variation, using 1600 offspring individuals, we found 95% empirical confidence intervals of 25 and 13 cM for the cross-specific and the bi-allelic QTLs, respectively.

Conclusions MPPs derived from less parents with few but large crosses generally increased the QTL detection power. Using a larger set of parents to cover a wider genetic diversity can be useful to detect QTLs with a reduced minor allele frequency when the QTL effect is large and when the total population size is increased.

4.1 Background

The use of multi-parent populations (MPPs) for QTL detection is growing in popularity. With respect to bi-parental crosses, MPPs address a larger genetic diversity. With respect to association panels, MPPs offer more control of the false positive rate due to better information about the pedigree and the population structure (Myles et al., 2009). Here, we focus on MPPs being a collection of crosses between a set of parents with no further intercrossing, unlike the situation in the multi-parent advanced generation inter-cross (MAGIC) populations (Cavanagh et al., 2008). Different statistical procedures have been proposed for QTL detection in MPPs. We performed simulations to evaluate MPP designs and QTL detection models in terms of power, of false positive detection, and of resolution assuming different levels of QTL allelic diversity.

4.1.1 QTL allelic diversity

In MPPs because crosses are derived from multiple parents, more alleles can potentially be segregating with effects that can be more or less diverse/consistent. We define four types of QTL allelic effects from the most diverse and specific effects to the most consistent and shared effects. In an MPP, the QTL effects can be defined in terms of allele origin and/or mode of action. The first QTL allelic effect is called cross-specific and represents an epistatic interaction between a parental allele and a cross genetic background (Blanc et al., 2006). Therefore, a cross-specific QTL has effects that can only be estimated within crosses (nested effects). The other QTL allelic effects are defined in terms of parental alleles, ancestral alleles, and SNP alleles with consistent effects across crosses. The QTL allelic effects can be common because the alleles are unique to: 1) a common parent (parental), 2) a common ancestral line (ancestral), or 3) a common causal SNP (bi-allelic). In the rest of the paper, we will refer to these four types of QTL effects calling them: cross-specific, parental, ancestral, and bi-allelic. Generally, the number of effects that needs to be estimated decreases from cross-specific to bi-allelic QTLs. Therefore, the sample size to estimate the individual alleles or cross-specific QTL effects increases from cross-specific to bi-allelic QTLs. We hypothesize that the QTL allelic diversity has a strong influence on QTL detection. In this paper, we evaluated the influence of the QTL allelic diversity on MPP QTL detection.

4.1.2 MPP design

Many MPP designs have been evaluated through simulation studies (Yu et al., 2008; Verhoeven et al., 2006; Jansen et al., 2003; Muranty, 1996). The nested association mapping (NAM) design is a collection of crosses between a central parent and peripheral lines (Yu et al., 2008). In a diallel design, every one of p parents is crossed with $p - 1$ other parents (Cockerham, 1963). In a factorial design, a set of parents, A, is fully or partially crossed with another set of parents, B (Cockerham, 1963).

To define an MPP design, we can look at the mating, or crossing, scheme, the number of parents, the number of crosses, and the number of individuals per cross. A larger set of parents allows to cover a wider genetic diversity by sampling more alleles (Xu, 1998). For a given amount of resources (fixed total population size), a larger set of parents implies more crosses and therefore reduces the number of individuals per cross. When more alleles are segregating, each individual allele will be present at a lower frequency within the population. Therefore, the number of parents represents a trade-off between the number of sampled alleles or cross-specific QTL effects and the sample size to detect the QTL allelic effects.

The general conclusion in the literature is that MPP designs with a reduced number of large crosses are more powerful (Xu, 1998; Xie et al., 1998; Verhoeven et al., 2006). Some articles attempted to determine an optimum between the number of parents and the number of individuals per cross. For example, in different MPP designs, Liu et al. (2013) found that the optimal number of parents corresponded to individual cross sizes of 100 individuals. Muranty (1996) concluded analytically that in MPPs with a fixed population size, the detection power was only influenced by the number of parents and not by the MPP design itself. She showed however that the power reached a plateau when more than six parents were involved. The trade-off between the covered genetic diversity and the sample size to detect the QTL allelic effects is a question that deserves further investigation.

4.1.3 Statistical models

The statistical model used for the QTL detection should take into account the MPP design and how to model the QTL effect. Xie et al. (1998), Xu (1998) and Verhoeven et al. (2006) assumed cross-specific QTL effects, which from a statistical point of view represents the saturated model. However, Muranty (1996) and Leroux et al. (2014) took into consideration the connection between the crosses due to common parents using consistent parental QTL effects, which implies a more parsimonious model for the QTL. Jansen et al. (2003) and Klasen et al. (2012) estimated multi-allelic QTL effects with the number of alleles being between two and the number of parents. Liu et al. (2013) used bi-allelic QTL models similar to the ones used in association studies. The Bayesian approach proposed by Wu & Jannink (2004) is an elegant solution because it allows to estimate for each QTL the number of alleles and the global QTL variance. However, it also suffers from limitations because, according to the authors, when using 600 individuals, it is only possible to distinguish effectively five alleles. The computation in a Bayesian approach can also be intensive.

We implemented QTL detection models assuming the different cross-specific, parental, ancestral and bi-allelic QTL allelic effects defined previously. The main objective of this study was to evaluate the influence of QTL allelic diversity on the detection of QTLs in MPPs. Therefore, we performed simulations using QTL genetic models with different levels of QTL allelic diversity. We evaluated the performance of our models on four common MPP designs (diallel, chessboard, factorial, and NAM) with five and nine parents to explore the ability to sample genetic diversity for QTL detection. We conclude with giving recommendations concerning the cross size and the situations where a larger set of parents is useful.

4.2 Methods

4.2.1 MPP design

We evaluated the detection of QTLs on four different MPP designs composed of F2 crosses : diallel, chessboard, factorial, and NAM (Figure 4.1). Given a fixed total population size, these designs represent different strategies to sample the QTL allelic diversity. The diallel design maximises the number of interactions between alleles and genetic background that can be estimated but limits the number of individuals per cross. The chessboard design is a compromise that samples a reduced number of allele by background interactions but allows more individuals per cross. The factorial design can be useful to cross two contrasting sets of parents (e.g. donors and recipients). Finally, the NAM design can be used to explore allelic diversity with respect to a reference parent (McMullen et al., 2009). In our case, the NAM design had the largest number of individuals per cross.

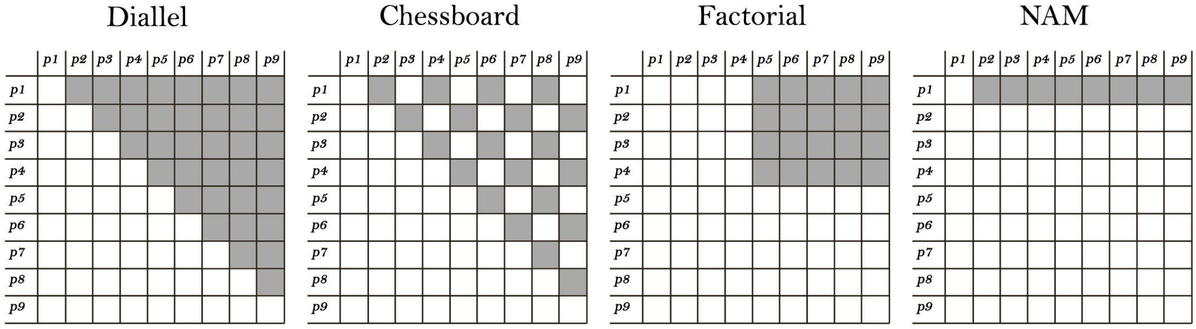


Figure 4.1: Illustration of the MPP designs used for simulation.

We simulated the MPP designs using the genotypic data from nine parents coming from the sugar beet breeding program of the company KWS SAAT SE. Six parents were almost fully inbred with less than 1% of heterozygous markers and three were partially inbred with around 18% of heterozygous markers. We simulated a reference full diallel population composed of the 36 possible F2 crosses between the nine parents. Each F2 cross consisted of 450 genotypes, with recombination and meiosis simulated using a random Poisson process based on the genetic map.

For each *in silico* QTL mapping experiment, we simulated the QTL effects on the reference population and sampled the genotypes from that population to form realizations of the tested MPP designs. We fixed the total population size to 800 or 1600 and "crossed" either five or nine parents. Depending on the crossing scheme, the number of crosses varied from 4 to 36 and the number of individuals per cross from 22 to 400 (Table 4.1). We used 5000 markers spread across nine chromosomes.

The use of an existing set of parents provided a realistic basis to our simulations in terms of genetic properties. The mating scheme and the generation of offspring populations and phenotypes were, however, simulated to give us the flexibility to cover various MPP designs and QTL genetic models. Therefore, although our simulation seems restricted by the particular set of parents, by the type of population (F2), and by the crop used (sugar beet), we believe that our conclusions could apply to other crops and to other type of populations (double haploid, recombinant inbred lines, etc.).

Table 4.1: MPP designs properties with the number of parents (N par), the number of crosses (N cr) and the number of individuals per crosses (N ind/cr).

MPP design	N par	N cr	N ind/cr (N = 800/1600)
Diallel	5	10	80/160
Diallel	9	36	22/44
Chessboard	5	6	133/266
Chessboard	9	20	40/80
Factorial	5	6	133/266
Factorial	9	20	40/80
NAM	5	4	200/400
NAM	9	8	100/200

4.2.2 QTL genetic model

We simulated seven types of QTLs (Table 4.2 and Supplementary file S1). Q1 and Q2 were cross-specific and had non-zero allelic effects in half and one third of the crosses respectively. Setting some cross-specific effects to zero implies that the QTL does not interact with those backgrounds, which seems a reasonable assumption to make from our experience of analysing real data. Q3 and Q4 were both parental, but while Q3 had a different allelic effect for each parent, Q4 only had a single non-zero parental allelic effect. Q5 and Q6 were ancestral, with ancestry groups determined by clustering the nine parental lines on local genetic similarity in a 10 cM window using the R package clusthaplo (Leroux et al., 2014). On average, we detected 3.9 ancestral alleles along the genome. While Q5 had a different allelic effect for each ancestral group, Q6 only had a single non-zero ancestral allelic effect. For Q4 and Q6, the non-zero allele was randomly assigned to one of the parents (ancestors). Q4 and Q6 were bi-allelic QTLs with a parental and ancestral basis. The last type of QTLs (Q7) were bi-allelic with an effect attached to the minor SNP allele.

Table 4.2: Simulated QTLs effects described by type of allelic effect, segregation, number of alleles or QTL effects, and QTL genetic model.

QTL	Allelic eff.	Segregation	N all. (Q. eff.)	Gen. mod.
Q1	cr. sp.	1/2 of the cr.	18	M1 (M5)
Q2	cr. sp.	1/3 of the cr.	12	M1 (M5)
Q3	parental	all parents	9	M2 (M5)
Q4	parental	1 parent	2	M2 (M5)
Q5	ancestral	all ancestors	4	M3 (M5)
Q6	ancestral	1 ancestor	2	M3 (M5)
Q7	bi-allelic	minor SNP allele	2	M4 (M5)

The non-zero QTL allelic effects were sampled from a uniform distribution, and the signs of the effects randomly assigned with equal probabilities. In each *in silico* QTL mapping experiment, we simulated a genetic model with eight QTLs located on different chromosomes but always leaving one chromosome without QTLs as control (to investigate the false discovery rate). We simulated four QTLs with a small effect and four with a big effect representing 2% and 6% of the phenotypic variation respectively. Therefore, the total genetic contribution was always equal to 32% of the phenotypic variance. The sampled QTL allelic values were standardized to have realized QTL variances equal to the ones defined. The remaining phenotypic variance represented the environmental and plot error and was simulated using a normal random term. For details about the phenotypic value simulation procedure see Supplementary file S2.

We simulated five QTL genetic models. The first four genetic models only used QTLs with a single type of allelic effect. The first model (M1) contained only cross-specific QTLs (Q1 and Q2), the second model (M2) only parental QTLs (Q3 and Q4), the third model (M3) only ancestral QTLs (Q5 and Q6), and the fourth model (M4 only bi-allelic QTLs (Q7). In the last multi-QTL effects (MQE) genetic model (M5), a combination of all previous QTL effects was used. It contained Q1 to Q6 and two times Q7 to have two QTLs of each type. In the rest of the paper, we will refer to the QTL genetic models calling them cross-specific, parental, ancestral, bi-allelic and MQE genetic models or we will refer directly to their types of QTL allelic effects (cross-specific, parental, ancestral, or bi-allelic QTLs).

4.2.3 QTL detection model and procedure

The QTL detection models had the following form:

$$\boldsymbol{y} = \boldsymbol{X_c}\boldsymbol{\beta_c} + \boldsymbol{X_Q}\boldsymbol{\beta_Q} + \boldsymbol{r}$$

where $\boldsymbol{X_c}$ represented a cross-specific intercept. $\boldsymbol{X_Q}$ and $\boldsymbol{\beta_Q}$ represented the QTL incidence matrix and the QTL effects that varied according to the number of QTL alleles or effects estimated (Garin et al., 2017). We evaluated the QTL detection performances of four models assuming: cross-specific, parental, ancestral and bi-allelic QTL allelic effects. The residual term $\boldsymbol{r}$ followed the linear model assumption of normality with an homogeneous variance $\boldsymbol{r} \sim N(0, \boldsymbol{I}\sigma_r^2)$. The QTL detection procedure was composed of a simple interval mapping scan to select cofactors followed by a composite interval mapping scan to build a multi-QTL model. The final list of QTLs was evaluated using a backward elimination. The cofactors were selected with a minimum in between distance of 50 cM to not select to many cofactors and avoid model overfitting. The QTLs were selected with a minimum distance of 30 cM. We fixed the cofactor and QTL detection thresholds to $-log10(p - value) = 4$. The choice of a value for the threshold was based on values determined by permutation in real data analyses. For example in Giraud et al. (2014), the threshold values varied between 3.4 to 5.6 for a type I error of 10%. The QTL detection scans were performed using the R package mppR (Garin et al., 2018).

4.2.4 Evaluation statistics

We evaluated the QTL detection performance calculating the true positive rate (TPR) as the fraction of correctly detected QTLs assuming a maximum distance of 5, 10 and 20 cM between the simulated QTL and the detected position. The TPR at the whole chromosome level (TPR chr) was the fraction of detected QTL on a chromosome with simulated QTL without restriction of distance with respect to the simulated QTL position. We calculated the corresponding false discovery rate (FDR) as the proportion of detected QTLs that were distant from a simulated QTL position by more than 5, 10 and 20 cM. The false discovery rate on the chromosomes without simulated QTL (FDR chr) was the percentage of runs where a QTL was wrongly detected. The TPR and the FDR did not sum to one because they were calculated with different denominators. The TPR denominator was the number of simulated QTLs (8) and the FDR denominator was the total number of detected QTLs. We evaluated the resolution of the QTL detection (dQTL) by measuring the distance between a simulated QTL and the largest significant peak on the chromosome.

We computed ANOVAs with the TPR at 10 cM as response and as explanatory factors: the MPP design (D), the number of parents (N), the QTL detection model (M), the QTL size (Qs), and the QTL effect (Qe) (Model 4.1). We included in the model the two-way interactions between the MPP design, the number of parents, the QTL detection model and the QTL size. The error term was normally distributed $e \sim N(0, \boldsymbol{I}\sigma_e^2)$.

$$\begin{aligned} TPR = & D + N + M + Qs + Qe + \\ & D \times N + D \times M + D \times Qs + \\ & N \times M + N \times Qs + \\ & M \times Qs + e \end{aligned} \tag{4.1}$$

We interpreted the ANOVA results for the TPR calculating least-squares means (LSMs), following a model selection procedure that retained significant interactions and significant main effects as well as non-significant main effects that underlie significant interactions. This allowed us to get predictions averaged over a set of factors to visualize and understand the effects of these factors on the TPR. We computed the LSMs using the R package emmeans (Lenth, 2018).

We performed 50 replications of each *in silico* QTL mapping experiment for a different QTL genetic model. For each replication, we sampled MPPs given the population sizes (N = 800 or N = 1600), the MPP design (diallel, chessboard, factorial and NAM), and the number of parents (five or nine). On each sampled MPP, we performed QTL detection by the cross-specific, parental, ancestral and bi-allelic models. It represented a total of 16,000 calculated QTL profiles with 128,000 simulated QTL positions.

4.3 Results

4.3.1 Global measurements

Table 4.3 contains the average TPR and FDR at 5, 10, and 20 cM, and the average dQTL per population size and per QTL genetic model. The results are averaged over all other factors (MPP design, number of parents, QTL detection model, QTL size, and QTL type). The TPR increased when the tolerance distance to the simulated QTL increased. However, the increase between the TPR at 20 cM and the TPR with no maximal distance to the simulated QTL (TPR chr) was limited.

The FDR on the chromosomes with no simulated QTL varied between 0.1 and 1.7%. On the chromosomes where QTLs were simulated, the FDRs were larger (e.g. at 10 cM around 25% and 20% for the N = 800 and N = 1600 populations respectively). The FDR decreased from the cross-specific and parental QTLs to the ancestral and bi-allelic ones. This tendency was more pronounced in the N = 1600 populations. For example, at 10 cM, the FDR was equal to 28%, 24%, 14% and 13% for the cross-specific, parental, ancestral, and bi-allelic genetic model respectively.

The dQTL followed a trend similar to the FDR and decreased from the cross-specific and parental QTLs to the ancestral and bi-allelic QTLs. For example, in the N = 1600 populations, dQTL decreased from 6.5 cM to 3.9 cM for the cross-specific and bi-allelic QTLs respectively. In Table 4.4, we calculated the 0.9, 0.95 and 0.99 quantile values of the dQTL empirical distribution per QTL size and population size for each QTL genetic model. For an illustration of the dQTL distributions, see Supplementary file S3. Table 4.4 allows to get an estimation of the QTL confidence interval (CI). For example, in a population of N = 800, with a mix of QTL effects (MQE) explaining each 6% of the phenotypic variation, 95% of the QTLs were detected at a distance of 0 to 23 cM from the simulated QTL. In the same situation, for a 2% QTL effect, the empirical 95% CI would be 0 to 35 cM. Finally, we noticed that increasing the population size, had a small effect on the resolution of the cross-specific and parental QTLs. The reduction of dQTL between N = 800 and N = 1600 populations was larger for the ancestral and bi-allelic QTLs.

Table 4.4 also contains the number and the percentage of detected QTLs in the different configurations. As expected, big QTLs represented the majority of detected QTLs, being detected two to three times more than the small QTLs. In general, the TPR, FDR, and dQTL results obtained for the MQE genetic model seemed to be an average of the individual types of genetic model.

Table 4.3: Average TPR FDR and dQTL results. The TPR and FDR are measured at 5, 10 and 20 cM. TPR chr is the TPR with no maximal distance to the true QTL. FDR chr is the average FDR on the chromosome with non simulated QTL. The results are presented per population size (N = 800 and N = 1600) and QTL genetic model (cross-specific, parental, ancestral, bi-allelic, and MQE).

	N = 800					N = 1600				
Genetic model	Cr. sp.	Par.	Anc.	Biall.	MQE	Cr. sp.	Par.	Anc.	Biall.	MQE
TPR (%)($5cM$)	24	16	25	26	23	42	34	48	50	42
FDR (%)($5cM$)	45	47	35	32	40	42	41	26	24	33
TPR (%)($10cM$)	32	22	30	31	29	52	43	56	56	51
FDR (%)($10cM$)	28	28	21	19	24	28	24	14	13	20
TPR (%)($20cM$)	38	27	35	35	34	61	51	61	61	58
FDR (%)($20cM$)	12	11	8	7	10	13	10	6	5	9
TPR chr (%)	41	30	37	37	37	66	56	64	63	62
FDR chr (%)	0.3	0.9	0.2	0.6	0.5	1.7	0.9	0.1	0.6	0.7
dQTL (cM)	7.1	8	6.1	5.3	7.1	6.5	6.8	4.3	3.9	5.4

Table 4.4: 90, 95 and 99 % quantile values in cM from the dQTL distribution per QTL size (small 2% and big 6%), per QTL genetic model (cross-specific, parental, ancestral, bi-allelic, and MQE), and per population size (N = 800 and N = 1600). N det. QTL is the number of detected QTL on which the dQTL distribution was build. % det. QTL is the percentage of detected QTLs in the configuration.

		small QTLs (2%)					big QTLs (6%)				
	Genetic model	Cr. sp.	Par.	Anc.	Biall.	MQE	Cr. sp.	Par.	Anc.	Biall.	MQE
		Cr. sp.	Par.	Anc.	Biall.	MQE	Cr. sp.	Par.	Anc.	Biall.	MQE
	.90	21	24	25	23	27	17	19	15	13	17
N=800	.95	30	31	33	33	35	24	26	21	19	23
	.99	57	55	53	71	55	41	52	39	32	41
	N det. QTL	1461	802	1030	954	1142	3768	3050	3734	3777	3627
	% det. QTL	23	13	16	15	18	59	48	58	59	57
	.90	18	22	17	16	19	17	16	9	8	13
N=1600	.95	27	29	24	23	26	25	23	13	13	19
	.99	48	47	46	40	44	43	39	28	24	34
	N det. QTL	3271	2130	2782	2624	2658	5113	5012	5365	5487	5217
	% det. QTL	51	33	43	41	42	80	78	84	86	82

4.3.2 MPP design

Tables 4.5 and 4.6 contain the F-statistics of ANOVAs describing the effect of the MPP design, the number of parents, the QTL detection model, the QTL size, and the QTL effect on the TPR at 10 cM for the N = 800 and N = 1600 populations, respectively. Looking at the ANOVAs, we observe that the MPP design was mostly influential for the cross-specific QTLs. The MPP design F-statistic of the cross-specific genetic model was equal to 91.9 and 44.8 in the N = 800, and N = 1600 populations respectively. In the other QTL genetic models except the MQE, the maximum MPP design F-statistic was equal to 4.4. In Figure 4.2, we plotted the LSM of TPR over the MPP design per QTL size. In these plots, we observe again that the cross-specific QTLs are the only QTL effects for which the MPP design influence the TPR. For the cross-specific QTLs, the TPR increased from the diallel to the NAM design.

4.3.3 Number of parents

The second studied factor was the number of parents involved in the MPP design. In ANOVA Tables 4.5 and 4.6, we notice that the magnitude of the main effect for the number of parents decreased from the cross-specific to the bi-allelic genetic models. For example, in the N = 800 populations, the F-statistic decreased from 188.6 to 22.4. In the N = 1600 population, the main effect for the number of parents became non-significant for the ancestral and bi-allelic QTLs.

Table 4.5: ANOVA degrees of freedom (df), and F-statistics with significance of the MPP design, the number of parents, the QTL detection model, the QTL size and the QTL effect on TPR per QTL genetic model (cross-specific, parental, ancestral, bi-allelic, and MQE) for the N = 800 populations. Adjusted and cross-validation R2 of the models.

	df	Cr. sp.		Parental		Ancestral		Bi-allelic		MQE	
MPP des	3	91.9	***	2.6		1.4		3.4	*	15.6	***
Nb. par	1	188.6	***	131.8	***	37	***	22.4	***	97	***
Det. model	3	75.5	***	8.9	***	87.4	***	99	***	5	**
QTL size	1	693.3	***	845.1	***	1372.9	***	1167.8	***	824	***
QTL type	$0-3^a$	8.5	**	0		0				10.9	***
MPP des x Nb. par	3	9.7	***	1.8		1.3		0.1		3.4	*
MPP des x Det. model	9	4.5	***	1.2		1.9		1.5		0.5	
MPP des x QTL size	3	2.4		3.1	*	11.2	***	0.8		0.3	
Nb. par x Det. model	3	3.2	*	2.5		13.9	***	13.3	***	2.3	
Nb. par x QTL size	1	5.3	*	22	***	1.1		1.4		3.5	
Det. model x QTL size	3	22	***	6.3	***	17.8	***	17.3	***	0.2	
Residuals	$94-97^a$										
adj. R2 (%)		92		89		93		96		67	
CV R2 (%)		90		85		91		92		65	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

a. The df of the QTL type term is equal to 1, 1, 1, 0, 3 from the Cr. sp. to the MQE genetic model. The df of the residuals change according to the df of the QTL type term.

Table 4.6: ANOVA degrees of freedom (df), and F-statistics with significance of the MPP design, the number of parents, the QTL detection model, the QTL size and the QTL effect on TPR per QTL genetic model (cross-specific, parental, ancestral, bi-allelic, and MQE) for the N = 1600 populations. Adjusted and cross-validation R2 of the models.

	df	Cr. sp.		Parental		Ancestral		Bi-allelic		MQE	
MPP des	3	44.8	***	1		4.4	**	0.8		4.4	**
Nb. par	1	69.6	***	4.5	*	0		1.6		13.2	***
Det. model	3	122.2	***	9.6	***	113	***	125	***	0.3	
QTL size	1	328.3	***	623.8	***	2614.8	***	2217.6	***	952.7	***
QTL type	$0-3^a$	28.2	***	17.1	***	13	***			10.9	***
MPP des x Nb. par	3	19.9	***	0.5		4.5	**	4	*	5.9	***
MPP des x Det. model	9	5.9	***	1		5.5	***	3.5	**	0.3	
MPP des x QTL size	3	4.9	**	0.1		0.4		0.5		1.9	
Nb. par x Det. model	3	8.9	***	1		17.6	***	16.5	***	1.3	
Nb. par x QTL size	1	17.3	***	28.7	***	242.9	***	54.9	***	44.7	***
Det. model x QTL size	3	2.6		8.2	***	22.7	***	20	***	6.8	***
Residuals	$94-97^a$										
adj. R2 (%)		89		85		96		98		69	
CV R2 (%)		85		80		95		95		66	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

a. The df of the QTL type term is equal to 1, 1, 1, 0, 3 from the Cr. sp. to the MQE genetic model. The df of the residuals change according to the df of the QTL type term.

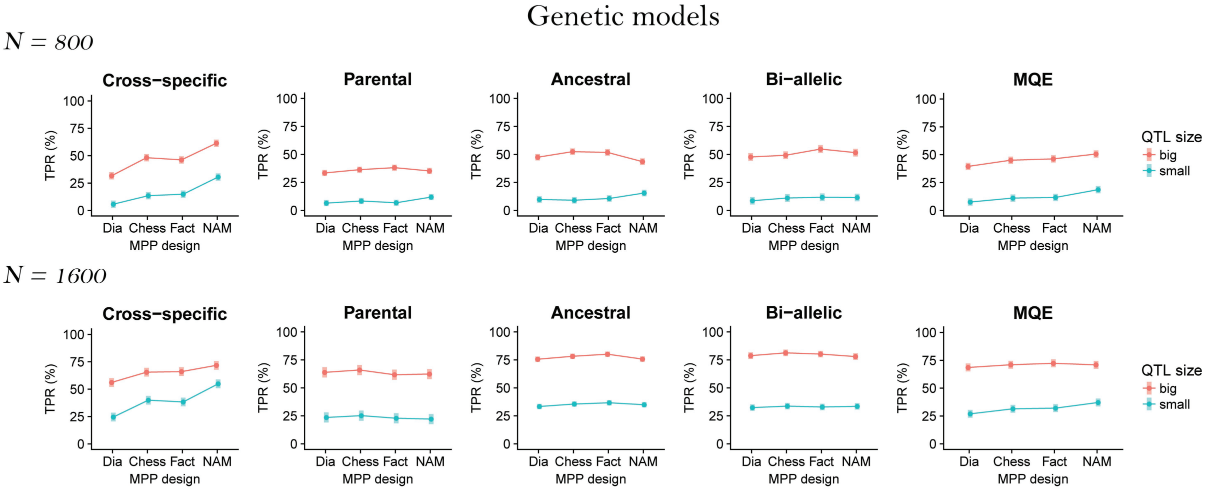


Figure 4.2: TPR LSM over the MPP designs (diallel, chessboard, factorial and NAM) and the QTL sizes (small 2% and big 6%) for all QTL genetic models (cross-specific, parental, ancestral, bi-allelic, and MQE) per population size (N = 800 and N = 1600).

In Figure 4.3, we plotted the TPR LSMs against the number of parents and the QTL size. In general, we observed that the MPP designs with nine parents had a lower TPR. This trend was consistent in the N = 800 populations. However, in the N = 1600 populations, for the big QTLs, the TPR increased from five to nine parents for the parental, ancestral, and bi-allelic QTLs. We investigated in more detail the situations where sampling nine parents increased the TPR. In Figure 4.4, we plotted the TPR LSMs against the number of parents for the big (6%) simulated QTLs (Q1 to Q7) in the N = 1600 populations. For the bi-allelic QTLs (Q7), we split the results in a low and a high minor allele frequency (MAF) category given that the QTL MAF was below or above the median. For the parental, ancestral and bi-allelic effects we noticed that sampling a larger number of parents was more useful for the QTLs with a reduced number of QTL allelic effects and a reduced minor allele frequency (MAF) (Q4, Q6 and Q7 low MAF). The most illustrative examples are the parental QTLs Q3 and Q4. Q3 had 9 parental alleles different from zero where Q4 only had one non-zero parental allele. The TPR of Q3 decreased while the TPR of Q4 increased when we sampled nine parents in place of five.

4.3.4 QTL detection model

The main effect for QTL detection model was highly significant for all QTL genetic models except for the MQE. In the MQE genetic model, the QTL detection model term was not significant for the N = 1600 populations but moderately significant in the N = 800 populations. In Figure 4.5, we plotted the TPR LSMs against the QTL detection model and the QTL sizes. The QTL detection model effect was consistent with the way we simulated the QTL genetic model. The QTL detection model that assumed the corresponding simulation QTL genetic model showed the best performance (e.g. cross-

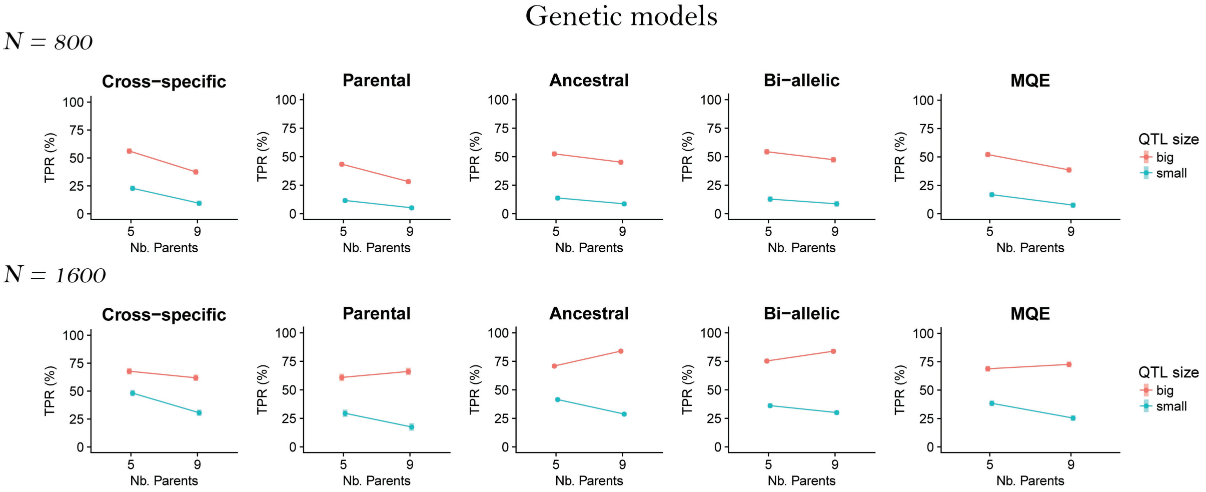


Figure 4.3: TPR LSM over the number of parents (5 and 9) and the QTL size (small 2% and big 6%) for all QTL genetic models (cross-specific, parental, ancestral, bi-allelic, and MQE) per population size (N = 800 and N = 1600).

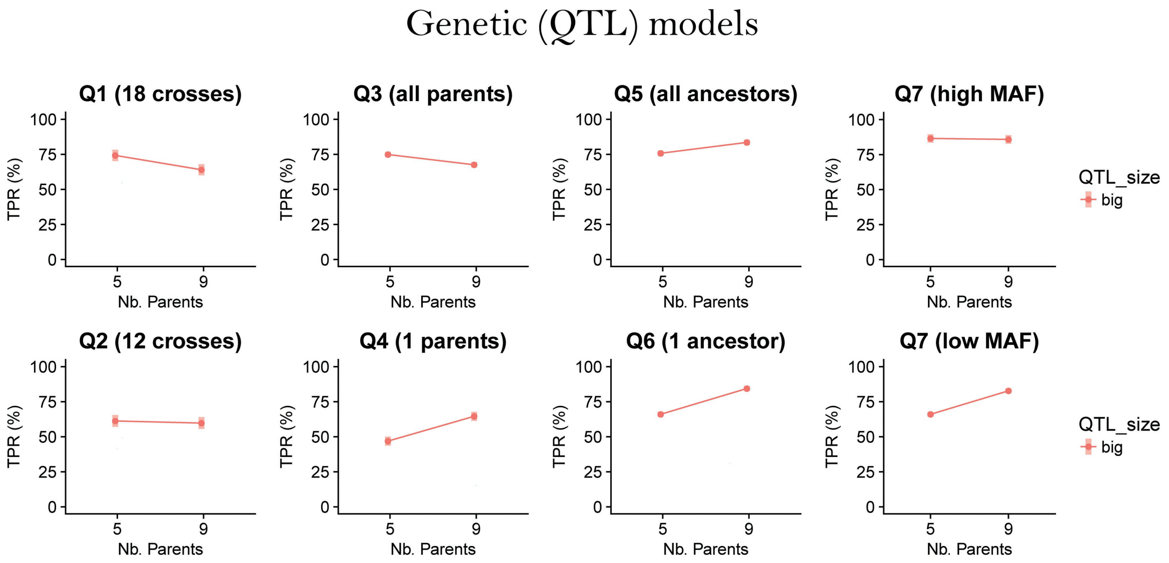


Figure 4.4: TPR LSM over the number of parents (5 and 9) for big (6%) simulated QTLs Q1-7 in populations N = 1600. The bi-allelic QTLs (Q7) were split into a low and a high MAF category given that their MAF was below or above the median.

specific QTLs were best detected with a cross-specific model). This result was especially true for the big QTLs.

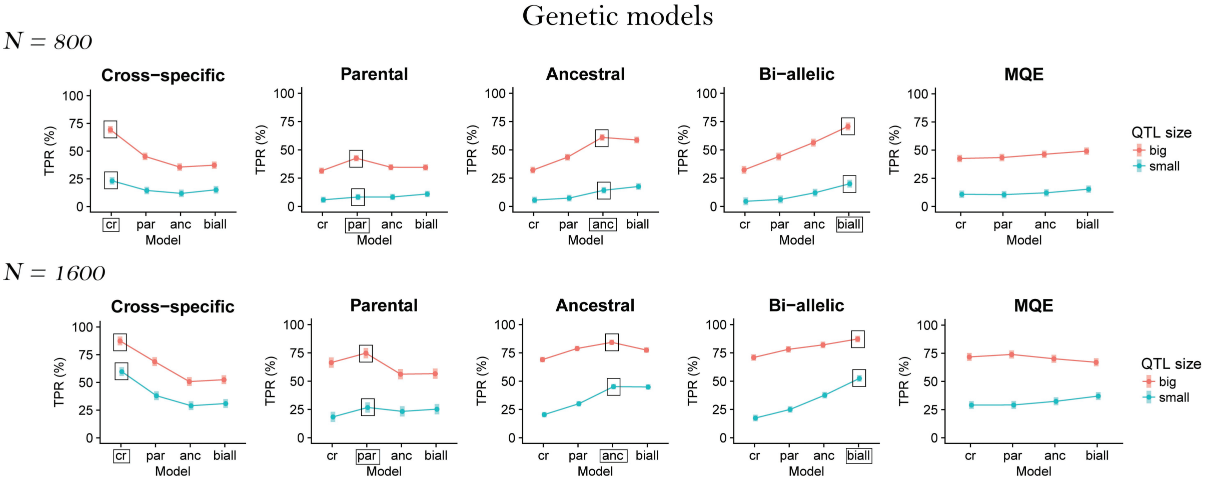


Figure 4.5: TPR LSM over the QTL detection model (cross-specific, parental, ancestral, and bi-allelic) and the QTL size (small 2% and big 6%) for all QTL genetic models (cross-specific, parental, ancestral, bi-allelic, and MQE) per population size (N = 800 and N = 1600). The framed results represent the QTL detection model corresponding to the simulated QTL genetic model.

4.4 Discussion

The main objective of this study was to evaluate the influence of the QTL allelic diversity on MPP QTL detection. We simulated QTLs with different levels of allelic diversity and evaluated the QTL detection power in different MPP designs. The designs were characterized by a specific form (diallel, chessboard, factorial, NAM), and by the use of five and nine parents. Changes in the number of parents used allowed to vary the covered genetic diversity and the number of individuals per cross.

4.4.1 MPP design

According to the ANOVA (Tables 4.5 and 4.6) and the LSM results (Figure 4.2), the MPP design was mostly important for the cross-specific QTLs. For the cross-specific QTLs, MPP designs with a reduced number of large crosses like the NAM performed better than designs with many small crosses like the diallel. Obviously, cross-specific QTL effects need large cross sample sizes to be detected. The recommendation for detecting effects that are potentially diverse and cross-specific would be to choose designs with large crosses rather than sampling many allele genetic background interactions. It seems more important to sample few QTL effects with a large enough sample size than many effects that can not be distinguished due to too small cross sample sizes. From a statistical point of view, the null hypothesis of no QTL will be rejected if a single cross-specific QTL effect or allele is different from zero. The detection of the QTL and the estimation of the allelic effects should therefore be distinguished. For example, an MPP design with few crosses will be efficient to detect the QTL but it will not allow to estimate properly all individual QTL allelic effects.

The MPP design was less important for detecting the other types of QTL allelic effects (parental, ancestral, bi-allelic). Contrary to the cross-specific QTL effects, the parental, ancestral and bi-allelic QTL alleles are consistently defined across crosses which allow them to have an increased sample size. The parental, ancestral and bi-allelic QTL alleles reached more easily the critical sample size for detection which made them less dependent on a particular MPP design. This result is consistent with the conclusion of Muranty (1996) who noticed that the form of the design did not influence the detection of parental and bi-allelic QTLs.

4.4.2 Number of parents

The number of parents used represents a trade-off between the number of sampled alleles and the sample size to detect the QTLs. Similarly, more parents imply better chances for identifying cross-specific QTL effects, but QTL detection power will decrease. For a fixed population size, MPP designs using more parents allows to cover a larger genetic diversity but reduce the number of individuals per cross. Generally, the TPR decreased with a larger set of parents and reduced cross sizes (Figure 4.3). The use of MPP designs with large crosses is therefore an important factor for MPP QTL detection. This result is consistent with several simulations (Xu, 1998; Xie et al., 1998; Wu & Jannink, 2004; Verhoeven et al., 2006) and empirical cross-validation studies (Han et al., 2016).

In MPPs composed of crosses, most of the QTL variance is probably due to the contrast between the parental alleles within the crosses. We used models with cross-specific intercepts to correct for between cross differences that can be both genetic and environmental. We prefer sampling strategies that increase the sizes of the segregating crosses. In our case, five parents probably already allowed to cover a sufficient part of the genetic diversity and resources are better spent on enlarging the sizes of the crosses. This finding is consistent with the conclusions of Muranty (1996) who showed that QTL detection power reaches a plateau after six parents. Therefore, assuming that few parents are already enough to sample genetic diversity, MPP designs should rather invest resources to increase the cross sizes than maximising the covered genetic diversity.

Still, in some situations we did find that using nine parents instead of five increased the TPR. This result was consistent with the conclusions of Muranty (1996), Wu & Jannink (2004), Yu et al. (2008), and Liu et al. (2013). In our case, we underlined that using a larger set of parents was mostly useful to increase the detection power of QTLs with a reduced MAF (Figure 4.4). The bi-allelic QTL Q7 with a low MAF, the parental QTL Q4, which segregated in crosses of a single parent, and the ancestral QTL Q6, which segregated in a single ancestral group, were the only cases where using nine parents improved significantly the TPR.

We hypothesise that when the QTL MAF was reduced (Q4, Q6, and Q7 low MAF), using a larger set of parents was useful to sample at least one cross where the QTL segregated. However, we noticed that the TPR only increased in the large populations (N = 1600) and for the big QTLs. Therefore, the effect of a larger set of parents should be combined with an increased total population size and will be conditioned by the size of the QTL effect. A large population and a big QTL effect will increase the chance of detection when only few individuals carry the QTL.

4.4.3 Mapping resolution

The results concerning the MPP design and the number of parents allowed to distinguish the cross-specific QTLs from the more consistent QTL allelic effects like the parental, ancestral and bi-allelic QTLs. The results obtained in terms of FDR and dQTL also reflected this distinction. In Table 4.3, we noticed that the cross-specific and parental QTLs were detected with larger FDR and dQTL than the ancestral and bi-allelic QTLs. This pattern was mostly observed for the big QTLs (Table 4.4). The increased resolution and reduced FDR observed for the ancestral and bi-allelic QTLs could again be due to an increased sample size to detect the consistent QTL allelic effects.

The FDR on the chromosomes with simulated QTLs decreased when the tolerated distance to the simulated QTL increased. For example, around 60% of the QTLs detected as false positive at 10 cM become true positive if we enlarge the tolerated distance to 20 cM. Many false positive QTL detections could be explained by the extent of the linkage disequilibrium, which is relatively large in F2 populations. The FDR on chromosomes with simulated QTLs indicates that, in MPPs composed of F2 crosses, the confidence region around a detected QTL should be wide to include the true QTL position. For example, in Table 4.4, we noticed that 95% of the 2% QTLs detected in the N = 800 MQE populations, were located between 0 to 35 cM from the simulated QTL. For the 6% QTLs the 95% CI was 0-25 cM. Using a CI of 20 cM around the detected QTLs seems to be a good option. At that level of CI, the FDR was maintained around 10% (Table 4.3). The TPR did not increase much ($2 - 5\%$) when the tolerated distance to the true QTL was relaxed from 20 cM to the whole chromosome (Table 4.3).

4.4.4 Simulation validity

From a general point of view, the low FDR on the chromosome with no simulated QTLs confirmed that our QTL detection procedure functioned properly. Extrapolating the FDR chr to the whole genome gave us values between 0.9 and 15.3% with an average of 5.9%. These values are comparable to the type I error of 10% of the empirical thresholds that supported our choice (Giraud et al., 2014).

We also compared the TPR we obtained with the results of power estimation studies existing in the literature. For example, Beavis (1998) estimated the QTL detection power of interval mapping method in F2 populations with 10 QTLs accounting for a total of 30% of the phenotypic variation. This situation was the closest to our genetic architecture simulation settings. In that case, Beavis (1998) used an empirically determined threshold with type I error of 25% and he obtained powers of 57% and 85% for total population sizes of 500 and 1000 individuals respectively.

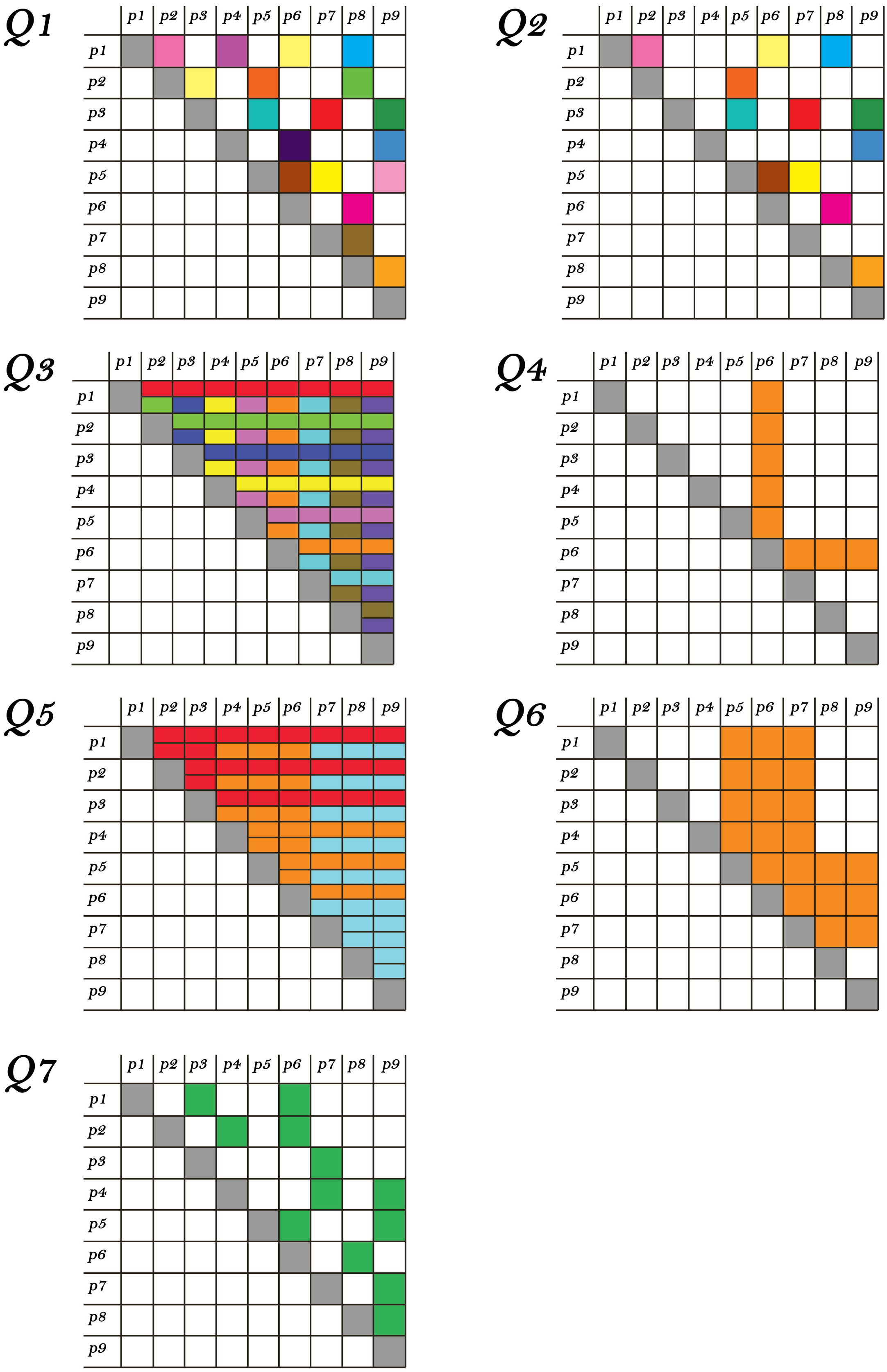
We compared these values to the TPR obtained with the bi-allelic genetic model scenario detected with the bi-allelic model in the N=800 populations because this scenario was the closest to the one used in Beavis (1998). We compared these values to the TPR with no maximal distance to the simulated QTL (whole chromosome interval) obtained in our simple interval mapping scan. We used $-log10(p-value) = 3$ as QTL significance threshold to take into consideration the larger type I error used by Beavis (1998). In that case we obtained a TPR value of 69%. Compared to the interval $57 - 85\%$ obtained for 500 and 1000 individuals, our result are in the same range, which supports the credibility of our simulation.

4.5 Conclusions

We tried to determine the most powerful MPP design to detect QTL in MPPs given various levels of QTL allelic diversity. According to our results, the design of the MPP is only important when the QTL effects are cross-specific. In that case, MPP designs with a reduced number of large crosses will be more powerful. Concerning the number of parents used, our results support the conclusion that sampling a reduced number of parents is already enough to cover a sufficient amount of the genetic diversity. Therefore, resources should be used to increase cross sizes rather than increasing the covered allelic diversity by sampling some extra parents. However, using a larger set of parents can be useful to detect QTLs with a reduced MAF and a big effect if the size of the whole population is also increased.

4.6 Supplementary Material

S1: Illustration of the simulated QTL effects on the reference diallel



S2: Phenotypic values simulation procedure

For each realization, phenotypic values were simulated at the level of the reference diallel population (36 crosses with 450 genotypes). The i^{th} realization included the following steps:

1. The phenotypic variance was expressed in terms of QTL and random error variance $\sigma_p^2 = \sigma_Q^2 + \sigma_e^2$. We assumed a strict additivity of these components.

2. We randomly sampled the 8 QTL positions $(q_1, ..., q_8)$. Each QTL was on a different chromosome. We assumed that the QTL positions were independent and that the global QTL variance (σ_Q^2) was the sum of each individual QTL variance contribution $(\sigma_Q^2 = \sum_{i=1}^{n_{QTL}} \sigma_{qi}^2)$. We calculated the individual QTL variance using $\sigma_{qi}^2 = V(\boldsymbol{X_{q_i}\beta_i})$ where X_{q_i} and $\boldsymbol{\beta_i}$ were the incidence matrix and the allelic effect of QTL i. The incidence matrix X_{q_i} took different forms according to the type of simulated QTL effect (cross-specific, parental, ancestral, bi-allelic). The form of $\boldsymbol{\beta_i}$ followed the definition of the simulated QTLs (Q1-7). The non-zero elements of $\boldsymbol{\beta_i}$ were sampled from a uniform distribution (1-10) with random sign assignment. We scaled the $\boldsymbol{\beta_i}$ values to make sure that the σ_{qi}^2 reached the desired phenotypic proportion (2 or 6 %). Finally we calculated the QTL contribution to the phenotype using $\boldsymbol{y_Q} = \boldsymbol{X_Q\beta}$.

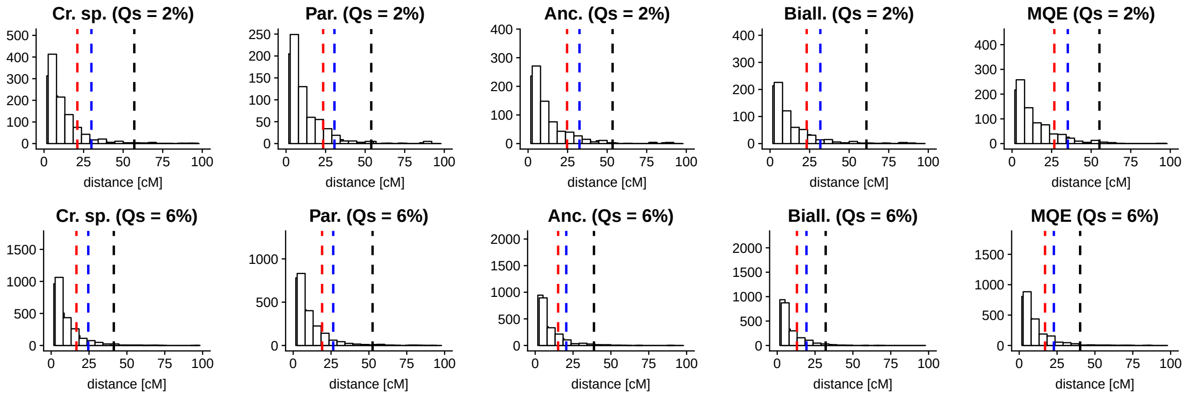
3. We determined the error variance contribution (σ_e^2) such that $\sigma_e^2 = ((1-h^2)/h^2) * \sigma_Q^2$. In all cases, h^2 was equal to 0.32. Given σ_e^2 we sampled the phenotypic variation due to the error using $\boldsymbol{y_e} \sim N(0, \sigma_e^2)$. The simulated phenotypic values were therefore expressed as $\boldsymbol{y_{sim}} = \boldsymbol{y_Q} + \boldsymbol{y_e}$.

4. In many cases, even if the QTLs were sampled on different chromosomes, there was still an important covariance between the QTL positions. Therefore, we sampled a large number of realizations and we kept only the one where the covariance between the QTL positions was inferior to 1% of the total phenotypic variance.

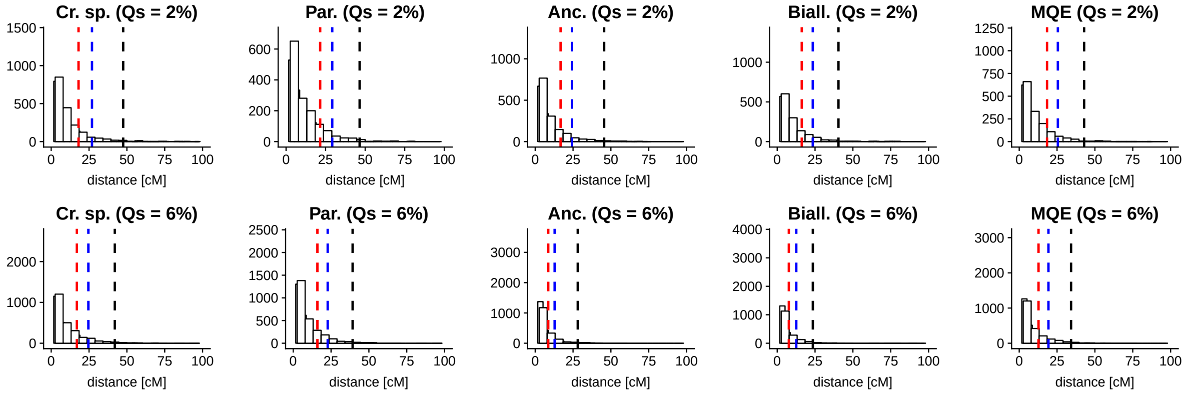
S3: Histograms of the distance between simulated and detected QTL distributions

The red, blue, and black dashed lines represent the .90, .95 and .99 distribution quantiles.

$N = 800$



$N = 1600$



Chapter 5

The usefulness of multi-parent multi-environment QTL analysis: an illustration in different NAM populations

Vincent Garin[1, 2], Marcos Malosetti[1], Fred van Eeuwijk[1]

1. Biometris, Wageningen University & Research Center

2. C.T. de Wit Graduate School for Production Ecology & Resource Conservation (PE & RC)

Abstract

Commonly QTL detection in multi-parent population (MPPs) data measured in multiple environments (ME) is done by a single environment analysis on phenotypic values 'averaged' across environments. This method can be useful to detect QTLs with a consistent effect across environments but it does not allow to estimate environment-specific QTL (QTLxE) effects. Running separate single environment analyses is a possibility to measure QTLxE effects but those analyses do not model the genetic covariance due to the use of the same genotype in different environments. In this paper, we propose methods to analyse MPP-ME QTL experiments using simultaneously the data from several environments and modelling the genotypic covariances. Using data from the EU-NAM and the US-NAM populations, we show that these methods allow to estimate the QTLxE effects and that they give a more precise description of the trait genetic architecture than separate within environment analyses. The MPP-ME models we propose can also be extended to integrate environmental indices (e.g. temperature, precipitation, etc.) to understand better the mechanisms behind the QTLxE effects. Therefore, our methodology allows to exploit the full potential of MPP-ME data: to estimate QTL effect variations a) within the MPP between sub-populations due to different genetic backgrounds; and b) between environments.

5.1 Introduction

The use of multi-parent populations (MPPs) to investigate biological questions becomes progressively a regular practice in plant genetics and plant breeding. Different MPPs have been developed like the nested association mapping (NAM) populations (McMullen et al., 2009) or the multi-parent advanced generation inter-cross (MAGIC) populations (Cavanagh et al., 2008). The collections of crosses between a set of parents used in breeding programs can also be analysed as MPPs (Würschum, 2012; Parisseaux & Bernardo, 2004). Different statistical approaches have been proposed to detect QTLs in NAM populations (Xavier et al., 2015), in MAGIC designs (Verbyla et al., 2014b) or in MPPs composed of crosses (Jourjon et al., 2005).

The plant phenotype is the result of cumulative interactions between the genotype and the environment (Malosetti et al., 2013). Therefore, researchers have developed statistical procedures to detect QTLs taking the genotype by environment (GxE) interactions into consideration (Boer et al., 2007; Korte et al., 2012). Several MPPs have been tested in multiple environments (MPP-ME) (Buckler et al., 2009; Giraud et al., 2014; Saade et al., 2016) but only few studies have proposed a proper MPP GxE QTL detection methodology (Piepho & Pillen, 2004; Verbyla et al., 2014a). Most of the researches average the phenotypic values by calculating adjusted means or predictions across the environments that represent an average phenotypic value (Giraud et al., 2014; Poland et al., 2011; Buckler et al., 2009). In other articles, people performed separate analyses in each environment (e.g. Saade et al. (2016)).

We consider that the main interest of an MPP-ME QTL experiment is to estimate genetic (QTL) variations at two levels: a) within the MPP between sub-populations due to different genetic background; and b) between environments. In previous researches, we developed a framework for QTL detection in MPPs composed of crosses between a set of parents with different assumptions about the QTL effects (Garin et al., 2017, 2018). The QTL effects were assumed to be more or less diverse/consistent in the different genetic backgrounds, which allowed to estimate QTL allelic variations within the MPP. In the present paper, we extended our methodology to the multi-environment context. We allowed the QTL effects to also vary between the environments to estimate QTL by environment (QTLxE) interactions. We further extended our models to integrate environmental information like temperature or precipitation to get a deeper understanding of the mechanisms behind the QTLxE effects. In the following sections, we present different methods to analyse MPP-ME QTL experiments. We illustrate the usefulness of our methodology with examples coming from the EU-NAM and the US-NAM populations.

5.2 Material and methods

5.2.1 Statistical methodology

In Table 5.1, we present a list of models to analyse MPP-ME data. The first set of models allow to model phenotypic variation. The second category represents genotypic models used to detect QTLs. The last section presents four methods for QTL detection in MPP-ME experiments. The models and their components are described in the following sections assuming that the data come from two environments.

5.2.2 Phenotypic models

Model 1 and 2 model phenotypic variations accounting for experimental design factors (des) such as replicates, incomplete blocks, row-columns, etc. In model 1, y_{icp} is the plot data phenotypic value of genotype i from cross c. G_{ic} is the genetic effect of genotype that can be separated into the tested genotypes and the check entries like in Boer et al. (2007). Finally, ϵ_{icp} is the plot error. Model 1 allows to calculate the genotype best linear unbiased estimates (BLUEs) treating G_{ic} as fixed. In model 1, we considered the data from each environment separately and calculated within environment genotype BLUEs ($\boldsymbol{y_{E1}}, \boldsymbol{y_{E2}}$).

In model 2, we model the phenotypic variation of the two environments jointly. y_{icep} is the phenotypic plot measurement taken on genotype i from cross c in environment e. The intercept μ_e, the design term $des_{(e)}$ become environment specific. The term GE_{ice} represents the genotype by environment interaction. In model 1 and 2, all terms were fitted as random except the genotype term G_{ic}. The genotype BLUEs ($\boldsymbol{y_{E(1,2)}}$) obtained with model 2 represent main phenotypic values across environments.

5.2.3 Genotypic models

Model 3 and 4 are MPP QTL detection models for single and multi-environment data, respectively. In model 3, y_{ic} is the phenotypic value of the i^{th} individual in cross c. μ_c is a cross intercept. The term G_{ic} (3.1) describes the random genotypic effect. G_{ic} can be partitioned into a fixed QTL part $\sum_q x_{iq}^T * \beta_q$ and a residual part g_{ic}. x_{iq}^T represents the expected number of QTL alleles received by individual i at position q and β_q are the QTL allelic substitution effects.

The vector x_{iq}^T is of dimensions $[1 \times n_{al}]$ and varies according to the number of alleles (n_{al}) assumed at the QTL position. A first model called parental assumes that each parent contributes a unique allele to the MPP ($n_{al} = n_p$). A second option called ancestral model

Table 5.1: Models table

Code	Model
	Phenotypic models
	Genotype BLUEs within environment
1	$y_{icp} = \mu + des + G_{ic} + \epsilon_{icp}$ BLUEs vectors: $\boldsymbol{y_{E1}}, \boldsymbol{y_{E2}}$
	Genotype BLUEs across environments
2	$y_{icep} = \mu_e + des_{(e)} + G_{ic} + GE_{ice} + \epsilon_{icep}$ BLUEs vector: $\boldsymbol{y_{E(1,2)}}$
	Genotypic models
	Single environment analysis
3	$y_{ic} = \mu_c + G_{ic} + \epsilon_{ic}$
3.1	$G_{ic} = \sum_q x_{iq}^T * \beta_q + g_{ic}$
3.2	$\epsilon_{ic} \sim N(0, \boldsymbol{R}) \quad \text{with} \quad \boldsymbol{R} = \bigoplus_{c=1}^{n_c} \sigma_{\epsilon_c}^2$
3.3	$\boldsymbol{y} = \boldsymbol{X_c \beta_c} + \boldsymbol{X_Q \beta_Q} + \boldsymbol{\epsilon}$
	Multi-environment analysis
4	$y_{ice} = \mu_{ce} + G_{ice} + \epsilon_{ice}$
4.1	$G_{ice} = \sum_q x_{iq}^T * \beta_{qe} + g_{ice}$
4.2	$g_{ice} \sim N(0, \sigma_g^2)$
4.3	$\epsilon_{ice} \sim N(0, \boldsymbol{R}) \quad \text{with} \quad \boldsymbol{R} = \bigoplus_{c=1}^{n_c} \sigma_{\epsilon_{c(e)}}^2$
4.4	$\begin{bmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_c} & 0 \\ 0 & \boldsymbol{X_c} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_{c_1}} \\ \boldsymbol{\beta_{c_2}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{X_Q} & 0 \\ 0 & \boldsymbol{X_Q} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_{Q_1}} \\ \boldsymbol{\beta_{Q_2}} \end{bmatrix} + \boldsymbol{g} + \boldsymbol{\epsilon}$
	MPP-ME QTL detection methods
	M1: MPP QTL analysis on BLUEs across environments
M1	$\boldsymbol{y_{E(1,2)}} = \boldsymbol{X_c \beta_c} + \boldsymbol{X_Q \beta_Q} + \boldsymbol{\epsilon}$
	M2: MPP QTL analysis on BLUEs within environment
M2	$\boldsymbol{y_{E1}} = \boldsymbol{X_c \beta_c} + \boldsymbol{X_Q \beta_Q} + \boldsymbol{\epsilon}$ $\boldsymbol{y_{E2}} = \boldsymbol{X_c \beta_c} + \boldsymbol{X_Q \beta_Q} + \boldsymbol{\epsilon}$
	M3: Two-stage MPP GxE QTL analysis
M3	$\begin{bmatrix} \boldsymbol{y_{E1}} \\ \boldsymbol{y_{E2}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_c} & 0 \\ 0 & \boldsymbol{X_c} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_{cE_1}} \\ \boldsymbol{\beta_{cE_2}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{X_Q} & 0 \\ 0 & \boldsymbol{X_Q} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_{QE_1}} \\ \boldsymbol{\beta_{QE_2}} \end{bmatrix} + \boldsymbol{g} + \boldsymbol{\epsilon}$
	M4: One-stage MPP GxE QTL analysis
M4	$y_{icep} = \mu_{ce} + des_{(e)} + G_{i'(e)} + \sum_q x_{iq}^T * \beta_{qe} + g_{ice} + \epsilon_{icep}$

assumes that genetically similar parents inherit their allele from a common ancestor. Parents are grouped in ancestral classes based on genetic similarity. We assume that each group represents one ancestral allele ($n_{al} = n_a$). The final possibility is a bi-allelic model assuming that genotypes with the same SNP score transmit the same allele.

The elements of x_{iq}^T take values between 0 and 2. For the parental model, they represent the expected number of parental allele copies estimated using IBD probabilities computed by the package R/qtl (Broman et al., 2003). For the ancestral model the vector x_{iq}^T specifying the parental allele distribution is modified taking identical by state (IBS) parental genetic relatedness into consideration. We used the R package clusthaplo (Leroux et al., 2014) to evaluate the parent genetic similarities along the genome and infer common ancestral classes. For the bi-allelic model, x_{iq}^T is a scalar taking values 0, 1 or 2 corresponding to the number of IBS copies of the minor SNP allele. The model is estimated fixing one QTL allele as reference. In NAM populations, it is convenient to set the central parent allele as reference and to interpret the additive allelic substitution effect β_q as the deviations with respect to the central parent.

G_{ic} (3.1) is also composed of a residual genetic effect g_{ic} that was not accounted by the QTLs. In the single environment analysis, g_{ic} is not directly modelled. The residual genetic variation is modelled by the error term ϵ_{ic} (3.2), which also contains the residual plot error. ϵ_{ij} follows a normal distribution $N(0, \boldsymbol{R})$. We assume that $\boldsymbol{R} = \bigoplus_{c=1}^{n_c} \sigma_{\epsilon_c}^2$ which means that the variance of the error term is different in each cross to take into consideration the heterogeneity that could exist between crosses.

Model 3 can be expressed in matrix notation (3.3) with $\boldsymbol{y}$ being the phenotypic values vector of dimension $[N \times 1]$. $\boldsymbol{X_c}$ is a $[N \times n_c]$ cross-specific intercept matrix with $\boldsymbol{\beta_c}$ representing the vectors of cross intercepts. $\boldsymbol{X_Q}$ is a $[N \times n_{al}]$ QTL incidence matrix and $\boldsymbol{\beta_Q}$ the corresponding vector of QTL additive allelic substitution effects.

Model 4 is a modification of model 3 to analyse jointly the data from several environments. Compared to model 3, the cross effects μ_{ce}, the QTL effects β_{qe} (4.1), and the residual genetic effect g_{ice} (4.2) are now indexed per environment. In expression 4.1, we modelled the environmental specific QTL allelic effects. These QTL allelic effects could also be partitioned into a main effect across environments and environment-specific components. This is a difference with respect to model 3. Another important difference between model 3 and 4, is the explicit modelling of the genotypic covariance between phenotypic observations measured on the same genotype in different environments via the term g_{ice} (4.2). g_{ice} is normally distributed $N(0, \sigma_g^2)$, which corresponds to a compound symmetry model assuming a uniform covariance between genotypes in all environments. More sophisticated variance covariance (VCOV) structures are possible like the unstructured model using a specific covariance term for each pair of environments.

The error term ϵ_{ice} (4.3), is normally distributed $N(0, \boldsymbol{R})$ with within environment cross-specific variance error terms $\boldsymbol{R} = \bigoplus_{c=1}^{n_c} \sigma^2_{\epsilon_{c(e)}}$. For an illustration purpose, the VCOV matrix of phenotypic observations coming from two different crosses in two different environments take the following form:

$$V \begin{bmatrix} \boldsymbol{y_{ic_1E_1}} \\ \boldsymbol{y_{i'c_2E_1}} \\ \boldsymbol{y_{ic_1E_2}} \\ \boldsymbol{y_{i'c_2E_2}} \end{bmatrix} = \begin{bmatrix} \sigma^2_g + \sigma^2_{\epsilon_{11}} & 0 & \sigma^2_g & 0 \\ 0 & \sigma^2_g + \sigma^2_{\epsilon_{21}} & 0 & \sigma^2_g \\ \sigma^2_g & 0 & \sigma^2_g + \sigma^2_{\epsilon_{12}} & 0 \\ 0 & \sigma^2_g & 0 & \sigma^2_g + \sigma^2_{\epsilon_{22}} \end{bmatrix}$$

It represents a compound symmetry with heterogeneous cross-specific environment variances. In Model 4, the genotype by environment variance (σ^2_{ge}) can not be distinguished from the error term ($\sigma^2_{\epsilon_{c(e)}}$). Model 4.4 is a matrix expression of model 4. The vector of phenotypic values $\boldsymbol{y} = [\boldsymbol{y_1 y_2}]^T$ includes the phenotypic observations of the two environments. The cross term $\boldsymbol{X_c \beta_c}$ and the QTL term $\boldsymbol{X_Q \beta_Q}$ are extended to model environment-specific cross and QTL effects.

5.2.4 MPP-ME QTL detection methods

Here we present four methods to perform a QTL detection in MPP-ME experiments. The first three methods are two-stage analyses using first a phenotypic model to calculate genotype BLUEs that are used afterwards in a genotypic QTL detection model. The last method is a one-stage analysis. Method M1 performs a QTL analysis on genotype BLUEs calculated across environments. It uses the BLUEs calculated with model 2 as response variable in QTL model 3. Method M1 tests for association between the genotype and $\boldsymbol{y_{E(1,2)}}$, which represents a main phenotypic effect across environments. Therefore, M1 does not allow to estimate the QTLxE effects.

Method M2 allows to model the QTLxE interactions by performing separate QTL analyses in each environment. M2 uses the BLUEs calculated within each environment ($\boldsymbol{y_{E1}}, \boldsymbol{y_{E2}}$) in QTL model 3. The methods M1 and M2 do not take the advantage to analyse jointly the phenotype data measured in different environments, which has been shown to provide a greater understanding of the GxE interactions (Malosetti et al., 2004; Alimi et al., 2013a). A proper MPP GxE QTL detection using mixed model allows to model the heterogeneity of genetic variance across environments and the genetic covariance between environments (Malosetti et al., 2013).

Method M3 analyses jointly the within environment genotypes BLUEs ($\boldsymbol{y_{E1}}, \boldsymbol{y_{E2}}$) using model 4 taking the covariance between the same genotype measured in different environments into consideration. The last method (M4) is a one-stage analysis on the plot data. For the one-stage analysis, we followed Boer et al. (2007) and separated the genetic effect terms of model 2 ($G_{ic} + GE_{ice}$) into a check entries term ($G_{i'(e)}$) and two terms to model the tested genotypes by the QTLs ($\sum_q x_{iq}^T * \beta_{qe}$) and the residual genetic effect (g_{ice}). Method M4 allows to simultaneously estimates the non-genetic effects due to the experimental design ($des_{(e)}$) and the QTL variations.

The variance of the error term $\epsilon_{ic(e)p}$ of models 1, 2 and M4 can be modelled by different VCOV structures taking for example spatial variations into consideration. In models 3, 4, and M4, the cross (μ_{ce}) and the QTL terms (β_{qe}) were fixed. The Wald test (McCulloch & Searle, 2001, 5.39) tests the global null hypothesis of all QTL allelic substitution effects being equal to zero. In models 4 and M4, the null hypothesis will be rejected if one allele is different from zero in at least one environment. The combination of four methods (M1-4) and three types of QTL effects (parental, ancestral, bi-allelic) represents 12 models to analyse MPP-ME QTL experiments.

5.2.5 Plant material

To illustrate our methodology, we focused on two examples showing significant and observable QTLxE interactions. The examples came from the EU-NAM and the US-NAM populations tested for a single trait in two environments.

EU-NAM data

The maize EU-NAM Flint population was composed of 811 double haploid (DH) lines coming from 11 crosses between UH007 and 11 peripheral parents representative of North Europe maize diversity (Bauer et al., 2013; Lehermeier et al., 2014). The EU-NAM Flint population was evaluated in six European locations for five traits. We used the raw phenotypic data provided by Lehermeier et al. (2014) `http://www.genetics.org/content/198/1/3/suppl/DC1`. We used the raw genotypic data provided by Bauer et al. (2013) available here `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50558`, and the consensus map from Giraud et al. (2014) available at: `http://maizegdb.org/data_center/reference?id=9024747`. After quality control, we kept 5949 markers spread on 10 chromosomes for a total map length of 1584 cM. For the ancestral model, we clustered the parental lines at each marker position using a two cM window around the marker with the R package clusthaplo (Leroux et al., 2014). We detected an average of 6.7 ancestral classes along the genome.

From the possible combinations of traits and environments, we focused on dry matter yield (DMY, decitons per hectare, $\frac{dt}{ha}$) measured at La Coruña (Spain - CIAM) and at Roggenstein (Germany - TUM). Within a location, the trials were laid out as augmented p-rep designs with one third of the genotypes replicated. The genotypes were laid out with parents and checks in 160 incomplete block consisting of eight plots. Therefore, in model 1, 2 and M4 $des(e) = rep_{l(e)} + block_{m(le)}$ to account for the l^{th} replicate, and the m^{th} block within replicate effects. The model 2 was equivalent to model 1 in Lehermeier et al. (2014). In M4, the variance of the error term was environment and cross specific (4.3). The average heritability on a line mean basis within cross over environments was $\bar{h}^2 = 47\%$ (supplemental material S1). The Pearson correlation between the two within environment BLUEs was equal to 0.35.

US-NAM data

The maize US-NAM population was composed of 4421 recombinant inbred lines (RIL) coming from 25 crosses between B73 and 25 peripheral parents representative of the international maize diversity (McMullen et al., 2009). It was evaluated for 19 traits in up to 11 environments (Hung et al., 2012). We used the phenotypic data provided by Hung et al. (2012). The genotypic marker data and map data came from Ogut et al. (2015). We downloaded the data from `http://www.panzea.org`. We used the map with 1478 markers (1 mk/cM) and a total map length of 1434 cM. For the US-NAM data, we did not have the IBS marker data. Therefore, we did not calculate the ancestral and the bi-allelic models, and we restricted our analyses to the parental model.

We focused on days to anthesis (DTA) measured in 2007 in the North Carolina (NC) and the New York (NY) environments. Within an environment, the experimental design was a set design where each set contained all lines of a cross. Each set was randomised across environments as an α-design. The α-design was augmented by including the two parental lines of the cross within each incomplete block. The row and column effects were defined at the environment level (Hung et al., 2012). Therefore, in model 1, 2 and M4 $des(e) = set_{l(e)} + block_{m(le)} + row_{n(e)} + col_{o(e)}$ to model the l^{th} set, m^{th} block within set, n^{th} row and o^{th} column effects. Model 2 was similar to the one used by Hung et al. (2012) removing the cross term. The set and the cross effects were not confounded because the parents and the checks were considered as being part of an 'extra' cross.

For the BLUEs computation in model 1 and 2, the VCOV structure of the error term was modelled by an environment-specific autoregressive correlation in the rows and columns (AR1 x AR1) (Gilmour et al., 1997). For the one-stage M4 analysis, the modelling of the spatial trend was however computationally too intensive, so we only used within environment cross-specific error terms (4.3). The average heritability on a line mean basis within cross over environments was $\bar{h^2} = 78\%$ (supplemental material S1). The Pearson correlation between the two within environment BLUEs was equal to 0.25.

5.2.6 QTL detection procedure

The QTL detection procedure was composed of a simple interval mapping scan to select cofactors followed by a composite interval mapping (CIM) scan to build a multi-QTL model. The final list of QTLs was evaluated using a backward elimination. The cofactors were selected with a minimum in between distance of 50 cM to avoid model overfitting. The QTLs were selected with a minimum distance of 20 cM. We fixed the cofactor and QTL detection thresholds to $-log10(p - value) = 4$. We performed the QTL detection extending the R package mppR (Garin et al., 2018) to the multi-environment situation. The mixed models were calculated using asreml-R (Butler et al., 2009). The R package programmed can be found here `https://github.com/vincentgarin/mppGxE`.

5.2.7 Cross-validation

We adapted the cross-validation (CV) procedure described by Utz et al. (2000) to the MPP-ME context to evaluate the performances of the QTL detections. For each combination of methods (M1-4) and type of QTL effects (parental, ancestral, bi-allelic), we ran three times a three-fold CV procedure, which gave us nine estimates for each parameter. We did not perform the CV for the one-stage (M4) QTL analysis of the US-NAM data due to computational limitations. We partitioned the full dataset at the within-cross level into an estimation set (ES) and a test set (TS). We used the estimated effects of the QTLs detected in the ES to predict phenotypic values of the ES and TS (e.g. $\boldsymbol{\hat{y}_{TS}} = \boldsymbol{X_{TS}}\boldsymbol{\hat{\beta}_{TS}}$).

We calculated the proportion of genetic variance explained in the ES by the QTLs ($\hat{p}_{ES}$) using the Pearson correlation between the reference values $\boldsymbol{y_{ES}}$ and the predicted values $\boldsymbol{\hat{y}_{ES}}$. We calculated the proportion of genetic variance predicted in the TS by the QTLs ($\hat{p}_{TS}$) using the Pearson correlation between the reference values $\boldsymbol{y_{TS}}$ and the predicted values $\boldsymbol{\hat{y}_{TS}}$. We used as reference values ($\boldsymbol{y_{ES}}$ and $\boldsymbol{y_{TS}}$) the within environment BLUEs (model 1) to compare our results across methods. The $\hat{p}_{ES}$ and $\hat{p}_{TS}$ were computed within crosses per environment. We estimated the $\hat{p}_{ES}$ and $\hat{p}_{TS}$ at the whole MPP level by calculating the average within cross values.

5.2.8 Modelling QTL effect in relation to environmental information

The methods M2, M3 and M4 allow to detect environmental differences of the QTL effects but they do not allow to understand how environmental characteristics influence the QTL effects. A natural extension is to integrate environmental information to understand better the QTLxE interaction. Inspired by the model 16 proposed by Malosetti et al. (2013), we can reformulate the QTL part of G_{ijk} (4.1) to describe the QTL effect of a single QTL (q^*) in term of an environmental covariate (Z_e).

$$\sum_q x_{iq}^T \beta_{qe} = \sum_{q \in Q; q^* \notin Q} x_{iq}^T \beta_{qe} + x_{iq^*}^T (\alpha_{q^*} + Z_e \beta_{q*} + a_{eq^*}) \tag{5}$$

In this formula the QTL part is the same as in 4.1 for all QTLs except for q^*. For the QTL q^*, we decompose its environmental effect (β_{q^*e}) into a main effect component α_{q^*} and a component β_{q*} describing the sensitivity to the environmental covariate Z_e. In the MPP context, α_{q^*} and β_{q*} will be vectors of dimension $[1 \times n_{al}]$ containing one element per QTL allele l. α_{q^*l} represents the QTL effect of allele l when Z_e is equal to zero. β_{q*l} is the environmental sensitivity of QTL allele l and represents the amount of change in trait quantity for one extra unit of Z_e. α_{q^*l} and β_{q*l} are both defined with respect to a reference allele (e.g. the central parent). Z_e is the value of the environmental covariate (e.g. temperature). Finally, $a_{eq^*} \sim N(0, \sigma^2_{aq^*})$ is the residual unexplained QTL effect.

5.3 Results

5.3.1 Cross-validation results

Table 5.2 contains the CV results for the two populations over the different combinations of methods (M1-M4) and QTL effects (parental, ancestral, bi-allelic). We did not detect large differences in terms of prediction power ($\hat{p}_{TS}$) between the different methods. For example, for the US-NAM data in the first environment (NC), the average $\hat{p}_{TS}$ across crosses was equal to 15, 16 and 17 for the M1, M2 and M3 methods, respectively. Similarly, for the EU-NAM ancestral model in the second environment (TUM), the $\hat{p}_{TS}$ was equal to 11, 11, 15 and 12 for the M1 to M4 methods, respectively. Concerning the number of QTLs, we observed that, on average, the separate single environment M2 method detected the lowest number of QTLs. The one-stage M4 analyses detected less QTLs than M1 and M3. In some cases, method M1 detected the largest number of QTLs like in the three EU-NAM complete data analyses.

Concerning the type of QTL effect (parental, ancestral, bi-allelic), We also did not detect any difference in terms of prediction power ($\hat{p}_{TS}$). For the EU-NAM data, the average within crosses $\hat{p}_{TS}$ of the parental, of the ancestral, and of the bi-allelic model varied between: 8-16, 10-15, and 9-14, respectively. We noticed however, that we detected on average more QTLs with the bi-allelic and the ancestral models.

5.3.2 -log10(p-values) scatter-plots

In Figure 5.1, we plotted the -log10(p-values) of the CIM profiles obtained with the method M4 with respect to the methods M1 to M3 for all complete data QTL analyses. We could observe that, in general, the -log10(p-values) were larger in M4. The differences between the M4 and the M2 profiles were the largest. Concerning M4 versus M1, an important fraction of the -log10(p-values) were superior in the M4 profiles with respect to the M1 profiles. However, for the most significant -log10(p-values), the M1 method gave sometimes slightly larger -log10(p-values) than the M4 method, for example for the EU-NAM parental model. The -log10(p-values) obtained with M3 and M4 were similar.

5.3.3 Detected QTLs

In Figure 5.2, we plotted the -log10(p-values) CIM profile of the US-NAM M4 parental QTL analysis with a representation of the genetic effect significance per parent and per environment along the genome. On that plot, we observed that the QTLs on chromosome eight, nine, and ten were detected with a large significance. Looking at the genetic effect distribution along the genome, we also noticed that these QTLs potentially had a different parental allelic series between the environments. The information about the significance of the QTL genetic effect along the genome should however be taken with caution because it is based on an incremental and conditional Wald test that can change given the order of the tested parameters (Butler et al., 2009).

In Figure 5.3, we plotted the -log10(p-values) CIM profile of the EU-NAM M4 ancestral QTL analysis with a representation of the genetic effect significance per parent and per environment along the genome. We noticed that the QTL on chromosome six had an interesting allelic series. Many parents were grouped in the same ancestral class and the QTL had a genetic effect specific to the second environment (TUM). The rest of the detected positions in all methods and type of QTL model combinations can be found in the supplementary material S2.

5.3.4 Estimation of the QTL allelic substitution effects

We represented the QTL allelic series of the QTL detected on chromosome six in the EU-NAM ancestral model and the QTLs detected on chromosome eight, nine and ten

Table 5.2: Cross-validation results of the EU-NAM and US-NAM data for each combination of methods (M1-4) and type of QTL effects (parental, ancestral bi-allelic). Number of QTLs detected in the complete data analyses ($N.QTL$). Average number of detected QTLs in the TS over the CV runs ($N.QTL_{cv}$). Average proportion of variance explained by the detected QTLs in the estimation set ($\hat{p}_{ES}$) over the crosses with minimum and maximum within cross values. Average proportion of variance predicted by the detected QTLs in the test set ($\hat{p}_{TS}$) over the crosses with minimum and maximum within cross values.

		EU-NAM				US-NAM			
		$N.QTL$	$N.QTL_{cv}$	$\hat{p}_{ES}$	$\hat{p}_{TS}$	$N.QTL$	$N.QTL_{cv}$	$\hat{p}_{ES}$	$\hat{p}_{TS}$
parental									
M1	E1	8	4.9	23 (13-42)	10 (5-19)	22	14.8	29 (15-45)	15 (4-37)
M1	E2	8	4.9	21 (6-44)	9 (2-19)	22	14.8	18 (4-47)	9 (1-35)
M2	E1	6	3.7	23 (13-53)	8 (2-36)	24	16.6	40 (27-56)	16 (4-31)
M2	E2	3	2.7	20 (6-35)	12 (3-25)	7	4.8	14 (3-45)	8 (1-39)
M3	E1	7	5.1	28 (15-54)	10 (2-24)	27	17.2	40 (27-55)	17 (6-38)
M3	E2	7	5.1	33 (11-61)	12 (3-21)	27	17.2	27 (15-54)	8 (0-33)
M4	E1	6	3.7	21 (10-39)	12 (3-24)	18			
M4	E2	6	3.7	22 (7-34)	16 (1-28)	18			
ancestral									
M1	E1	9	5.2	17 (4-36)	11 (1-24)				
M1	E2	9	5.2	19 (7-33)	11 (3-29)				
M2	E1	6	4.6	21 (8-58)	10 (3-20)				
M2	E2	4	3.8	21 (8-41)	11 (1-19)				
M3	E1	6	7.2	28 (6-59)	10 (1-23)				
M3	E2	6	7.2	30 (6-57)	15 (4-30)				
M4	E1	7	3.7	15 (5-24)	10 (3-17)				
M4	E2	7	3.7	19 (4-33)	12 (3-24)				
bi-allelic									
M1	E1	12	8.4	18 (5-28)	11 (2-31)				
M1	E2	12	8.4	20 (7-32)	12 (6-30)				
M2	E1	7	5.7	20 (7-32)	10 (3-21)				
M2	E2	5	5.1	20 (7-33)	10 (1-22)				
M3	E1	10	7.8	20 (6-34)	9 (1-14)				
M3	E2	10	7.8	19 (3-35)	14 (7-23)				
M4	E1	9	7	15 (4-27)	12 (2-25)				
M4	E2	9	7	18 (3-32)	12 (3-25)				

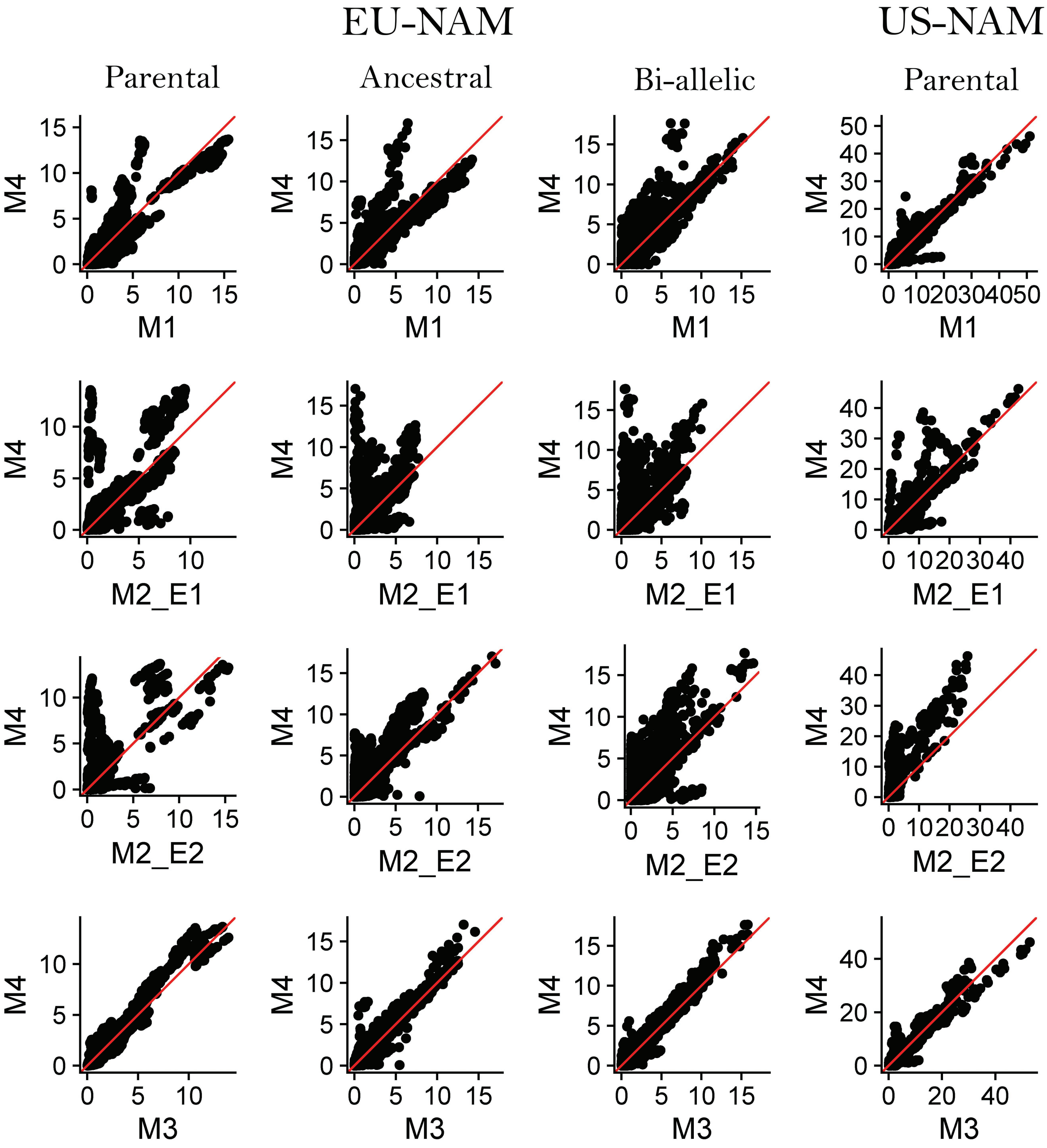


Figure 5.1: CIM -log10(p-values) scatter-plots of methods M1-M3 compared to M4 for both the EU-NAM and the US-NAM whole population QTL analyses.

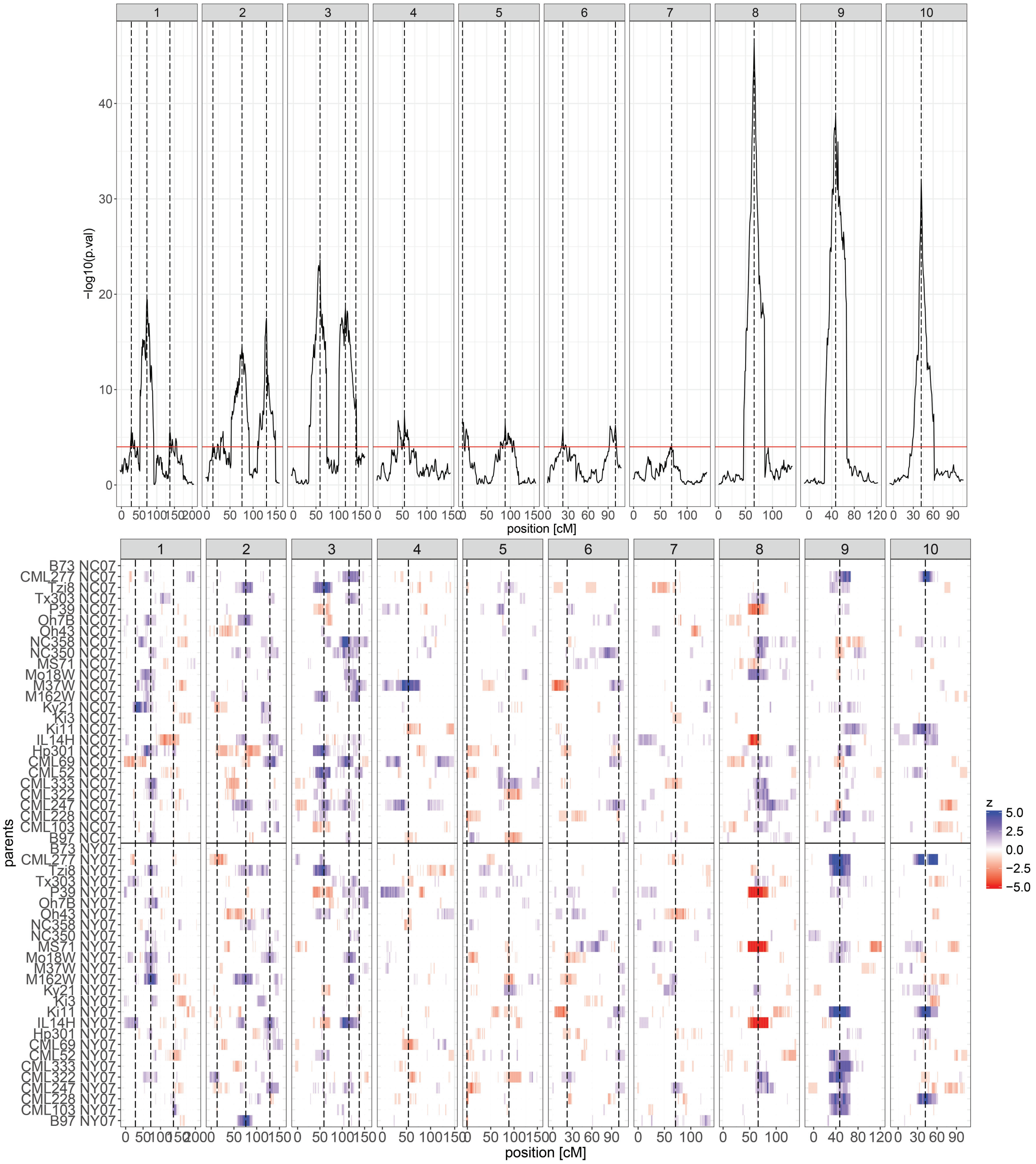


Figure 5.2: CIM -log10(p-values) profile of the US-NAM M4 parental QTL analysis. The lower part of the figure represents the within environment parental QTL allelic significance along the genome. The Wald test p-values of the parental allelic substitution effects are converted into a colour code from > 0.05 (1) to $> 10^{-5}$ (5). The colours red (negative) and blue (positive) correspond to the sign of the QTL effect

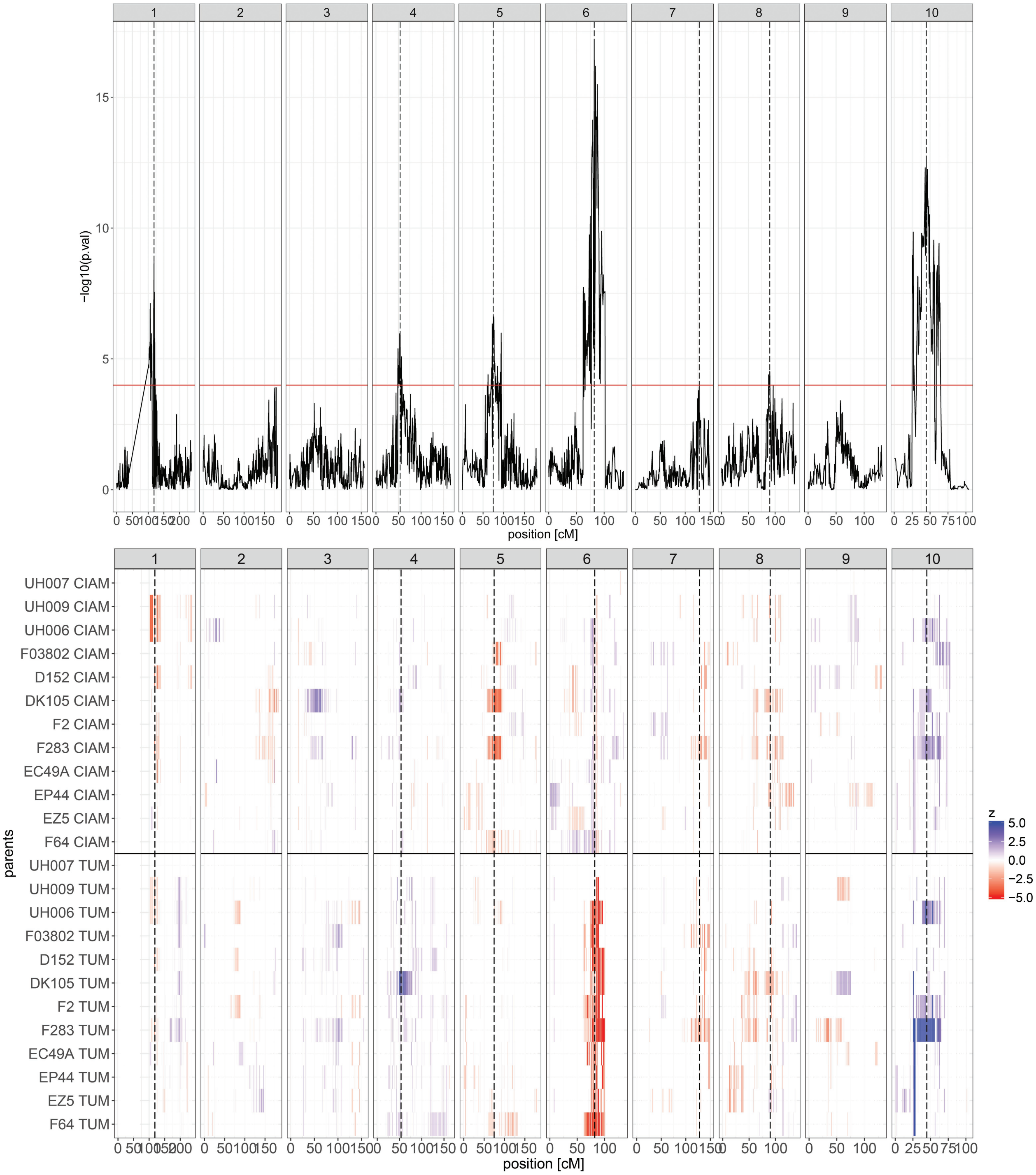


Figure 5.3: CIM -log10(p-values) profile of the EU-NAM M4 ancestral QTL analysis. The lower part of the figure represents the within environment parental QTL allelic significance along the genome. The Wald test p-values of the parental allelic substitution effects are converted into a colour code from > 0.05 (1) to $> 10^{-5}$ (5). The colours red (negative) and blue (positive) correspond to the sign of the QTL effect

in the US-NAM data in Figure 5.4. The estimated QTL effects were conditioned on the list of cofactors detected with the corresponding final model. The numerical QTL allelic substitution effect values can be found in the supplementary material S3. In the text, the standard errors of the allelic effects are given in parentheses. In Figure 5.4, we observed the differences in terms of QTL effect estimation between method M1 using BLUEs representing an average phenotypic effect across environments, and the method M4, which allows to estimate environment-specific QTL effects. The results obtained with method M4 were comparable to the ones obtained with methods M2 and M3. A full comparison between the estimated QTLs effects obtained with all the methods can be found in the supplementary material S4.

The QTL detected on chromosome six with the ancestral model in the EU-NAM population (Figure 5.4-A), was the most illustrative example of environment-specific QTL effects. In the second environment (TUM), an ancestral allele inherited by parents D152, EC49A, F03802, F2, F283, UH006 and DK105 had a strong negative effect of $-7.2(0.8)\frac{dt}{ha}$ with respect to the ancestral group containing parents UH007, EZ5, and UH009. In the first environment (CIAM), the effect of the main ancestral group was substantially reduced $(-1.2(0.8)\frac{dt}{ha})$ and non-significant. In the M1 method, the main ancestral allele took an average values $(-4.1(0.7)\frac{dt}{ha})$ across the two environments.

In example 5.4B, we observed that the three most significant parental alleles had an effect that was stronger in the second environment (NY) compared to the first one (NC). The estimated allelic substitution effects (in days) of parental alleles IL14H, MS71, and P39 were -1.3(0.1) vs -0.8(0.1), -1.2(0.1) vs -0.5(0.1), and -1.6(0.2) vs -1.0(0.2), respectively. The estimated QTL effects obtained with M1 were again approximately averaged between the two environments with -1.1, -1.1 and -1.3 days respectively. Similarly, in example 5.4C, we observed that several parental alleles had a stronger positive expression in NY compared to NC. For example, the allelic substitution effect of Ki11 was equal to $+1.9(0.2)$ days in NY and $+0.3(0.2)$ days in NC. The allelic substitution effect of CML277 was equal to $+1.9(0.2)$ days in NY while it was equal to $+0.8(0.2)$ days in NC. Finally, in example 5.4D, the most significant parental alleles had a stronger effect in the second (NY) environment. For example, the estimated allelic substitution effect of CML277 was equal to $+3(0.2)$ days in NY and $+1.4(0.2)$ days in NC. Once again, in the examples 5.4C and 5.4D, the allelic substitution effects estimated with method M1 were approximately averaged across the two environments.

5.3.5 QTL effect in relation to environmental information

To illustrate the extension of our models with environmental covariates, we re-analysed the effect of the QTL detected on chromosome six (84.2 cM) in the EU-NAM with the

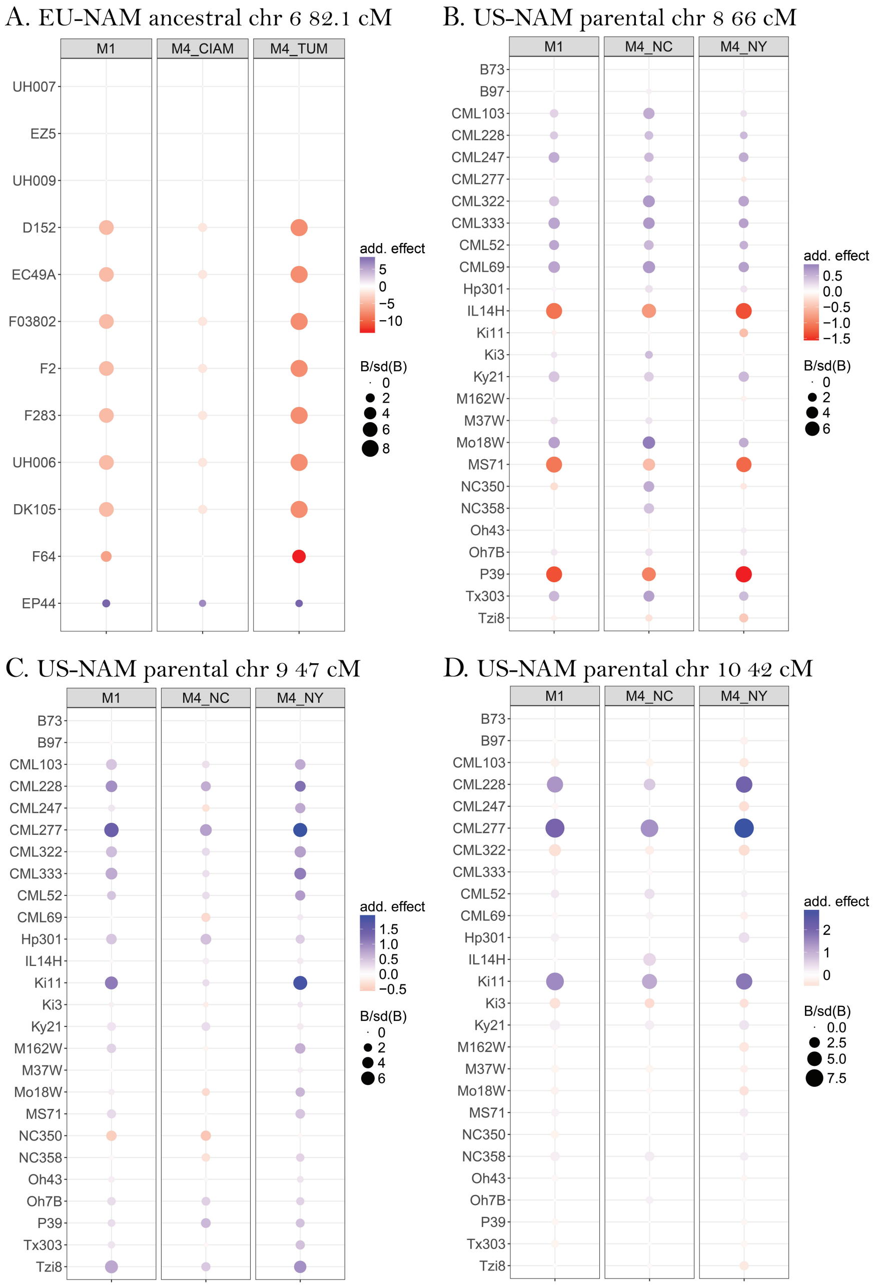


Figure 5.4: Comparison of the allelic substitution effect series between M1 and M4 for four QTL positions: A) EU-NAM ancestral model chromosome 6 82.1 cM; B) US-NAM parental model chromosome 8 66 cM; C) US-NAM parental model chromosome 9 47 cM; D) US-NAM parental model chromosome 10 42 cM. The colour intensities are proportional to the allelic effect. The allelic effects are deviations in decitons per hectare (A) and days (B-D) with respect to the central parent (UH007 or B73). The sizes of the dots are proportional to the ratio between the allelic effect and its standard error.

M3 ancestral model including the effect of water precipitation (Z_e). This QTL had five alleles: allele A (UH007- central parent), allele B (D152, EC49A, EP44, F2, F64,

Table 5.3: Parameter estimates of the environment-specific QTL effects expressed in terms of the QTL sensitivity to average water precipitation during July and August.

	Estimates	Std. Err	Units	Wald	Df	P(Wald)	
α_A	0			0	0	-	
α_B	-0.75	1.07		12.5	1	<0.001	***
α_C	2.02	1.39	$dt * ha^{-1}$	3.5	1	0.06	.
α_D	1.82	1.44		1.1	1	0.3	
α_E	-9.02	4.6		2.7	1	0.1	
β_A	0			0	0	-	
β_B	-0.06	0.03		4	1	0.05	*
β_C	-0.12	0.04	$dt * ha^{-1} * Wat.(mm)$	8.4	1	0.003	**
β_D	-0.03	0.04		0.7	1	0.42	
β_E	0.17	0.14		1.6	1	0.21	

Table 5.4: Extra yield homozygous genotype with allele A versus B given water precipitation.

Location	Z_e (mm)	Yield (Std. Err) $[dt * ha^{-1}]$
La Coruna	0	1.50 (2.13)
Roggenstein	42	6.60 (1.64)
Ploudaniel	28	4.80 (1.35)
Einbeck	33	5.46 (1.40)

UH006), allele C (F03802, F283), allele D (UH009, DK105), and allele E (EZ5). We used the final QTL model and the average water precipitation in mm ($Wat.(mm)$) at each location between July and August obtained from `https://en.climate-data.org/`. To increase the range of the environmental covariate, we included four environments (La Coruña, Roggenstein, Einbeck, Ploudaniel). We considered the precipitation of the driest location (La Coruña - 35 mm) as the reference level. The precipitation in the other environments were expressed as the difference with respect to the reference.

In Table 5.3, we can observe the estimates of the QTL main effect (α) and QTL x environment effect (β) on the trait. We noticed that, in the driest environment (La Coruña), the allele B reduced the yield by 0.75 $\frac{dt}{ha}$ compared to allele A. When the level of precipitation increased, this difference was accentuated by 0.06 $\frac{dt}{ha} * Wat.(mm)$. In Table 5.4, we calculated the difference in yield between an homozygous genotype with allele A versus B in the four environments ($2 * (\alpha_B + Z_e * \beta_B)$). In this table, we observed that the difference was equal to 1.5 in the driest reference environment (La Coruña). The extra yield given by allele A increased with more precipitation (e.g. 6.6 $\frac{dt}{ha}$ at Roggenstein).

5.4 Discussion

Several MPPs have been characterized in multiple environments but, most of the time, the QTL analyses were performed on genotype BLUEs calculated across environments, which represent average phenotypic values (Giraud et al., 2014; Buckler et al., 2009). We called that method M1. M1 does not estimate the QTLxE effects. Therefore, M1 does not allow to use the full information potential of MPP-ME QTL experiments, measuring QTL variations: a) within the MPP between sub-populations due to different genetic backgrounds; and b) between environments.

An alternative to estimate the QTLxE effects in MPPs is to perform separate QTL analyses within each environment (M2). However, this method does not model the covariance due to the repeated measurements on the same genotype. Therefore, we proposed two methods (M3 and M4) that allowed to model properly the QTLxE effects in MPPs. Those methods used simultaneously the phenotypic data from multiple environments taking into consideration the covariance existing between the same genotype measured in different environments. With respect to M3, M4 also integrated the sources of variation due to experimental design elements performing a one-stage analysis on the plot data. Using M3 and M4 we were able to detect important QTLs and we showed that these QTLs had environmental specific allelic effects. Finally, we extended our models to integrate environmental information and better characterize the QTLxE effects.

5.4.1 Comparison of the detected QTLs with the existing literature

Using the different methods presented, we detected several interesting QTLs for DMY and for DTA. The QTL on chromosome six at 82.1 cM detected by the ancestral model in the EU-NAM population (Figure 5.4A) was also detected by Giraud et al. (2014). They detected a QTL at 83.5 cM, 90.5 cM and 90.7 cM using the connected, the 2cM-LDLA and the 1mk-LDLA models, respectively.

The QTLs detected for DTA in the US-NAM population on chromosomes eight, nine, and ten, at 67, 47, and 42 cM, respectively (Figure 5.4B, C and D), were also detected by Buckler et al. (2009). They detected corresponding QTLs on chromosomes, eight, nine, and ten, at 67, 45.2, and 42.9 cM, respectively. According to Buckler et al. (2009), the QTL on chromosome eight is close to the vegetative to generative transition 1 (vgt1) gene. Concerning the QTL on chromosome ten (42 cM), Giraud et al. (2014) also detected a QTL with a strong effect on flowering time between 45 to 50 cM. According to them, it corresponds to the ZmCCT gene.

5.4.2 Estimation of the QTLxE effects

In Figure 5.4 and appendix S4, we showed that important QTLs had environment-specific allelic substitution effects. In those cases, the QTL effects estimated with method M1 were inaccurate. Those QTL effects were overestimated in one environment and underestimated in the other. Therefore, in presence of QTLxE effect, the use of method M2, M3, or M4 is necessary to estimate properly the QTL environmental differences.

The QTL detected on chromosome six at 82.1 cM in the EU-NAM population (Figure 5.4A) is a good illustration of the possibility to estimate QTL effect variations at two levels: within the MPP between sub-populations, and between environments. At that position, the QTL effect was rather consistent within the MPP because an important group of parents showed a negative effect that could be due to a common ancestral allele. For that QTL, we could also observe environmental variation because the allelic effects were only significant in the second environment. This example illustrates the interest of using a sound MPP GxE QTL detection methodology.

5.4.3 QTL effect in relation to environmental information

In Table 5.3 and 5.4, we showed that methods M3 or M4 could be extended to integrate environmental information to better characterize the QTLxE effects. The estimation of the water precipitation effect on a single QTL was a simple case with a unique environmental covariate but we could imagine more complex models with more QTLs and/or environmental covariates (Malosetti et al., 2004). We assumed a linear relationship between the QTL effect and the environmental covariate. More complex relationships such as a quadratic form or splines could also be assumed (van Eeuwijk et al., 2007). The ultimate goal of such an approach is to unravel the physiological mechanisms behind the QTL effects. This possibility to integrate environment information make methods M3 or M4 more attractive than M2.

5.4.4 Full data analyses and cross-validation results

In terms of prediction power, the CV results (Table 5.2) did not allow to make a difference between the four methods. The $\hat{p}_{TS}$ were mostly similar. We showed in the full data analyses and the CV that, on average, method M2 detected less QTLs. The prediction power ($\hat{p}_{TS}$) of M2 was also reduced with respect to the other methods, even if the differences were small. According to our results, the use of separate within environment analyses is therefore not an optimal strategy. The joint analysis of multi-environment data, as performed in M3 and M4, accounted better for the shared effects across environments, which was beneficial for the QTL analysis.

In the full data analysis, we also noticed that in all EU-NAM analyses (parental, ancestral, bi-allelic), method M1 detected the largest number of QTLs. In those cases, the extra power of M1 could be explained by the fact that this method uses a reduced number of degrees of freedom (df) to estimate the QTL effect. Indeed, in M1, the QTL term uses $n_{al} - 1$ df, while in M2, M3 or M4, the QTL term uses $N_{Env} * (n_{al} - 1)$ dfs. When the QTLxE effect is strong, the loss in power due to the extra df for the QTLxE term is compensated by a better modelling of the GxE effect. However, when the QTL effect is consistent across the environments, the additional modelling of the QTLxE variation penalizes the test for a QTL. With consistent QTL effects, method M1 is more parsimonious.

We would like to emphasize that we selected examples with significant observable GxE interactions but that these situations did not represent the majority of the cases we tested. This observation is in agreement with Buckler et al. (2009) who found that, for flowering time traits in the US-NAM population, the environment-specific QTL effects were small compared to the main QTL effects across environments. The potential weakness of QTLxE effects with respect to the main QTL effects in the US-NAM and EU-NAM populations, could explain why, in some cases, method M1 obtained better results than the GxE analyses (M2-M4). The existence of GxE effects could be more important in MPP-ME QTL experiments where the environmental conditions are more controlled. For example, in experiments testing the same population in control versus heat or salt stress conditions (Saade et al., 2016).

As observed in Table 5.2, method M3 detected on average more QTLs than M4. Looking at the list of detected QTLs (appendix S2), we noticed that, in the EU-NAM analyses, the positions of the QTLs in M3 and M4 were consistent. In the US-NAM analyses, the positions of the QTLs with a large significance were consistent between M3 and M4. However, several QTLs with a low or medium significance were either only detected in M3, or distant by 10-15 cM between the two methods.

The difference between M3 and M4 could be explained by the modelling of the experimental design variation in M4. Piepho & Pillen (2004) showed that even if the variance of the experimental design factors were smaller than other elements like the genetic covariance, it could still reduce substantially the QTL effect when it was used in a one-stage analysis. This reduction of the QTL effect could explain why we detected less QTLs in M4 with respect to M3. We should still remember that in the US-NAM analysis we used an AR1 x AR1 covariance structure for the error term to calculate the within environment BLUEs used in the M3 analyses. Due to computational limitations, we did not use the AR1 x AR1 covariance structure in M4. This could also explain the differences between M3 and M4 in the US-NAM analyses.

5.4.5 Other extensions

To present our methodology, we used examples from NAM populations but our methods and the reasoning behind it are also valid for any MPP composed of crosses like diallel population or factorial designs used in breeding programs. Our methods can also be adapted to the multi-trait situation. The analysis of longitudinal traits measured at different time points using a VCOV reflecting the time dependence could be a possibility. For an illustration purpose, we only used data coming from two environments but we could increase this number. However, the fitting of mixed models on large datasets can be computationally intensive. For example, it took us four days to perform the M4 one-stage QTL detection in the US-NAM population on a personal computer (Intel Core i7-3770 CPU 3.4 GHz).

5.4.6 Conclusions

We proposed mixed model methods to detect QTL in MPP-ME data. These models analysed jointly data from several environments and modelled the genetic covariance due to repeated measurements on the same genotype. Our methods allowed to estimate the QTLxE effect while the methods using genotype BLUEs calculated across environments did not. However, the methods focusing on the main QTL effects remain useful if the QTL effects are consistent across the environments. Moreover, we showed that our methods could be extended to integrate environmental information and understand better the mechanisms behind the QTLxE effects. The methods we proposed are therefore an interesting tool to exploit the full information potential of MPP-ME data. They allow to estimate the QTL variations: a) within MPP between sub-populations due to different genetic backgrounds; and b) between environments.

5.5 Supplementary Material

S1: Heritability computation

We computed the trait heritabilities on a line mean basis within the j^{th} cross using the formula 3 from Hung et al. (2012)

$$h^2 = \frac{\sigma^2_{g(cr_j)}}{\sigma^2_{g(cr_j)} + \frac{\sigma^2_{gxE(cr_j)}}{N_{env}} + \frac{\sigma^2_{e(cr_j)}}{N_{env} * N_{rep}}} \tag{5.1}$$

EU-NAM data

The model used for the computation of the heritabilities was the following:

$$\begin{aligned} y_{ijklm} =& \mu + env_k + rep_{l(k)} + block_{m(lk)} + cross_j + \\ & G_i(cross_j) + G_i(cross_j) * env_k + e_{ijklm} \end{aligned} \tag{5.2}$$

In models 5.2, all terms were random and the variance of the error term was cross-specific.

	σ^2_g	se(σ^2_g)	σ^2_{ge}	se(σ^2_{ge})	h^2(%)
D152	208.36	65.38	93.36	12.26	60.65
EC49A	0.11	93.29	80.07	63.86	0.04
EP44	257.42	139.59	29.11	176.75	66.64
EZ5	41.23	190.46	0.00		19.14
F03802	145.96	47.45	528.20	282.09	27.36
F2	302.10	79.10	75.92	66.24	71.97
F283	361.53	70.04	0.00		77.90
F64	222.86	80.69	100.30	61.39	60.92
UH006	224.28	60.12	138.10	81.42	57.17
UH009	20.12	37.53	75.42	62.59	12.70
DK105	270.12	75.60	106.34	61.26	59.76

US-NAM data

The model used for the computation of the heritabilities was the following:

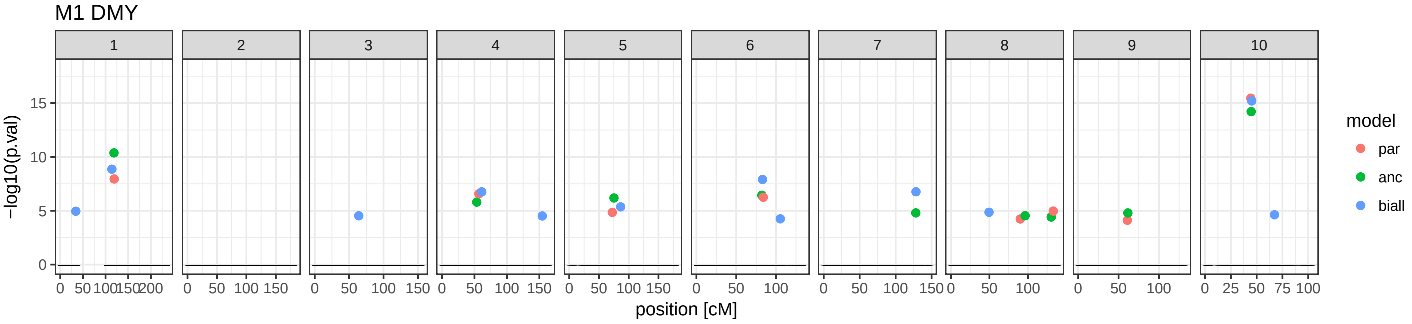
$$
\begin{aligned}
y_{ijklmno} =& \mu + env_k + set_{l(k)} + block_{m(lk)} + row_{n(k)} + col_{o(k)} + cross_j + \\
& G_i(cross_j) + G_i * env_k + e_{ijklm}
\end{aligned} \tag{5.3}
$$

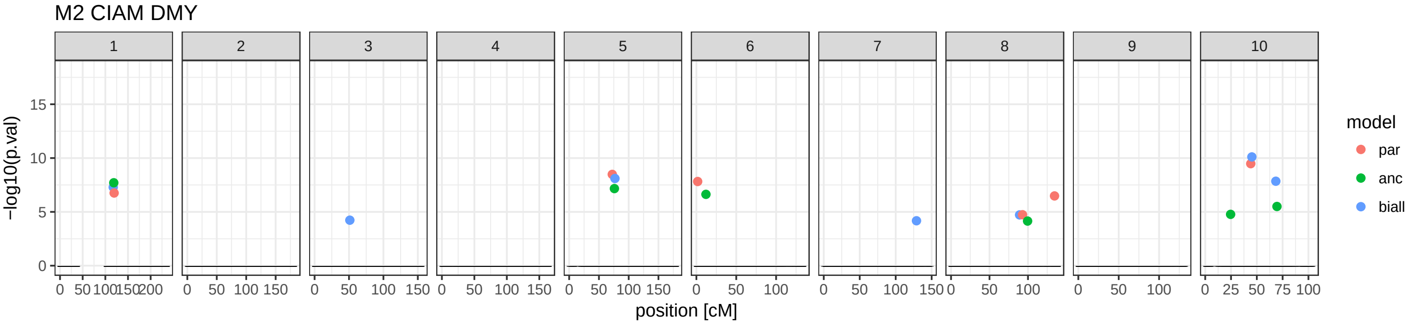
In models 5.3, all terms were random and the variance of the error term was modelled by an environmental specific autoregressive correlation in the rows and columns (AR1 x AR1) (Gilmour et al., 1997).

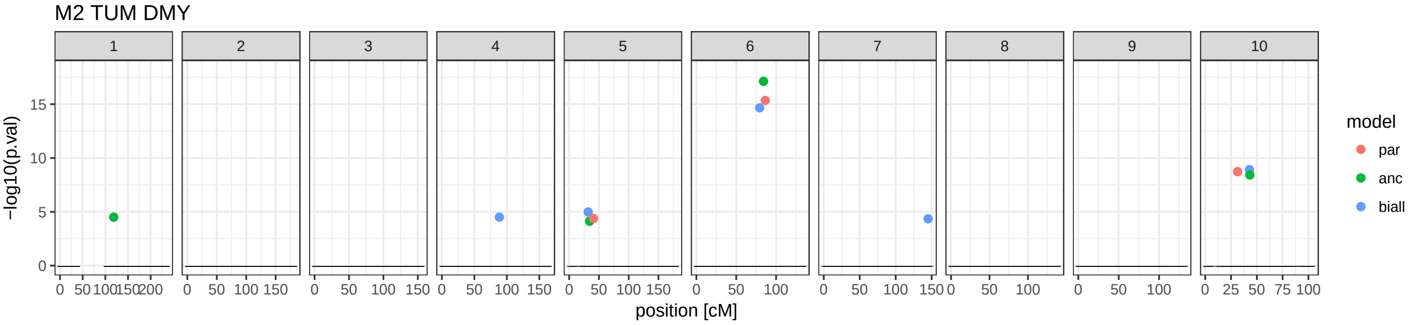
	σ^2_g	se(σ^2_g)	σ^2_{ge}	se(σ^2_{ge})	h^2(%)
B97	1.77	0.32	0.97	0.10	75.64
CML103	1.69	0.37	0.97	0.10	64.63
CML228	7.09	1.13	0.97	0.10	73.13
CML247	7.19	0.94	0.97	0.10	85.04
CML277	10.02	1.44	0.97	0.10	79.75
CML322	3.57	0.54	0.97	0.10	82.26
CML333	4.18	0.62	0.97	0.10	80.23
CML52	6.20	0.83	0.97	0.10	86.47
CML69	2.13	0.47	0.97	0.10	61.21
Hp301	3.24	0.44	0.97	0.10	86.98
IL14H	4.32	0.54	0.97	0.10	89.91
Ki11	6.50	1.06	0.97	0.10	70.35
Ki3	3.79	0.64	0.97	0.10	86.53
Ky21	1.64	0.32	0.97	0.10	71.62
M162W	3.09	0.50	0.97	0.10	78.33
M37W	3.09	0.53	0.97	0.10	73.31
Mo18W	4.84	0.76	0.97	0.10	76.34
MS71	2.22	0.32	0.97	0.10	82.07
NC350	2.77	0.51	0.97	0.10	70.48
NC358	1.67	0.33	0.97	0.10	73.09
Oh43	1.73	0.33	0.97	0.10	72.38
Oh7B	2.51	0.42	0.97	0.10	80.19
P39	5.47	0.68	0.97	0.10	91.86
Tx303	2.89	0.50	0.97	0.10	74.29
Tzi8	4.60	0.68	0.97	0.10	85.03

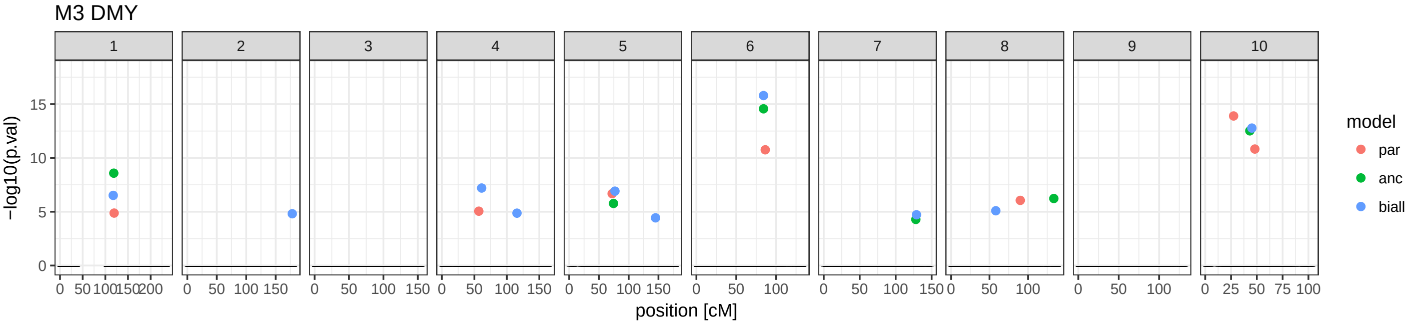
S2: List of detected QTLs

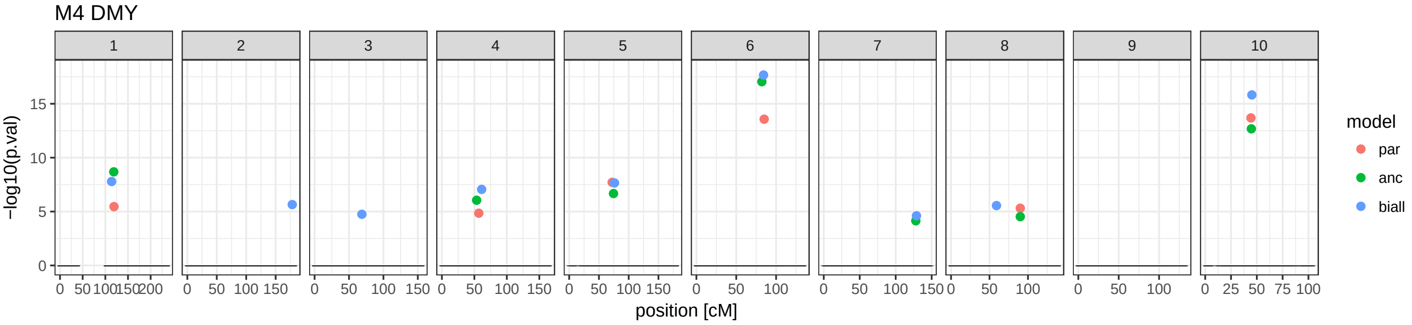
EU-NAM data - List of detected QTLs











EU-NAM M1

	n	Mk. names	chr	pos [cM]	-log10(pval)
parental					
	1	PZE.101146834	1	119.1	8
	2	PZE.104027223	4	56.8	6.6
	3	PZE.105054186	5	72.4	4.8
	4	PZE.106101027	6	83.9	6.3
	5	PZE.108099840	8	90.2	4.2
	6	PZE.108131921	8	133.1	5
	7	PZE.109054632	9	60.9	4.1
	8	PZE.110048720	10	44.3	15.5
ancestral					
	1	PZE.101144216	1	118.6	10.4
	2	PZE.104029507	4	53.5	5.8
	3	PZE.105068880	5	75.2	6.2
	4	PZE.106098066	6	82.1	6.4
	5	PZE.107128534	7	128	4.8
	6	PZE.108109731	8	96.4	4.5
	7	PZE.108131479	8	130.4	4.4
	8	PZE.109058296	9	61.5	4.8
	9	PZE.110049068	10	44.7	14.2
bi-allelic					
	1	PZE.101024519	1	34.3	5
	2	PZE.101143233	1	113.9	8.9
	3	PZE.103096063	3	64	4.5
	4	PZE.104052802	4	61.2	6.8
	5	PZE.104153023	4	154.5	4.5
	6	PZE.105103875	5	86.4	5.4
	7	PZE.106097991	6	83	7.9
	8	PZE.106114241	6	105.3	4.3
	9	PZE.107128336	7	128.3	6.8
	10	PZE.108027746	8	49.5	4.9
	11	PZE.110049474	10	45.2	15.2
	12	PZE.110086343	10	67.3	4.6

EU-NAM M2 CIAM

	n	Mk. names	chr	pos [cM]	-log10(pval)
parental					
	1	PZE.101147104	1	119.4	6.8
	2	PZE.105054186	5	72.4	8.5
	3	PZA00606.3	6	1.6	7.8
	4	PZE.108104106	8	92.9	4.7
	5	PZE.108133621	8	134.5	6.5
	6	PZE.110049572	10	44.1	9.5
ancestral					
	1	PZE.101144216	1	118.6	7.7
	2	PZE.105063383	5	76	7.2
	3	PZE.106009233	6	12	6.6
	4	PZE.108110343	8	99.5	4.1
	5	PZE.110009558	10	24.6	4.8
	6	PZE.110087849	10	69.5	5.5
bi-allelic					
	1	PZE.101144248	1	117.4	7.3
	2	PZE.103072486	3	51.2	4.2
	3	PZE.105074287	5	76.9	8.1
	4	PZE.107128846	7	128.9	4.2
	5	PZE.108099415	8	89.1	4.7
	6	PZE.110049474	10	45.2	10.1
	7	PZE.110088931	10	68.4	7.8

EU-NAM M2 TUM

	n	Mk. names	chr	pos [cM]	-log10(pval)
parental					
	1	PZE.105019465	5	40.9	4.4
	2	PZE.106102395	6	86.4	15.4
	3	PZE.110013764	10	31.4	8.7
ancestral					
	1	PZE.101144216	1	118.6	4.5
	2	PZE.105017551	5	34.3	4.1
	3	PZE.106101278	6	84.2	17.1
	4	PZE.110049040	10	43.2	8.4
bi-allelic					
	1	PZE.104094429	4	88.6	4.5
	2	PZE.105017975	5	32	5
	3	PZE.106095383	6	79.3	14.7
	4	PZE.107136612	7	145	4.4
	5	PZE.110049406	10	42.8	8.9

EU-NAM M3

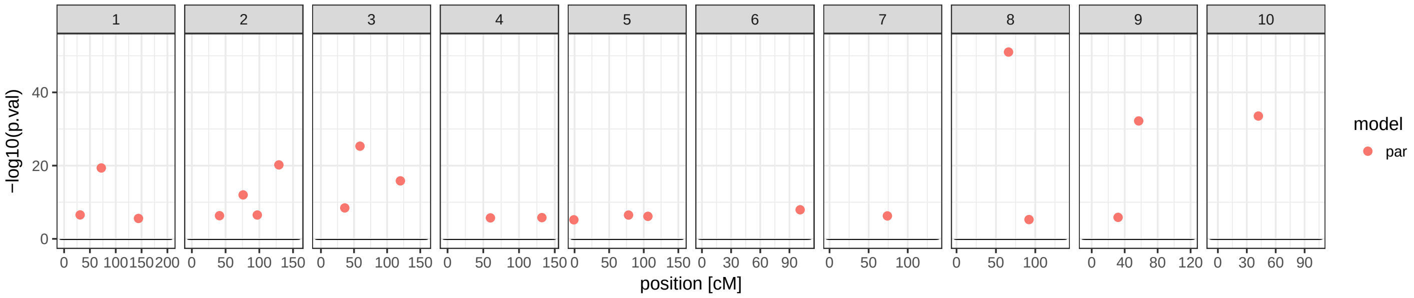
	n	Mk. names	chr	pos [cM]	-log10(pval)
parental					
	1	PZE.101147104	1	119.4	4.9
	2	PZE.104027223	4	56.8	5.1
	3	PZE.105062183	5	72.2	6.7
	4	PZE.106102395	6	86.4	10.8
	5	PZE.108099425	8	90	6.1
	6	PZE.110010098	10	27.5	13.9
	7	PZE.110049922	10	48.1	10.8
ancestral					
	1	PZE.101144216	1	118.6	8.6
	2	PZE.105065789	5	74.5	5.8
	3	PZE.106101278	6	84.2	14.6
	4	PZE.107128534	7	128	4.3
	5	PZE.108133100	8	133.6	6.2
	6	PZE.110049040	10	43.2	12.5
bi-allelic					
	1	PZE.101144248	1	117.4	6.5
	2	PZE.102192367	2	178.2	4.8
	3	PZE.104052802	4	61.2	7.2
	4	PZE.104110016	4	115.6	4.9
	5	PZE.105074287	5	76.9	6.9
	6	PZE.105160757	5	145.1	4.4
	7	PZE.106101278	6	84.2	15.8
	8	PZE.107128846	7	128.9	4.7
	9	PZE.108057679	8	58.2	5.1
	10	PZE.110049474	10	45.2	12.8

EU-NAM M4

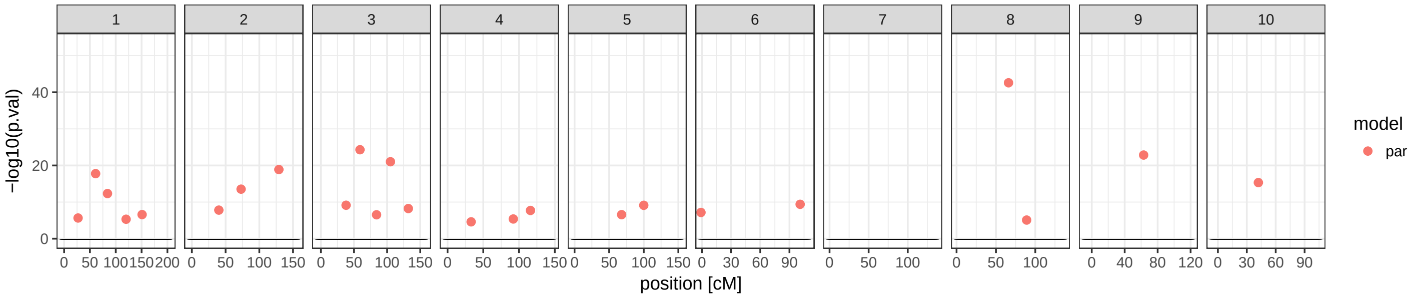
	n	Mk. names	chr	pos [cM]	-log10(pval)
parental					
	1	PZE.101146834	1	119.1	5.5
	2	PZE.104027223	4	56.8	4.9
	3	PZE.105062183	5	72.2	7.7
	4	PZE.106099144	6	85	13.6
	5	PZE.108099425	8	90	5.3
	6	PZE.110048720	10	44.3	13.7
ancestral					
	1	PZE.101144216	1	118.6	8.7
	2	PZE.104029507	4	53.5	6.1
	3	PZE.105065789	5	74.5	6.7
	4	PZE.106098066	6	82.1	17
	5	PZE.107128534	7	128	4.1
	6	PZE.108099425	8	90	4.5
	7	PZE.110049068	10	44.7	12.7
bi-allelic					
	1	PZE.101143233	1	113.9	7.8
	2	PZE.102192367	2	178.2	5.6
	3	PZE.103106593	3	68.7	4.8
	4	PZE.104052802	4	61.2	7.1
	5	PZE.105063758	5	76.1	7.7
	6	PZE.106101278	6	84.2	17.7
	7	PZE.107128846	7	128.9	4.6
	8	PZE.108058577	8	59.1	5.6
	9	PZE.110049474	10	45.2	15.8

US-NAM data - List of detected QTLs

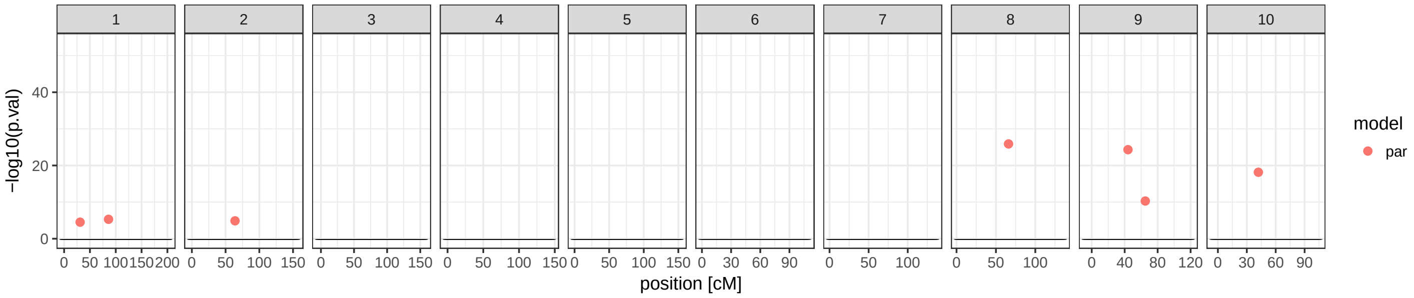
M1 DTA



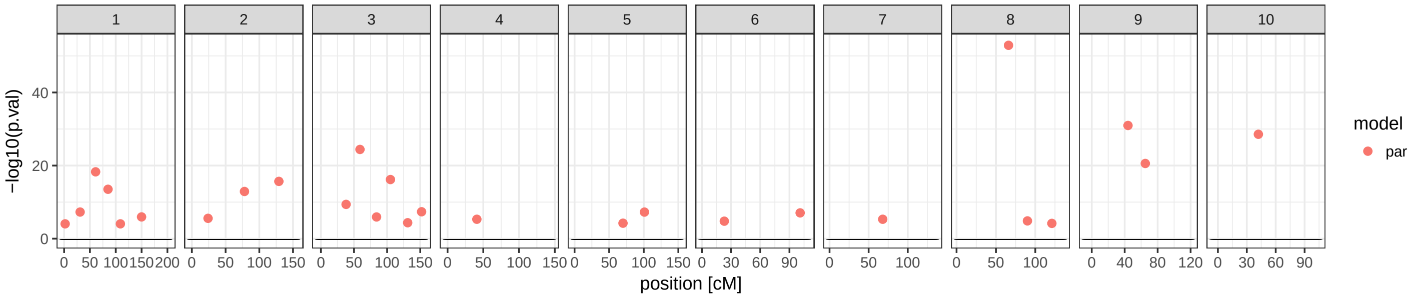
M2 NC DTA



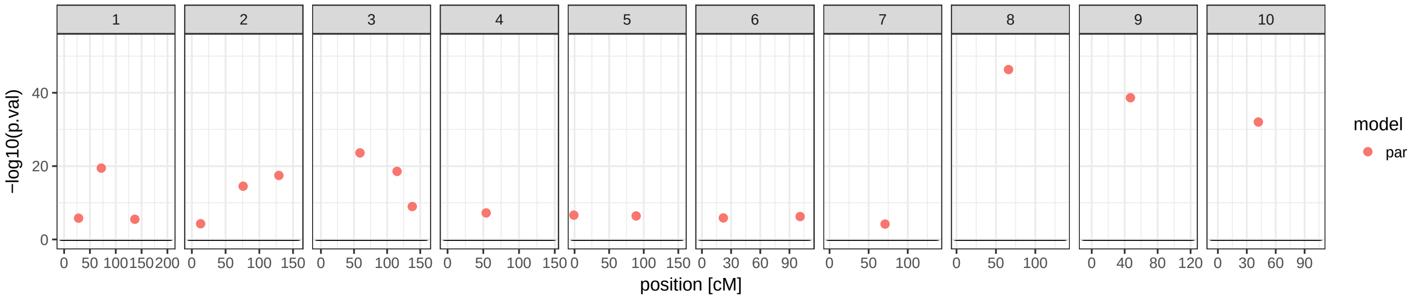
M2 NY DTA



M3 DTA



M4 DTA



US-NAM M1

	n	Mk. names	chr	pos [cM]	-log10(pval)
parental					
	1	m36	1	31	6.5
	2	m77	1	72	19.4
	3	m149	1	144	5.6
	4	m255	2	41	6.3
	5	m290	2	76	12
	6	m311	2	97	6.5
	7	m343	2	129	20.2
	8	m413	3	36	8.5
	9	m436	3	59	25.3
	10	m497	3	120	15.8
	11	m599	4	60	5.7
	12	m671	4	132	5.8
	13	m689	5	-1	5.2
	14	m768	5	78	6.5
	15	m796	5	106	6.2
	16	m947	6	101	7.9
	17	m1032	7	74	6.3
	18	m1162	8	66	51
	19	m1188	8	92	5.3
	20	m1275	9	32	5.9
	21	m1300	9	57	32.2
	22	m1414	10	42	33.5

US-NAM M2 New York

	n	Mk. names	chr	pos [cM]	-log10(pval)
parental					
	1	m32	1	27	5.7
	2	m66	1	61	17.8
	3	m89	1	84	12.3
	4	m125	1	120	5.3
	5	m156	1	151	6.6
	6	m254	2	40	7.8
	7	m287	2	73	13.5
	8	m343	2	129	18.9
	9	m415	3	38	9.1
	10	m436	3	59	24.3
	11	m461	3	84	6.5
	12	m482	3	105	21
	13	m509	3	132	8.2
	14	m572	4	33	4.6
	15	m631	4	92	5.4
	16	m655	4	116	7.7
	17	m758	5	68	6.6
	18	m790	5	100	9.1
	19	m845	6	-1	7.2
	20	m947	6	101	9.4
	21	m1162	8	66	42.6
	22	m1185	8	89	5.1
	23	m1306	9	63	22.9
	24	m1414	10	42	15.3

US-NAM M2 North Carolina

	n	Mk. names	chr	pos [cM]	-log10(pval)
parental					
	1	m36	1	31	4.5
	2	m91	1	86	5.3
	3	m278	2	64	4.9
	4	m1162	8	66	25.9
	5	m1287	9	44	24.3
	6	m1308	9	65	10.3
	7	m1414	10	42	18.2

US-NAM M3

	n	Mk. names	chr	pos [cM]	-log10(pval)
parental					
	1	m7	1	2	4.1
	2	m36	1	31	7.3
	3	m66	1	61	18.3
	4	m90	1	85	13.5
	5	m114	1	109	4.1
	6	m155	1	150	6
	7	m238	2	24	5.6
	8	m292	2	78	12.9
	9	m343	2	129	15.7
	10	m415	3	38	9.4
	11	m436	3	59	24.4
	12	m461	3	84	5.9
	13	m482	3	105	16.2
	14	m508	3	131	4.4
	15	m529	3	152	7.4
	16	m580	4	41	5.3
	17	m760	5	70	4.2
	18	m791	5	101	7.3
	19	m869	6	23	4.8
	20	m947	6	101	7.1
	21	m1026	7	68	5.3
	22	m1162	8	66	52.9
	23	m1186	8	90	4.9
	24	m1217	8	121	4.2
	25	m1287	9	44	31
	26	m1308	9	65	20.6
	27	m1414	10	42	28.6

US-NAM M4

	n	Mk. names	chr	pos [cM]	-log10(pval)
parental					
	1	m33	1	28	5.8
	2	m77	1	72	19.5
	3	m142	1	137	5.5
	4	m227	2	13	4.3
	5	m290	2	76	14.5
	6	m343	2	129	17.5
	7	m436	3	59	23.6
	8	m492	3	115	18.6
	9	m515	3	138	9
	10	m593	4	54	7.2
	11	m689	5	-1	6.6
	12	m779	5	89	6.4
	13	m868	6	22	5.9
	14	m947	6	101	6.3
	15	m1029	7	71	4.2
	16	m1162	8	66	46.3
	17	m1290	9	47	38.6
	18	m1414	10	42	32

S3: QTL additive effects allelic series

EU-NAM chr 6 82.1 cM

Table 5.5: QTL additive effects and standard deviations

	β_{M1}	β_{M4-E1}	β_{M4-E2}	sd(β_{M1})	sd(β_{M4-E1})	sd(β_{M4-E2})	β/sd(β)(M1)	β/sd(β)(M4-E1)	β/sd(β)(M4-E2)
UH007	0.00	0.00	0.00	0.00	0.00	0.00			
EZ5	0.00	0.00	0.00	0.00	0.00	0.00			
UH009	0.00	0.00	0.00	0.00	0.00	0.00			
D152	-4.07	-1.19	-7.20	0.71	0.83	0.83	-5.75	-1.42	-8.63
EC49A	-4.07	-1.19	-7.20	0.71	0.83	0.83	-5.75	-1.42	-8.63
F03802	-4.07	-1.19	-7.20	0.71	0.83	0.83	-5.75	-1.42	-8.63
F2	-4.07	-1.19	-7.20	0.71	0.83	0.83	-5.75	-1.42	-8.63
F283	-4.07	-1.19	-7.20	0.71	0.83	0.83	-5.75	-1.42	-8.63
UH006	-4.07	-1.19	-7.20	0.71	0.83	0.83	-5.75	-1.42	-8.63
DK105	-4.07	-1.19	-7.20	0.71	0.83	0.83	-5.75	-1.42	-8.63
F64	-6.13	0.52	-12.81	2.13	2.40	2.40	-2.88	0.22	-5.34
EP44	-0.94	5.65	10.19	5.72	5.63	5.63	-0.17	1.00	1.81

US-NAM chr 8 67 cM

Table 5.6: QTL additive effects and standard deviations

	β_{M1}	β_{M4-E1}	β_{M4-E2}	sd(β_{M1})	sd(β_{M4-E1})	sd(β_{M4-E2})	β/sd(β)(M1)	β/sd(β)(M4-E1)	β/sd(β)(M4-E2)
B73	0.00	0.00	0.00	0.00	0.00	0.00			
B97	0.04	0.09	0.08	0.14	0.14	0.14	0.27	0.66	0.56
CML103	0.29	0.55	0.21	0.16	0.16	0.16	1.83	3.46	1.31
CML228	0.36	0.43	0.45	0.22	0.22	0.22	1.61	1.92	2.03
CML247	0.54	0.45	0.55	0.17	0.20	0.20	3.12	2.28	2.83
CML277	-0.05	0.26	-0.17	0.22	0.18	0.18	-0.20	1.44	-0.94
CML322	0.41	0.68	0.60	0.16	0.18	0.18	2.62	3.91	3.45
CML333	0.60	0.69	0.62	0.17	0.19	0.19	3.55	3.65	3.30
CML52	0.59	0.47	0.52	0.22	0.19	0.19	2.65	2.49	2.73
CML69	0.61	0.67	0.64	0.16	0.16	0.16	3.70	4.16	3.98
Hp301	0.08	0.20	0.19	0.14	0.14	0.14	0.54	1.43	1.31
IL14H	-1.09	-0.82	-1.32	0.14	0.14	0.14	-7.56	-5.80	-9.35
Ki11	-0.11	-0.06	-0.52	0.19	0.20	0.20	-0.54	-0.29	-2.61
Ki3	0.17	0.44	-0.07	0.20	0.29	0.29	0.86	1.51	-0.23
Ky21	0.38	0.32	0.46	0.12	0.13	0.13	3.09	2.43	3.49
M162W	0.03	-0.03	-0.11	0.13	0.16	0.16	0.25	-0.17	-0.72
M37W	0.20	0.17	0.00	0.16	0.15	0.15	1.24	1.14	0.02
Mo18W	0.62	0.87	0.53	0.18	0.20	0.20	3.54	4.32	2.64
MS71	-1.06	-0.54	-1.17	0.14	0.13	0.13	-7.54	-4.13	-8.99
NC350	-0.27	0.56	-0.20	0.18	0.18	0.18	-1.45	3.11	-1.10
NC358	0.04	0.40	-0.05	0.12	0.14	0.14	0.32	2.81	-0.34
Oh43	-0.02	-0.07	0.12	0.13	0.14	0.14	-0.14	-0.50	0.81
Oh7B	0.15	0.22	0.20	0.15	0.16	0.16	1.01	1.33	1.24
P39	-1.32	-0.99	-1.56	0.18	0.18	0.18	-7.46	-5.48	-8.69
Tx303	0.48	0.64	0.44	0.17	0.20	0.20	2.87	3.24	2.23
Tzi8	-0.11	-0.23	-0.44	0.15	0.20	0.20	-0.74	-1.17	-2.18

US-NAM chr 9 47 cM

Table 5.7: QTL additive effects and standard deviations

	β_{M1}	β_{M4-E1}	β_{M4-E2}	sd(β_{M1})	sd(β_{M4-E1})	sd(β_{M4-E2})	β/sd(β)(M1)	β/sd(β)(M4-E1)	β/sd(β)(M4-E2)
B73	0.00	0.00	0.00	0.00	0.00	0.00			
B97	-0.05	0.01	0.02	0.12	0.13	0.13	-0.40	0.06	0.18
CML103	0.47	0.25	0.69	0.13	0.16	0.16	3.62	1.57	4.44
CML228	0.93	0.67	1.20	0.22	0.23	0.23	4.26	2.95	5.30
CML247	0.20	-0.29	0.71	0.16	0.19	0.19	1.25	-1.52	3.70
CML277	1.43	0.78	1.94	0.21	0.18	0.18	6.85	4.41	11.00
CML322	0.53	0.30	0.79	0.14	0.17	0.17	3.73	1.79	4.65
CML333	0.70	0.26	1.08	0.16	0.19	0.19	4.39	1.39	5.71
CML52	0.48	0.28	0.82	0.21	0.18	0.18	2.30	1.55	4.62
CML69	0.03	-0.36	0.18	0.14	0.15	0.15	0.19	-2.42	1.23
Hp301	0.46	0.51	0.41	0.13	0.14	0.14	3.42	3.56	2.81
IL14H	-0.05	0.15	0.18	0.14	0.14	0.14	-0.37	1.07	1.30
Ki11	1.10	0.28	1.85	0.20	0.21	0.21	5.43	1.33	8.70
Ki3	0.12	-0.19	0.20	0.18	0.27	0.27	0.65	-0.69	0.76
Ky21	0.23	0.29	0.18	0.10	0.13	0.13	2.23	2.20	1.35
M162W	0.36	-0.11	0.65	0.14	0.18	0.18	2.56	-0.62	3.66
M37W	-0.04	-0.02	0.15	0.15	0.16	0.16	-0.24	-0.13	0.96
Mo18W	0.16	-0.36	0.60	0.17	0.21	0.21	0.93	-1.72	2.82
MS71	0.31	0.04	0.46	0.13	0.13	0.13	2.42	0.32	3.58
NC350	-0.48	-0.57	-0.09	0.15	0.18	0.18	-3.28	-3.17	-0.51
NC358	-0.07	-0.29	0.36	0.12	0.15	0.15	-0.54	-2.01	2.47
Oh43	0.14	0.06	0.22	0.13	0.14	0.14	1.07	0.45	1.51
Oh7B	0.28	0.38	0.36	0.14	0.17	0.17	2.03	2.20	2.08
P39	0.28	0.55	0.50	0.17	0.19	0.19	1.62	2.85	2.57
Tx303	0.21	-0.09	0.51	0.15	0.20	0.20	1.42	-0.42	2.52
Tzi8	0.72	0.45	0.92	0.13	0.18	0.18	5.42	2.45	5.02

US-NAM chr 10 42 cM

Table 5.8: QTL additive effects and standard deviations

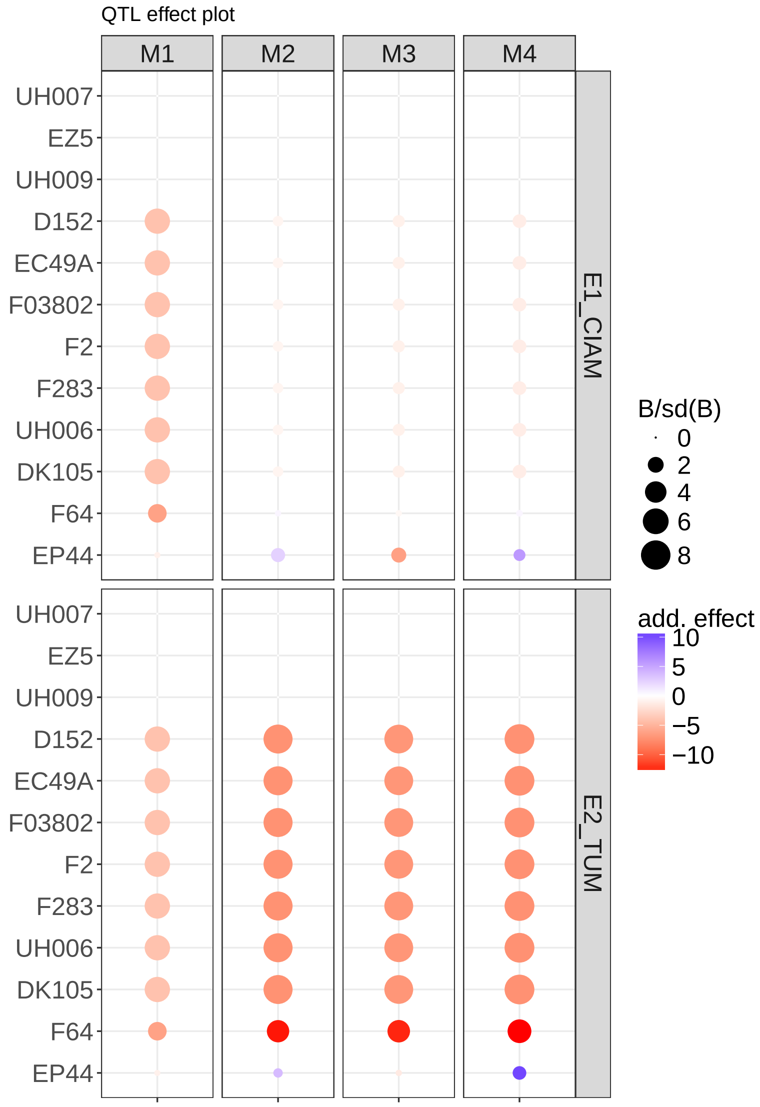
	β_{M1}	β_{M4-E1}	β_{M4-E2}	sd(β_{M1})	sd(β_{M4-E1})	sd(β_{M4-E2})	β/sd(β)(M1)	β/sd(β)(M4-E1)	β/sd(β)(M4-E2)
B73	0.00	0.00	0.00	0.00	0.00	0.00			
B97	-0.05	0.04	-0.20	0.13	0.13	0.13	-0.42	0.30	-1.49
CML103	-0.21	-0.20	-0.36	0.13	0.16	0.16	-1.61	-1.21	-2.21
CML228	1.34	0.66	2.10	0.21	0.22	0.22	6.27	2.98	9.41
CML247	-0.14	0.01	-0.47	0.16	0.19	0.19	-0.87	0.07	-2.43
CML277	2.03	1.39	2.99	0.22	0.18	0.18	9.22	7.54	16.26
CML322	-0.45	-0.26	-0.50	0.14	0.17	0.17	-3.22	-1.57	-2.97
CML333	0.16	-0.12	0.14	0.16	0.19	0.19	1.02	-0.66	0.74
CML52	0.29	0.41	0.22	0.20	0.19	0.19	1.45	2.20	1.21
CML69	-0.12	0.16	-0.26	0.14	0.16	0.16	-0.85	1.04	-1.63
Hp301	0.20	-0.01	0.41	0.14	0.15	0.15	1.44	-0.04	2.79
IL14H	-0.05	0.50	-0.02	0.13	0.14	0.14	-0.37	3.62	-0.15
Ki11	1.48	1.09	1.76	0.19	0.20	0.20	7.65	5.37	8.69
Ki3	-0.43	-0.55	-0.44	0.18	0.28	0.28	-2.38	-1.97	-1.57
Ky21	0.23	0.24	0.33	0.10	0.13	0.13	2.22	1.76	2.44
M162W	-0.08	-0.03	-0.37	0.14	0.18	0.18	-0.61	-0.20	-2.12
M37W	-0.17	-0.17	-0.23	0.14	0.15	0.15	-1.22	-1.13	-1.56
Mo18W	-0.21	-0.14	-0.42	0.17	0.21	0.21	-1.23	-0.65	-1.98
MS71	0.17	0.04	0.26	0.14	0.14	0.14	1.21	0.31	1.86
NC350	-0.20	-0.02	-0.07	0.15	0.18	0.18	-1.31	-0.13	-0.43
NC358	0.22	0.27	0.26	0.12	0.14	0.14	1.87	1.90	1.81
Oh43	-0.12	-0.05	-0.15	0.13	0.15	0.15	-0.91	-0.33	-1.01
Oh7B	-0.01	0.18	0.09	0.13	0.16	0.16	-0.07	1.11	0.53
P39	-0.15	0.00	-0.16	0.16	0.18	0.18	-0.90	0.02	-0.89
Tx303	-0.17	-0.04	-0.15	0.14	0.20	0.20	-1.19	-0.20	-0.75
Tzi8	-0.03	-0.08	-0.35	0.14	0.19	0.19	-0.23	-0.40	-1.82

S4: QTL additive effects allelic series

Comparison of the additive effect allelic series between M1, M2, M3 and M4 and two environments for two individual QTL positions. The upper panel contain the first environment results and the lower the one from the second environment.

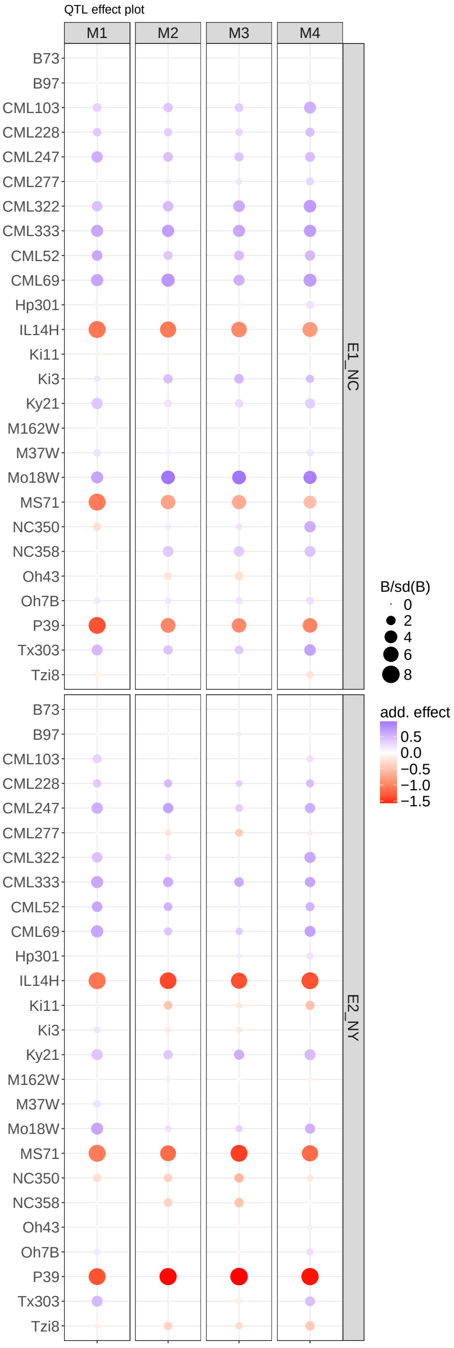
EU-NAM ancestral model chr 6 82.1 cM

Comparison of the additive effect allelic series between M1, M2, M3 and M4 and two environments



US-NAM parental model chr 8 67 cM

Comparison of the additive effect allelic series between M1, M2, M3 and M4 and two environments



Chapter 6

General Discussion

6.1 Is it still useful to detect QTLs (in MPPs)?

The objective of this thesis is to propose a statistical methodology to detect QTLs using multi-parent populations (MPPs). Since Mendel (1866) and the development of (plant) genetics, finding links between the genes and the phenotypic variation has been a central research question. The idea that continuous phenotypic variation results from the action of several Mendelian factors is an early hypothesis put forward by Fisher (1918) who was one of the first to establish the notion of quantitative trait locus (QTL). According to Doerge (2002, p. 44), "A QTL is a region of any genome that is responsible for variation in the quantitative trait of interest. Developments and knowledge in several areas such as DNA structure Watson et al. (1953), sequencing technologies Sanger et al. (1977), and bioinformatics allowed to perform increasingly complex QTL experiments and so, refine our understanding of the underlying genetic mechanisms explaining the phenotypic variation. The major features of the genetic models are often summarized in the concept of genetic architecture stating that: a) the phenotypic variation are due to the joint action of multiple QTLs with effect size varying from very small to large; b) the QTLs can act in different ways (additive, dominant, epistatic, influenced by the environment); c) the QTLs can be distributed randomly along the genome or can also segregate together (Wu et al., 2007; Mackay, 1996). Therefore, the concept of QTL is important for the quantitative genetics theory.

The usefulness of QTL detection for breeding improvements is however debated. Some QTLs with large effects on traits controlled by few QTLs were successfully introgressed in species, improving for example disease resistance (Bernardo, 2016; Leung et al., 2015). In several cases, it was possible to identify genes associated to the QTLs (Cobb et al., 2019). However, according to Bernardo (2016), the main limitation that prevents from using QTLs is their interaction with different genetic background and/or the environment.

MPPs are useful to address these issues because MPPs allow detect QTLs with more consistent effects by testing simultaneously the QTL effects in several genetic backgrounds (Blanc et al., 2006; Li et al., 2011; Bardol et al., 2013). In that case, the estimation of the QTL effects should be more reliable. The framework we proposed allows to test if a QTL allele interacts with the genetic background or if it has a consistent effect across backgrounds. Such a test can be done by comparing the estimated effects obtained with a so-called cross-specific model with the ones obtained with the parental or the ancestral model. In that case, it is possible to test if the estimated parental (ancestral) alleles have the same effect in all the crosses where they segregate or if the cross-specific allelic effects diverge.

Concerning the environmental influence, the detection of QTL in MPPs characterized in multiple environments (ME), as we proposed in chapter five, is also a progress. The use of an MPP GxE model allows to evaluate the QTL genetic variability between the crosses and between the environments. Using such a methodology, the breeder can determine if QTLs have allelic effects that are consistent in several genetic backgrounds and across the environments. Those QTLs are certainly more stable and so, they should be introgressed with priority.

Beside genetic background and environmental interactions, the quantitative nature of complex traits like yield can also explain the difficulty to improve plants using QTLs. Indeed, when a trait is controlled by many independent QTLs, it is difficult to build an ideotype with all favourable alleles because the frequency of such a plant can be extremely low (Bernardo, 2016). For highly quantitative traits, techniques like genomic prediction (GP) can be useful. Unlike QTL mapping, the main objective of GP is not to identify regions of the genome showing statistical association with the trait but to describe the phenotypes using the whole genomic information (Meuwissen et al., 2001). GP allows to predict genetic trait values of phenotyped and unphenotyped lines and therefore to perform a recurrent selection on the best genotypes. Many studies have shown the superiority of the GP models over the QTL models to predict traits (Alimi, 2016; Bustos-Korts et al., 2016).

In presence of highly quantitative traits, we can even consider that the GP models are more informative than the QTL models. Indeed, authors like Breiman et al. (2001), support the idea that a better prediction accuracy is associated with more reliable assumptions concerning the underlying process generating the data. For example, for highly quantitative traits, the GP model assumption considering that the phenotypic variation is due to many QTLs with a small effect seems more accurate than the one of a QTL model assuming the effect of 10-20 QTLs. For traits that are controlled by a reasonable amount of QTLs with a large or medium effects, the QTL model can still give a good description. The QTL models can also be combined with gene network models to improve the understanding of the underlying genetics (Alimi, 2016). Therefore, in some cases, the QTLs models keep a reasonable information predictive power.

In table 6.1, we compared the prediction accuracy of QTL models (parental, ancestral, and bi-allelic) and GP models (G-BLUP and Bayes C) in the EU-NAM Dent panel for dry matter yield ($\frac{dt}{ha}$). The genetic map contained 6285 markers. We performed 10 replicates of a 5-folds cross-validation (CV) procedure. The partition of the data between the estimation and the test set was done within cross. Therefore, for all models we used within cross and between cross information to estimate the marker effects. We measured

Table 6.1: Comparison of the prediction accuracies between QTL and GP models for dry matter yield in the EU-NAM Dent population

	QTL models			GP models	
crosses	parental	ancestral	bi-allelic	G-BLUP	Bayes C
CFD02	0.35	0.42	0.45	0.50	0.50
CFD03	0.35	0.37	0.37	0.46	0.46
CFD04	0.21	0.26	0.27	0.44	0.45
CFD05	0.24	0.28	0.25	0.37	0.37
CFD06	0.53	0.52	0.47	0.58	0.58
CFD07	0.33	0.33	0.26	0.36	0.35
CFD09	0.33	0.33	0.38	0.27	0.27
CFD10	0.36	0.31	0.33	0.49	0.49
CFD11	0.16	0.21	0.23	0.19	0.19
CFD12	0.38	0.35	0.37	0.40	0.40
average	0.32	0.34	0.34	0.41	0.41

the within cross accuracies using the Pearson correlation between the observed test set values and the model predictions. In table 6.1, we notice that the GP results are superior but the differences with the QTL models is not huge. In some crosses like CFD09, the accuracy of the QTL based prediction was superior to the one obtained with GP. In that case, the LD pattern between crosses can be different, which explain the low accuracy of the GP models. On the other hand, the QTL models can detect and use selective QTLs that show some consistency between crosses, which helps to improve the prediction accuracy.

The GP approach also has some limitations. For example, GP requires important genotyping efforts Bernardo (2016). GP is not pertinent for crops with low density genetic maps (Muchero et al., 2009; Xu et al., 2011). For those crops, QTL mapping still represents a first possibility to characterize the genetic effects. The cost factor should also be considered when we compare QTL and GP. Indeed, in a breeding and selection perspective, the genotyping of few QTLs is less expensive than the characterisation of thousands of markers (Cobb et al., 2019). Therefore, focusing on a limited number of QTL positions, we could produce and screen a larger number of candidate lines. The genotyping cost of GP could however be reduced by using marker imputation (Crossa et al., 2014). New technologies like CRISPR-CAS9 could also facilitate the introgression of QTLs making marker assisted selection easier Cobb et al. (2019).

From a didactic point of view, the QTL models can be more intuitive than the infinitesimal models generally assumed by the GP methods. In the end, the notion of QTL stays a fundamental concept for quantitative genetics theory that is widely used even in GP (Habier et al., 2011). Apparently, no genetic model is purely infinitesimal, GP methods attempting to weight differently regions of the genome like the Bayes B model (Meuwissen et al., 2001) or the multi-BLUP model (Speed & Balding, 2014) have shown their usefulness. These models assume the presence of (QTL) genome regions that have a larger influence on the trait.

To conclude this section, we summarize the reasons that justify the research on QTL detection and that support the use of MPPs to improve this operation. First, for moderately quantitative traits, the QTL models still represent descriptions with good informative and predictive abilities. From a breeding point of view, in several situations, QTLs with large or medium effects could be introgressed with success. Second, MAS based on QTL models can be cheaper than selection using GP. Third, for several crops that are not extensively genetically characterized, the QTL mapping approach is a first attempt to describe the genetic effects. Fourth, the notion of QTL stays a fundamental concept for quantitative genetics theory that is widely used even in GP. From a didactic point of view, the QTL models allow to get a first image of the genetic effects at the genome level. Finally, we want to underline the usefulness of the MPPs to detect QTLs with consistent effects, a major limitation of single cross QTL experiments. The MPP-ME experiments allow to estimate the consistency of the QTL effects in different genetic backgrounds and between environments. MPPs can therefore detect QTLs with more stable effects. All these reasons justify the use of MPPs for QTL detection in a breeding perspective, which is the topic of this thesis.

6.2 The proposed framework contrasted with other approaches

The main contribution of this thesis is to propose a statistical framework to detect QTLs in MPPs. The use of MPPs is growing in popularity and several statistical procedures have already been proposed to analyse MPP QTL experiments. In chapter two, we started with the following general model:

$$y_{ij} = \mu_j + Q_{ij} + g_{ij} + e_{ij} \tag{6.1}$$

where, y_{ij} was the phenotypic value of line i coming from cross j. μ_j was the contribution of cross j. Q_{ij} was the tested QTL effect and g_{ij} the polygenic effect of other QTLs elsewhere in the genome. And e_{ij} represented the environmental error term. Model 6.1, can be rewritten in matrix notation:

$$\boldsymbol{y} = \boldsymbol{X_c\beta_c} + \sum_{i=1}^{n_{QTL}} \boldsymbol{X_{Q_i}\beta_{Q_i}} + \boldsymbol{r} \tag{6.2}$$

With $\boldsymbol{y}$ being the vector of phenotypic values, $\boldsymbol{X_c\beta_c}$ represents the cross-specific intercepts, $\sum_{i=1}^{n_{QTL}} \boldsymbol{X_{Q_i}\beta_{Q_i}}$ is a number of QTLs, and $\boldsymbol{r}$ is the vector of residual terms.

6.2.1 The QTL term in the linear models

MPPs allow to define the QTL effect ($\boldsymbol{X_{Q_i}\beta_{Q_i}}$ in model 6.2) with a larger flexibility than in bi-parental crosses and association panels. The bi-parental crosses are limited to bi-allelic QTLs with allele origin determined by identical by descent (IBD) estimates tracing back to one of the parents (Rakshit et al., 2012; Broman et al., 2003). In association panels, the QTLs are generally also bi-allelic with alleles corresponding to the SNP marker classes, which represent identical by state (IBS) estimates (Astle & Balding, 2009).

In MPPs, it is possible to define QTLs with multiple alleles coming from the parents tracing back their origin using IBD estimates. We can assume that the parental allelic effects have a consistent effect in the whole MPP or that they interact with the cross genetic background (Blanc et al., 2006). In MPPs, we can also use the IBS marker information to cluster parents based on genetic similarity and infer a reduced number of ancestral QTL alleles (Leroux et al., 2014; Jansen et al., 2003). Finally, it is possible to treat MPPs as association panels and to use the same bi-allelic QTL metric corresponding to the IBS SNP marker scores.

These different ways to define the QTL term ($\boldsymbol{X_{Q_i}\beta_{Q_i}}$) are the central element of our methodology. They correspond to different ways to define the genetic relatedness at the QTL position. The genotypes are grouped in the same QTL allelic group because they share a common parent, ancestor or the same SNP marker score. The linear version of the QTL models we proposed already existed in the literature. The cross-specific and the parental models correspond to the disconnected and connected models implemented by Jourjon et al. (2005) and illustrated in Blanc et al. (2006). The ancestral model is similar to the LDLA model proposed by Leroux et al. (2014) and used by Bardol et al. (2013). Finally, the bi-allelic model corresponds to models used in association panels like model B from Würschum (2012).

In chapter three, we compared the parental, ancestral and bi-allelic models using subsets of the EU-NAM population (Bauer et al., 2013) characterized by different levels of genetic relatedness. We hypothesized that more parsimonious models like the ancestral and the bi-allelic one should perform better in MPPs with a reduced genetic basis. Indeed, if the parents are genetically similar we can expect that the number of segregating alleles will be reduced. In that case, the use of a model assuming a reduced number of QTL allele should increase the detection power.

We performed CV to evaluate our hypotheses but we could not verify them. It is possible that the differences in terms of genetic similarity between our populations were not large enough to verify our assumptions. Another possibility is that the MPPs are mostly characterized by complex allelic series that can be more easily described by cross-specific QTL effects. The parental, ancestral and bi-allelic models assume that the QTL allelic effects are consistent between crosses. However, phenomena like: a) multiple alleles, b) difference of allele frequency between crosses, c) cross-specific differences of linkage disequilibrium (LD) between markers and QTL, or d) some interaction with the cross genetic background, can explain the difficulty to detect consistent QTL effects (Steinhoff et al., 2012; Blanc et al., 2006). In many studies, the models assuming the cross-specificity of the QTL effects gave better description of the phenotypic variance than models assuming consistent QTL effects across the MPP (Li et al., 2005; Steinhoff et al., 2011; Würschum, 2012; Giraud et al., 2014).

In chapter four, we evaluated all the QTL models (cross-specific, parental, ancestral, bi-allelic) by simulation. As expected, we detected with a larger power the simulated QTLs for which the genetic assumption (cross-specific, parental, ancestral, bi-allelic) corresponded to the one of the statistical model. When the QTLs were simulated with an equal chance to be cross-specific, parental, ancestral or bi-allelic, the models performed equally well. This is generally the situation we face in practice. The user does not know a priori if the QTL has a cross-specific effect, or if it is characterized by a reduced number of ancestral alleles. For that reason, in chapter two, we proposed a QTL detection procedure aiming at describing each QTL with the type of effect (cross-specific, parental, ancestral, bi-allelic) that corresponded the most.

6.2.2 The QTL term in the IBD-variance mixed model framework

The use of different types of QTL effects corresponding to different definitions of the genetic relatedness is a central idea present in all approaches to detect QTL in MPPs (linear model, mixed model, or Bayesian approach). The IBD-variance component approach is a very useful theoretical framework to illustrate the question of the genetic relatedness at the QTL location. Haseman & Elston (1972) showed, from a theoretical point of view, that individuals with similar phenotypes were more likely to share alleles IBD. This led statisticians to propose mixed models where the QTL term Q_{ij} of model 6.1 was considered as random with an associated variance covariance (VCOV) structure representing the IBD relationship between the genotypes (Xu & Atchley, 1995). Therefore, $Q_{ij} \sim N(0, \mathbf{\Pi}\sigma_q^2)$ with π_{ij}, the individual elements of $\mathbf{\Pi}$, representing an (IBD) estimate of genetic relatedness between genotypes i and j. Such an approach was successfully implemented in plant MPPs (Xie et al., 1998; Xu, 1998; Crepieux et al., 2004).

One of the main advantages of the IBD-variance component approach is its capacity to handle fragmented MPP designs. When the population pedigree is known, and the number of crosses and/or parents is not too large (e.g. < 20), it is possible to make assumptions about the number of alleles, their origin and to properly estimate the genetic effect of the founder lines (Würschum, 2012). In many MPPs however, the combination of material makes the use of fixed QTL effects difficult because the number of parameters to estimate increases with the number of crosses and/or parents. The information to estimate these parameters can also be reduced when the crosses are too small (Xie et al., 1998; Xu, 1998). This situation can be encountered in outbred populations (Wu et al., 2007) or in MPPs coming from breeding programs (Crepieux et al., 2005). In such a situation, the difficulty to specify a finite set of alleles makes the use of random terms to model the QTL effect attractive because it only requires estimating a global variance term associated to the QTL effect.

The VCOV matrix $\mathbf{\Pi}$ can be any interesting measurement of genetic relatedness (Xu, 1996). In the IBD-variance component framework, Xie et al. (1998) proposed to model $\mathbf{\Pi}$ accounting only for the full-sibs relationship, which can be compared to our cross-specific model. Crepieux et al. (2004) extended the methodology proposed by Xie et al. (1998) by integrating the half-sibs relatedness due a shared parent between crosses. This is similar to the parental model. Crepieux et al. (2004) also proposed a VCOV integrating the relatedness between parents due to ancestral relationship and a shared pedigree. This model is similar to the ancestral model we proposed.

6.2.3 The QTL term in Bayesian approaches

The question of the genetic relatedness is also central for the methods based on a Bayesian approach. In that framework, methods allowing to use complex pedigree relationship to determine the genetic relatedness could be integrated in the estimation procedure. Bink et al. (2002) used an algorithm to describe the gene flow between the founders and the non-founders to determine the genotype of bi-allelic QTLs. Yi & Xu (2001) developed a similar approach for multi-allelic QTLs. In that case, they determined the allele origin of the final lines in the founder using the pedigree relationships. Those methods were integrated in the Bayesian modelling procedure and allowed to take advantage of using complex pedigree relationship to detect the

QTLs. Ter Braak et al. (2010) proposed a latent ancestral allele model (LAAM) method to decompose an IBD relationship matrix into a matrix that links genotypes to independent ancestral alleles. The LAAM allows to obtain a probabilistic distribution of the ancestral alleles in the population and was successfully used to sample ancestral allele origins in the Bayesian approach used by Bink et al. (2012) and van Eeuwijk et al. (2010b).

The philosophy of the LAAM is similar to the reduction of the parent IBD matrix done in the ancestral model using clusthaplo (Leroux et al., 2014). However, the LAAM allows to obtain probabilistic estimations of the ancestral allele origin when clusthaplo proposes a discrete classification. In the Bayesian context, the question of the allelic configuration at the QTL could also be treated as a random parameter determined by the model. Such an approach was proposed by Jannink & Wu (2003) who implemented a model where the number of alleles at the QTL position was treated as a random parameter. Their method also determined possible assignments of the parents to allele classes. This is an elegant way to explore the space of possible allelic configurations at the QTL location.

6.2.4 The multi-QTL effect model: a larger flexibility to define genetic relatedness at the QTL position

The definition of the genetic relatedness is a central question in all MPP QTL detection methods. These methods generally use a diversity of genetic relatedness measurement at the QTL position. Each metric can have its own justification and can be adapted to the description of a specific QTL. The flexibility offered by the method of Jannink & Wu (2003), which explores the space of possible allelic configuration, seems ideal. However, such a method can reach some limit in the number of alleles that can be effectively distinguished (Wu & Jannink, 2004). The Bayesian methods can also be computationally intensive (van Eeuwijk et al., 2010b).

In chapter two, we proposed a multi-QTL effect (MQE) model to gain in flexibility for the definition of the genetic relatedness at the QTL position. Indeed, so far, the proposed QTL detection methods generally restricted the model to a single type of QTL effect, keeping the same type of incidence matrix for all tested loci (e.g. Giraud et al. (2014)). We proposed a procedure to define a multi-QTL model where different loci could be modelled by different types of QTL effects. As we could see in chapter three, in some situations, the MQE model allowed to obtain a better prediction of the phenotypic variation. This is probably due to the capacity of the MQE model to better reflect the variety of allelic configurations present across loci. The MQE models were fitted in the linear model framework to keep a reasonable computational time.

6.2.5 Alternative ways to define the genetic relatedness at the QTL position

Since genetic relatedness is an important element for QTL detection (in MPPs), it could be useful to define the detection of QTLs as the search for statistical association between a measurement of genetic relatedness and the phenotypic variance. The IBD-variance component method is an approach that underlines the importance of the genetic relatedness definition for QTL detection in MPPs. Indeed, in that case, the central task is to define the VCOV matrix $\mathbf{\Pi}$ associated with the random QTL term Q_{ij} from model 6.1. We explored the idea to replace $\mathbf{\Pi}$ by a local genetic similarity matrix (GSM) build with dense SNP marker information to approximate the local IBD structure around the QTL position. The possibility to use dense SNP marker scores to approximate the IBD structure has been theoretically established since long time (Toro et al., 2011). Using SNP marker to locally define coancestry is supposed to benefit from the increasing amount of SNP markers available (Speed & Balding, 2015). According to Powell et al. (2010), the characterisation of genetic relatedness by dense SNP markers scores can be a simple method to integrate in one measurement several sources of relatedness. It allows to summarize recent relationships and more subtle variation due to a common ancestor far away in the pedigree.

We performed QTL scans for plant height in the EU-NAM Dent population using 32K markers. At each tested QTL position, we calculated a local GSM ($\mathbf{\Pi}$) with the formula from Astle & Balding (2009) using one, five or ten markers. We tested the QTL effect after correcting for the polygenic background removing the markers from the scanned chromosome (Rincent et al., 2014) using the R package SKAT (Wu et al., 2011). The local GSM scan with one marker corresponded to a standard GWAS analysis using the bi-allelic IBS SNP marker information. By increasing the number of markers used to calculate the local GSM, we increased the amount of local IBD information contained in $\mathbf{\Pi}$.

As we can see in figure 6.1, the differences between the QTL profiles, using one, five or ten markers GSMs, are small. Therefore, the information contained in a single SNP marker scores seems to be enough to describe the QTL genetic relatedness. The addition of markers on the left and the right of the tested position to capture the local IBD structure by stretches of IBS was unnecessary, at least in this example. It is possible that the marker density we used (26 markers/cM) was not large enough to distinguish variation. We can also imagine that our QTL segments where too short to infer properly IBD relationship using local GSMs and that a SNP based GSM could only be discriminant on medium or large segments (Speed & Balding, 2015, 2014).

The IBD framework developed by Zheng et al. (2014) and Zheng et al. (2015) is another

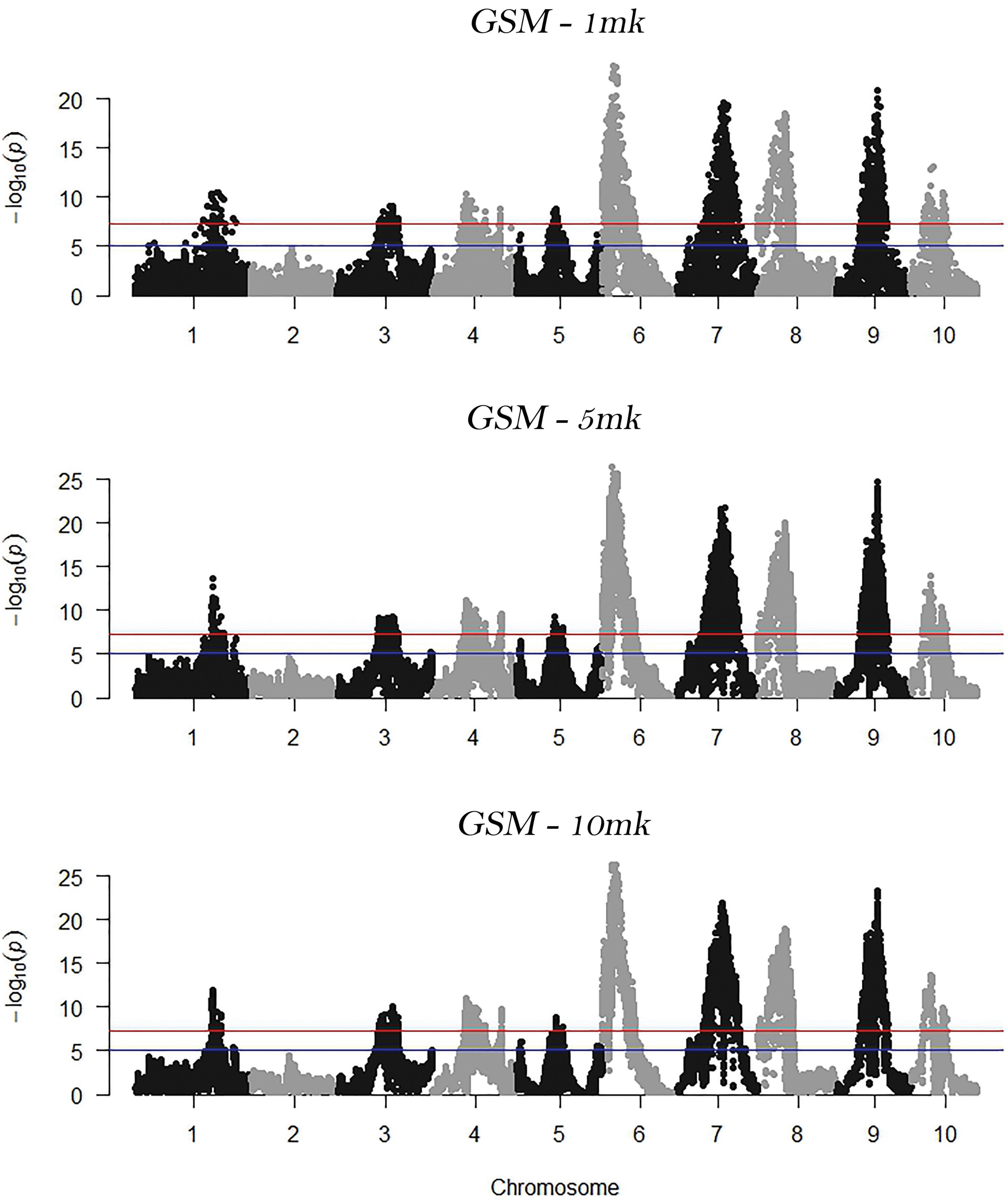


Figure 6.1: -log10(p-val) QTL profile of the GSM association scans for plant height in the EU-NAM Dent panel for local GSM composed of one, five and 10 markers.

useful alternative to define the genetic relatedness and detect QTLs in MPPs. Zheng et al. (2014) implemented a general framework to reconstruct ancestry block bit by bit

(RABBIT) in MPPs. Given the genotype information of the founders, the pedigree, and the genotypic information of the final lines, RABBIT determines the probabilistic allele origin of the final line in the founder alleles. The RABBIT method can handle a wide range of mating designs like diallel, breeding populations or the multi-parent advanced intercross (MAGIC) population, and model the allele transmission over a large number of generations (e.g. 100).

The IBD relationship matrices computed by RABBIT could be used to replace $\boldsymbol{X_{Q_i}}$ in model 6.1 for QTL detection. The use of a framework like RABBIT has several advantages compared to SNP based local GSMs. First, RABBIT makes a smart use of the available pedigree information, which has been shown to improve the QTL detection Yi & Xu (2001). A second advantage of RABBIT is the possibility to identify concrete founder alleles and express their frequency in the final line. Finally, we could use the RABBIT framework to visualize the gene flow along the pedigree and provide interesting information to the breeders.

6.2.6 The estimation of the QTL allelic effect

According to Xu (1998) and Wei & Xu (2016), fixed and random QTL effects have similar detection power. In our framework, we chose to fit the QTL term as fixed. Treating QTL effects as fixed allowed us to test for the significance of the whole QTL and of its allelic components with widely accepted statistical tests (F-test and Wald test). We could also have used random QTL terms estimating the significance of the QTL effect with a mixture of two chi-square distributions (Wei & Xu, 2016). Treating the QTL as fixed requires to define a constraint to estimate the effects. A first possibility is to set an allele as reference. For example in a NAM population, it is convenient to consider the central parent as the reference allele. However, in other designs, the choice of a reference is complicated. Thus, another option is to constrain the allelic effects to sum to zero. In that case, the QTL allelic effects will be defined as deviations with respect to the central tendency. Concerning the constraint, the use of a random QTL term can be convenient because, in that case, the BLUP estimates of the QTL allelic effects are sampled from a normal distribution with mean equal to zero. The interpretation of these allelic effects is also interesting because they are assumed to be drawn from a normal distribution representative of the whole population of potential QTL effects. These random QTL effects can therefore be extrapolated to other populations while fixed QTL effects are only meaningful in the particular MPP in which they were estimated.

In fixed QTL effect models, the estimation of the QTL effect and the constraint must be defined with respect to an interconnected part of the MPP design (Rebaï & Goffinet, 2000). An interconnected part can be drawn as a graph where the nodes represent the parents and the edges the crosses. In an interconnected part, it is possible to reach every nodes i starting from any node j (Weeks & Williams, 1964). The QTL allelic effects will be estimated independently in each interconnected part. When the MPP design becomes more complicated and contains several interconnected parts, it can be difficult to manage all the constraints and to produced meaningful estimates of the QTL effects. In such a case, we advise to visualise the graph of the design before realizing the QTL experiment to see if the addition of a few crosses can connect the different parts. Such a small intervention can make easier the interpretation of the QTL effects. Another possibility would be to use a random term for the polygenic background (g_{ij} in model 6.1) with a kinship matrix associated to define the VCOV. Such a strategy can create some connections between the different interconnected parts.

6.2.7 The polygenic effect

The polygenic effect represents the effect on the trait due to other QTLs than the tested QTL position, elsewhere in the genome. From a statistical point of view, it is important to model the QTL effect conditioned on the polygenic effect because it can reduce the error term and increase the QTL detection power (Zeng, 1993, 1994). In our case, we used a number of cofactor positions selected from a simple interval mapping (SIM) scan. Thus, we defined the QTL term of model 6.1 as $\sum_{i=1}^{n_{QTL}} \boldsymbol{X_{Q_i} \beta_{Q_i}} = \sum_{i=j}^{n_{cof}} \boldsymbol{X_{q_j} \beta_{q_j}} + \boldsymbol{X_Q \beta_Q}$. The use of a composite interval mapping (CIM) scan also allows to evaluate if the effect of a QTL is already explained by another QTL in the genome (Zeng, 1993). At the end of our procedure, we used a backward elimination to test simultaneously all the QTL positions we selected. An alternative to our procedure is to use a forward selection strategy (Li et al., 2011). In that case, at each QTL detection run, the most significant position is selected and integrated in the model as cofactor. A disadvantage of such a procedure is to make the QTL selection dependent from the positions that have been already selected in the process.

More generally, the selection of cofactors and the determination of a multi-QTL model face the problem of the multiplicity of equally good data models (Breiman et al., 2001). The general philosophy of model determination is to assign weights to the list of available variables (marker positions) to evaluate their importance in the explanation of the phenotypic variation. In many cases, the procedure can point out to many models with an equivalent goodness-of-fit or explanatory power. This problem becomes even more drastic with the increase in the amount of available markers. In that sense, the use of GP approaches generally offers the advantage to avoid the model selection problem by using all genetic information available (Meuwissen et al., 2001).

An alternative to the selection of cofactors performed in CIM or a forward selection is the use of a random polygenic term with a kinship matrix associated as VCOV. We could estimate the polygenic term g_{ij} of model 6.1 with a random term $g_{ij} \sim N(0, \boldsymbol{\Phi}\sigma_g^2)$ where $\boldsymbol{\Phi}$ is a kinship matrix representing the genetic relatedness existing between the genotypes. $\boldsymbol{\Phi}$ can be calculated using marker data and the formula from Astle & Balding (2009) or VanRaden (2008). The computation of the kinship matrix can also take into consideration the local LD pattern and assign weights to each marker proportional to their amount of tagging to balance the contribution of the genomic regions (Speed et al., 2012). Listgarten et al. (2012) also proposed to only use the subset of markers that are correlated to the trait to compute $\boldsymbol{\Phi}$. According to Rincent et al. (2014) and Yang et al. (2014), the kinship matrix $\boldsymbol{\Phi}$ should also be calculated without the markers of the scanned chromosome. The use of a random polygenic effect with kinship term allows to perform a single scan avoiding cofactor selection. However, according to Würschum et al. (2012) who analysed MPPs as association panels, the use of a kinship term was too stringent and only allowed to detect the QTLs with a large effect. The use of cofactors seemed a better option to increase the QTL detection power by detecting also the QTLs with a medium and small effect.

Another alternative to control for the polygenic effect without cofactor selection is the whole genome average interval mapping (WGAIM) procedure developed by Verbyla et al. (2007). The WGAIM approach tests if the genetic variance of all markers is equal to zero. If the null hypothesis is rejected, an outlier statistics allows to determine the most significant QTL, which is moved to the fixed part. The procedure continues until the genetic variance is estimated to be zero. A final option is a Bayesian method that incorporates all the markers simultaneously (Bink et al., 2012). In those Bayesian models, each marker receives a probability of being included in the model as QTL, and so, also avoids the intermediate step of cofactor selection.

6.2.8 The error term

Most of the procedure proposed for QTL detection in MPPs use linear models, which assumes an homogeneous variance residual term (HRT) (Blanc et al., 2006; Bardol et al., 2013; Giraud et al., 2014). This means that the residual variance of model 6.2 is assumed to be the same in every cross ($V(\boldsymbol{r}) = \boldsymbol{R} = \boldsymbol{I}\sigma_r^2$). The assumption of a constant residual variance can however be unrealistic. Indeed, when the MPP addresses a large genotypic and phenotypic diversity, we can expect a large heterogeneity of genetic and phenotypic distances between the parents of the crosses. If the MPP crosses have different levels of genetic and phenotypic diversity, this should produce heterogeneous levels of polygenic effects and therefore one would expect some heterogeneity in the residual variances.

Some authors were already aware of the potential heterogeneity of residual variance in MPPs. Some of them proposed to manage this problem by some transformation of the phenotypic data (Walling et al., 2000; Li et al., 2005). Others directly integrated in the model the possibility to have cross-specific residual variances. For example, Xu (1998) proposed to fit the model using iteratively re-weighted least squares. Li et al. (2011) fitted a maximum likelihood QTL model assuming cross-specific residual variances in the alternative hypothesis. Finally, Jannink & Wu (2003) also included cross-specific residual variances in their Bayesian procedure.

In chapter two, we proposed to address the heterogeneity of the residual variances between crosses by using models with cross-specific variance residual terms (CSRT) ($V(\boldsymbol{r}) = \boldsymbol{R} = \bigoplus_{c=1}^{n_c} \sigma^2_{r_c}$). We illustrated the usefulness of the CSRT in the supplemental material of chapter three. From a statistical point of view, we can show that the Wald test for the QTL effect $W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}})$ has the following property:

$$W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}}) \propto \sum_{c=1}^{n_c} \sum_{n=1}^{N_c} \frac{y_n \hat{y}_n}{\sigma^2_{r_c}} \tag{6.3}$$

The elements that compose $W(\hat{\boldsymbol{\beta}}_{\boldsymbol{Q}})$ are weighted by the cross-specific variances $\sigma^2_{r_c}$. In the HRT situation, the $\sigma^2_{r_c}$ would be replaced by a single term (σ^2_r) that represents the average level of uncertainty across the MPP. Let us assume that we have two crosses with $\sigma^2_{r_1} = 190$ and $\sigma^2_{r_2} = 10$. In that case, σ^2_r would be approximately be equal to 100. Then, if we use the HRT instead of the CSRT model, the contributions of cross one will be overestimated because the uncertainty of the HRT model ($\sigma^2_r = 100$) is lower than the uncertainty of the CSRT model ($\sigma^2_{r_1} = 190$). In cross two, we would have the opposite situation with underestimated contributions. This example illustrates the usefulness of the CSRT model to estimate the significance of the QTL term and of its allelic components. The larger the MPP genotypic and phenotypic heterogeneity, the more useful the CSRT model will be.

In chapter three, we evaluated the usefulness of the CSRT model for QTL detection in MPPs. We sampled parents of the EU-NAM Dent population to compose MPPs with various levels of genetic distances between the parents. We expected that the predictive ability of the CSRT model would be larger than the one of the HRT model when the diversity between parents increased. We could not observe such a difference. Several elements can explain this absence of difference. First, the genetic distances between the parents were not extremely large. Second, according to Hung et al. (2012), the phenotypic difference between parents is a more important criterion than the genetic distance to predict the within cross genetic variance and evaluate the diversity of an MPP. Finally, we could also imagine that the CSRT model needs large crosses (at least

100 individuals) to estimate precisely the within cross variances terms.

The use of the CSRT model should still be justified in MPPs covering a larger genotypic and phenotypic diversity with larger crosses. For example, in figure 6.2, we plotted the genome-wide significance of the parental QTL effects for the HRT and CSRT parental models used to characterize day to anthesis (DTA) in the US-NAM population (McMullen et al., 2009). The US-NAM population covers a wide genetic diversity and has large crosses (121-189 individuals). The phenotypic data represented the adjusted means calculated across two trials (New York and North Carolina 2007). The detected QTLs are indicated with dashed lines. In figure 6.2, we can notice some differences between the HRT (top) and CSRT (down) models. The HRT model detected 21 QTLs when the CSRT model detected 29. The significance pattern of the parental alleles is also different between the two analyses. The CSRT model shows more significant allelic regions for example on chromosomes three and seven. These extra significant QTL regions might be false positive but, as we showed it from a theoretical point of view, the CSRT model is more precise to estimate QTL significance than the HRT model.

An alternative to define more precisely the residual variance is to fit a model with a random pedigree term. In chapter two, we proposed to use a model where the polygenic term g_{ij} of model 6.1 was fitted as a random term with $g_{ij} \sim N(0, \boldsymbol{G}\sigma_g^2)$ where $\boldsymbol{G}$ was a genetic relationship matrix calculated from pedigree records using the method from Luo et al. (1992). Such an approach allows to model the variance due to a shared parent or to other pedigree relationships when this information is available. In chapter two, we also proposed a model combining random pedigree term and CSRT to model the phenotypic VCOV.

6.3 Framework extension

In chapter two, three and four, we proposed a methodology to detect QTLs in MPPs evaluated in a single environment. The most important features of our framework was the flexibility on the assumptions for the QTL term. The QTL effect could be considered as cross-specific or it could be described by parental, ancestral or SNP alleles with a consistent effect between crosses. These different models allowed to estimate the genetic variability of the QTLs due to different genetic backgrounds, which is the main advantage of the MPPs over the bi-parental crosses. In chapter five, we extended our methodology to also model the phenotypic variation due to the interactions between the QTL and the environment (QTLxE). We proposed a one-stage analysis on the plot data modelling simultaneously non genetic variation due to experimental design factors and the QTL effects. We finally showed that our models could be extended to integrate environmental indices to understand better the physiological mechanisms behind the QTLxE effects.

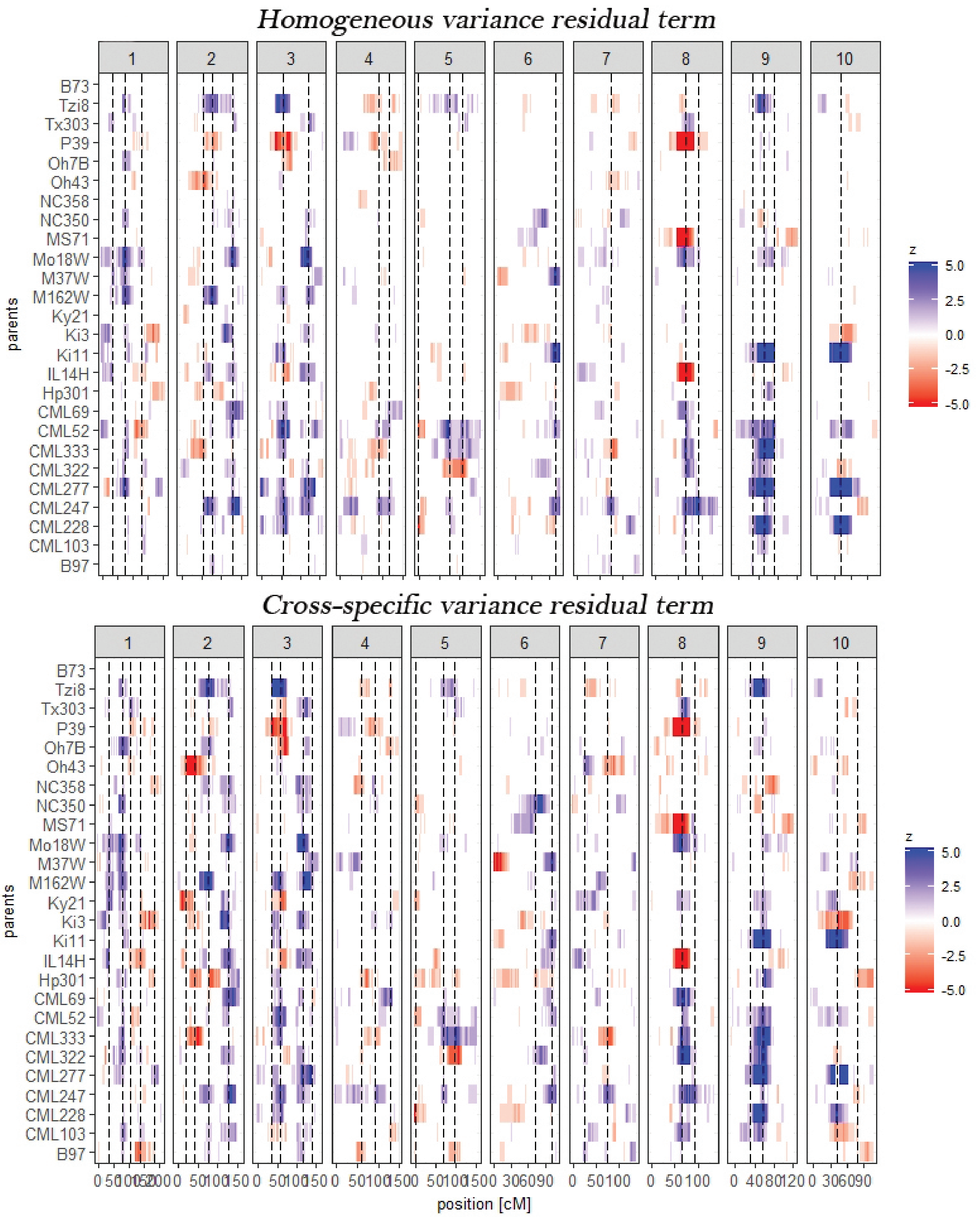


Figure 6.2: Genome-wide parental allelic effect significance in the US-NAM population for day to anthesis characterized at New York and North Carolina 2007. The results were obtained fitting a parental model with HRT (top) and CSRT (bottom). The dashed lines represent the detected QTLs.

6.3.1 MPP GxE QTL detection

Several MPPs have been characterized in multiple environments but generally the authors performed separated single environment QTL detections (Saade et al., 2016) or QTL detections on adjusted means or predictions across the environments that represent main phenotypic values (Buckler et al., 2009; Giraud et al., 2014). Only few authors have developed a proper methodology for MPP data characterized in multiple environments (MPP-ME) (Piepho & Pillen, 2004; Verbyla et al., 2014a). In chapter five, we extended our QTL detection framework to analyse MPP-ME data. We proposed MPP genotype by environment (GxE) models and illustrated them with examples from the EU-NAM Flint and the US-NAM populations characterized in two environments.

We compared several methods characterized by different ways to integrate the environmental variation. We first performed QTL detection on adjusted means (BLUEs) across environments representing a main phenotypic effect (M1). In M2, we performed separated QTL analyses per environment on the BLUEs calculated within environment. In M3, we did a two-stage MPP GxE QTL analysis by analysing jointly the within environment BLUEs taking into consideration the correlation existing between the measurements realized on the same genotype. Finally, we also performed a one-stage analysis on the plot data modelling simultaneously both the non-genetic variance due to experimental design factors and the QTL genetic variance (M4).

We showed that in presence of QTLxE effect, the user has to use M2, M3 or M4, since by definition M1 cannot estimate the QTLxE effects. M1 measures a main QTL effect that represents an average QTL effect across environments. We showed that for important QTLs some alleles had their effect influenced by the environment. In those cases, method M1 gave an average effect that did not reflect the environmental variability. Compared to M3 and M4, method M2 did not analyse the phenotypic data jointly and did not take advantage of the genetic covariance between environments. We also showed that method M2 detected fewer QTLs than M3 and M4. Even if these QTLs could be false positive, we are confident that elements of the joint analysis like the extra sample size allow to increase the detection power. Another advantage of methods M3 and M4 was the possibility to extend the models by integrating environmental indices like the temperature or the water precipitations. This allowed to estimate the effect of environmental covariates on the QTL effect to understand better the physiological mechanisms behind the QTLxE effects. From a prediction perspective, we showed that the four methods were equivalent. In terms of QTL detection power, we noticed that, in the example data we used, M1 often allowed to detect more QTLs than the other methods. This is due to the fact that method M1 is more parsimonious than the other methods. Method M1 uses less degrees of freedom for the QTL effect because it only estimates the main QTL effect across environments while the other methods estimate environmental specific QTL effects. Therefore, when the QTL effects are consistent across environments, which was mostly the case in our data, M1 is still useful.

The extension of our framework to analyse MPP-ME QTL experiments is an interesting contribution to QTL detection methodology. We know that the plant phenotype is the result of cumulative interactions between the genotype and the environment (Malosetti et al., 2013). The statistical models we propose are a first step to integrate the environmental information in the detection of QTLs in MPPs. Such a methodology allows to exploit the full potential of MPP-ME data by modelling both the genetic variability due to different crosses and the environmental variability.

The first results we obtained analysing the EU-NAM and the US-NAM data were interesting but we had to admit that, contrary to our expectations, we detected less QTLxE interactions than expected. Indeed, the examples we presented were the most illustrative selected from a large combination of traits and environments analysed. The large majority of these examples did not contain as clear QTLxE interactions as the one we presented in chapter five. Buckler et al. (2009) and Poland et al. (2011) found similar results. They noticed that, for flowering time and northern leaf blight in the US-NAM, the main QTL effects were larger than the QTLxE components. Even if flowering time is generally a trait with consistency across environments, such results are surprising. We could have expected that the wide genetic diversity present in the MPPs, especially in the US-NAM, would be more sensitive to the different environmental stimuli.

One possibility to increase the chance of detecting some significant QTLxE interactions is to perform MPP-ME experiments where the differences between environments are more clearly controlled and identified. For example, we also tested our MPP GxE QTL detection framework on the Barley HEB-NAM population developed by Maurer et al. (2015) characterized in control versus heat stress environments. The phenotypic data were kindly provided by Stephanie Saade and Mark Tester. In figure 6.3, we can observe the genome-wide parental allelic significance effect in control (top) and heat stress (bottom) environments for the QTL detection performed with an ancestral one-stage analysis for plant height. An interesting result is the QTL detected on chromosome 3 at 107.8 cM. At that location, the ancestral clustering defined two main allele classes. The first allele represented the German elite cultivar Barke (central parent) and the second regrouped 20 out of 25 of the exotic peripheral lines. We noticed that the ancestral allele carried by the exotic lines had a positive effect on plant height that is stronger in the heat stress environment. Maurer et al. (2015) also detected a QTL in that region (108.4 cM). According to them, it corresponds to the *denso* gene involved in the synthesis of gibberellic acid, an important component for flowering time.

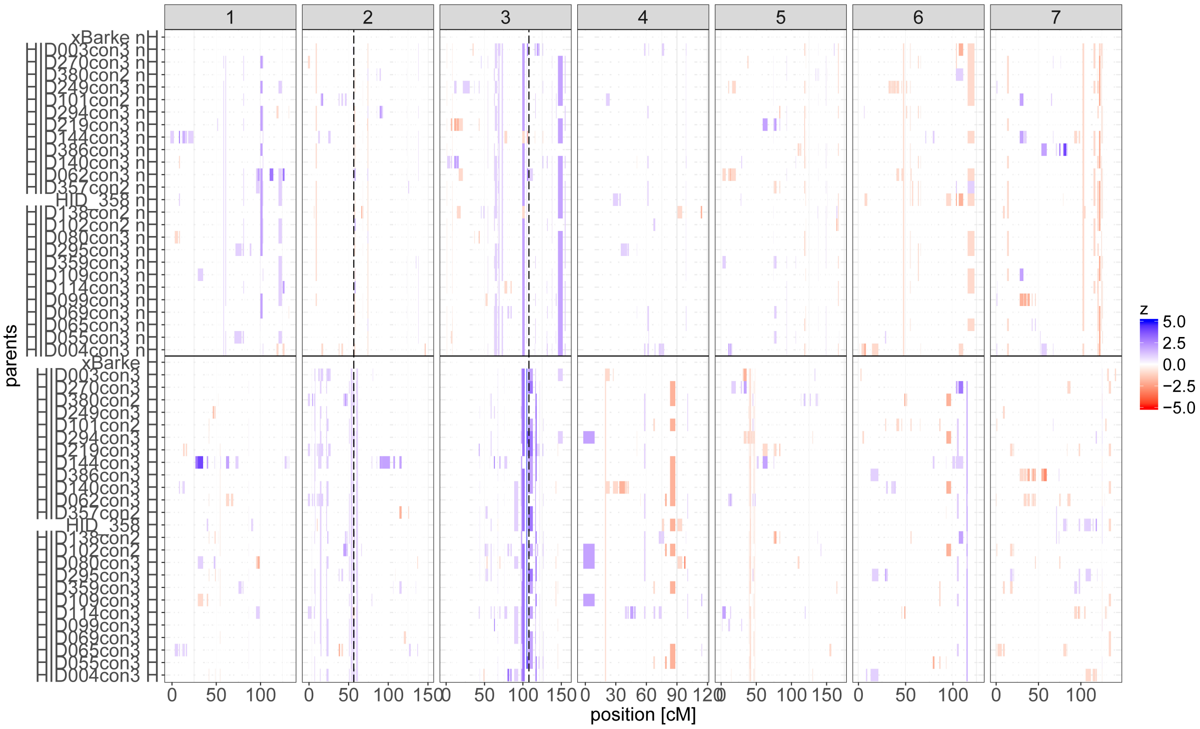


Figure 6.3: Genome-wide parental allelic effect significance in the HEB-NAM population for plant height characterized in control (top) vs heat stress (bottom) environments. The results were obtained fitting a one-stage ancestral model (M4). The dashed lines represent the detected QTLs.

The use of GxE QTL detection methods in MPPs is still a relatively unexplored area but we expect the attention for this topic to grow. As we have shown in our last example, MPP-ME experiments can be an appropriate way to characterize the environmental influence on the available genetic variability. The use of MPPs instead of association panels could also be a safer choice because the control for population structure via a kinship term in multi-environment GWAS is difficult (Korte et al., 2012). The use of MPP-ME QTL experiments is a promising strategy to develop new traits because it allows to test a structured population representative of the target genetic space into the target population of environments (Leung et al., 2015). One limitation of this approach is the growth of the datasets which could be difficult to analyse with many environments.

6.3.2 Extension of the MPP GxE models to a MPP multi-traits model

We presented the extension of our MPP QTL detection model using multi-environment data. The models we developed in chapter five can also be used to analyse multi-trait (MT) data with few adaptations or even without modifications. In many situations, researchers have dissected complex traits in a number of physiological component traits (Alimi et al., 2013b). In MPPs, the dissection of complex traits into component traits has already been practised by Hung et al. (2012) or Saade et al. (2016). Such a MPP-MT experiments could be analysed with the M3 and M4 approach we proposed to express the most complex traits as a function of the component traits. Multi-trait models can also find QTLs that are adapted to multiple sources of resistances. The choice of trait combinations analysed in the MPP-MT model could be motivated by a crop growth model (Alimi, 2016). However, as we could see in chapter five, the inclusion of many environments (traits) can quickly make the computation difficult, especially with a complex VCOV.

An interesting extension for the model we propose in chapter five is the modelling of traits measured at different time points (longitudinal traits). Several traits, like disease resistance are measured multiple times during a certain period (Macgregor et al., 2005). Generally, the trait value of the whole process is summarized in a single value that can be an average or the final measurement. However, for such longitudinal traits, it would be more interesting to have a statistical methodology allowing to describe the dynamics of the trait. A possibility would be to consider different time point measurements on the same trait as multiple traits and to analyse them with the method M3 or M4. In such a case, we should only modify slightly the VCOV describing the relationship between the traits to take into consideration the longitudinal dependence.

To illustrate our idea, we can imagine a MPP with a disease resistance trait measured several times $(t_1, ..., t_n)$. To cover the trait development dynamic we could analyse jointly the phenotypic measurements. In that case, the time dependency structure between a measurement taken on the same genotype at different time points could be described by an autoregressive process of order 1. The covariance between two measurements on the same genotype measured at points l and l^* is expressed by $\sigma_g^2 \rho^{|l-l^*|}$. The unknown ρ takes values between zero and one. It represents the rate at which the covariance decays between two time points measurements. The VCOV of such an experiment could also be described using and unstructured VCOV. In that case, $cov(y_{l..}, y_{l^*..}) = \sigma_{g_{l,l^*}}$. The model will estimate a unique covariance for each combination of time points.

The time series multi-trait model, the eco-physiological models, and the application of our MPP GxE models to control versus heat stress data are possible extensions of the models from chapter five. These models or experiments aim at integrating more information about the biological processes to better understand how the QTLs function.

6.4 The usefulness of MPPs for marker assisted selection

Marker assisted selection (MAS) is the most obvious application of our results to plant breeding. MAS uses the genetic information contained in DNA markers to select superior genotypes. QTL mapping is an essential component of MAS because it provides information about the markers by evaluating their association with the traits of interest. According to Collard et al. (2005), MAS is composed of three steps: 1) detect QTLs with high resolution; 2) validate the markers in an independent population; 3) transfer the markers in other genetic backgrounds. In the following sections, we will discuss the implications of our results for the different steps of MAS.

6.4.1 QTL detection power in MPPs

A first advantage of the MPPs for MAS is their larger QTL detection power with respect to the bi-parental crosses. In the present thesis, we did not compare the QTL detection power of the MPPs and the bi-parental crosses because several studies already treated this question (Rebaï & Goffinet, 1993, 2000; Muranty, 1996; Jansen et al., 2003). MPPs allow to increase the total population size and the sample size to estimate QTL allelic effects shared between crosses (Rebaï & Goffinet, 1993, 2000; Li et al., 2005). MPPs cover a larger genetic diversity, which reduces the chance of fixation at the QTL position (Xu, 1998). The larger genetic diversity present in MPPs also allows to sample a fraction of the genetic variance that is more representative of the one present in the hypothetical reference population, especially for multi-allelic QTLs (Wu & Jannink, 2004). All these elements increase the QTL detection power, which is an important success factor of MAS (Blanc et al., 2008).

Breeding populations are often composed of a large number of small crosses, which allow the breeder to test a large number of parents (Würschum, 2012). If diverse populations can be helpful to detect more QTLs, the multiplication of small crosses prevents from detecting QTLs due to too small sample sizes to estimate the QTL effects. This tension between creating diverse populations and increasing the QTL detection power was reflected in the simulation results from chapter four. We showed that, in general, it was more interesting to increase the number of individuals per cross than increasing the number of parents to cover a wider genetic diversity. These results agreed with other studies (Wu & Jannink, 2004; Muranty, 1996).

Increasing the number of parents is beneficial because it allows to avoid QTL fixation and to sample the QTL variation in a more representative way. However, it seems that after a certain number of parents, every possible QTL segregating in the population can be detected because it segregates in at least one cross. After this limit, increasing the number of parents is useless because it only increases marginally the chance of QTL segregation while it reduces the number of individuals per cross. Using crosses with few individuals reduces the chance to detect significant within cross QTL variance due to the small sample size. The determination of the optimal number of parents to form an MPP is an important question that certainly depends on many factors like the population size, the genetic architecture, the crop, etc.

The MPPs can be a compromise between increasing the population genetic diversity, and increasing the QTL detection power. A large unique bi-parental cross population is ideal to detect the QTLs because it guarantees large and balanced samples to estimate the QTL variance. On the other hand, the allelic diversity of such a unique cross is small. Thus, the use of MPPs with multiple crosses allows to increase the QTL allelic diversity while maintaining reasonable sample size to measure the QTL effects given that the crosses are not too small.

We also showed that increasing the number of parents was still beneficial for QTLs with a small allele frequency, given that the QTL effects were large (at least 6% of the phenotypic variance) and that the total population size was also large. In that situation, a larger number of parents and crosses allowed to sample at least one cross where the QTL was segregating. QTLs with rare alleles and large phenotypic effects represent an important part of the genetic variance (Leung et al., 2015). Therefore, increasing the covered genetic diversity by using more parents is still an interesting strategy given that the total population size is large. The negative effect of cross size reduction was less important for QTLs with shared effects between crosses like the ancestral QTLs. For those QTL alleles, the reduction of the within cross sample size was compensated by an increased between cross sample size due to consistent effects across crosses.

6.4.2 QTL detection resolution in MPPs

After the QTL detection power, the detection resolution is a second important element for MAS (Collard et al., 2005). Detecting QTL with a high resolution increases the chances that the markers are tightly linked to the true QTL and the underlying gene. If significant markers are detected far away from the QTL, it increases the probability of recombination between the marker and the QTL, which reduces the correlation marker-gene (Cobb et al., 2019). A low resolution also increases the chance to transfer undesirable characteristics (linkage drag) because the confidence region containing the interesting QTL is too wide (Leung et al., 2015).

As we could see in the simulations of chapter four, the resolution in MPPs composed of F2 crosses was not high. For example, in a N=800 population, assuming an equal distribution of QTL effects (cross-specific, parental, ancestral, and bi-allelic), we noticed that 95% of the QTLs explaining 2% (6%) of the phenotypic variation were detected between 0-35 (0-23) cM. Such a result can be explained by the fact that the LD extent in F2 populations is large. Indeed, F2 populations do not go through enough recombination cycles to break their genome in small pieces (Rakshit et al., 2012). The approximated resolution obtained in our MPP simulation was lower than the one obtained by Darvasi & Soller (1997) (33 and 11 cM) in single F2 populations of comparable sizes for QTL explaining the same proportion of phenotypic variance.

The use of a single large unique F2 population could give a higher QTL detection resolution because, at the QTL position, the within cross allele frequencies are expected to be balanced. This will increase the QTL detection power and the strength of the signal, which will improve the resolution. In MPPs however, we combine material coming from several crosses where the QTL segregation is not uniform. For example, the QTL may only segregate in few crosses. Therefore, the QTL detection power as the resolution can be reduced. We should also emphasize that the results of Darvasi & Soller (1997) were obtained for a single QTL model whereas our results came from QTL genetic models with eight QTLs. In our study, the noise introduced by undetected QTLs could reduce the QTL detection resolution.

MPPs without intercrossing, like the populations used in this thesis, are maybe not the best choice to increase QTL detection resolution. The MAGIC populations intercrossing DNA from several founders for a number of generations are probably more suitable (Valdar et al., 2006). MPP designs combining a larger number of DNA origins in the final generation increase the QTL detection resolution and the power with respect to designs where the final line are composed of DNA coming from only two parents (Klasen et al., 2012). Such populations are interesting tools to develop new traits (Leung et al., 2015) but they are more difficult to develop than the MPPs composed of crosses (Ladejobi et al., 2016).

6.4.3 Validation of the QTL effects in MPPs

According to Collard et al. (2005), a second important step in MAS is the validation of the QTL effect in different populations. Such a validation is a critical step. Indeed, according to Bernardo (2016), the detected QTLs have often allelic effects that are cross and/or environment specific, which prevents from introgressing them in other genetic backgrounds. The MPP designs are very useful to evaluate the QTL effect in different genetic background and therefore provide more reliable QTL effects for MAS (Cobb et al., 2019).

One of the main advantage of the MPPs over the bi-parental crosses is the possibility to evaluate the QTL effects in several genetic backgrounds (Blanc et al., 2006). Such a test will allow to better evaluate the marker-trait correlation Cobb et al. (2019). Using MPPs allows to test if the QTL effects are cross-specific or if they are consistent over crosses. The possibility to test for specificity versus consistency of the QTL effects is a major characteristic of the framework we proposed. The different models we defined in chapter two assume an increasing degree of QTL effect consistency. The cross-specific model assumes that the QTL effects are specific to a unique genetic background while the parental or ancestral models assume that the QTL allelic effects are consistent in the whole MPP. The comparison of the QTL effects estimated with the different models allows the user to determine if a QTL effect is specific or consistently defined within the MPP. The ancestral model is particularly interesting to detect consistent QTL effects. In this thesis, we tested a unique strategy to infer ancestral haplotypes based on the local clustering of the parental lines using genetic similarities (Leroux et al., 2014). Other haplotyping strategies based on recombination and/or LD could also be tried (Barrett et al., 2004).

The information about the consistency of the QTL effect in different genetic backgrounds could be used in a breeding perspective. If the breeder wants to maintain the variability or introduce some diversity in the population, then the selection of cross-specific QTLs with an effect restricted to one or two crosses would be better. On the other hand, if the breeder wants to introgress more stable effects, he could use ancestral QTLs with consistent effects in several genetic backgrounds. Such QTL effects will tend faster to fixation.

Compared to the estimation of the QTL effects in individual crosses, the joint estimation realized in MPPs allows a direct comparison of the QTL allelic effects. Such estimates should be more robust and more representative of genetic architecture at the whole population level. The QTL effects estimates obtained in MPPs give a better idea about the size of the QTL effect at the whole population level. A large QTL effect estimate in a single cross might not have the same strength in another cross. The joint estimation of the QTL allelic effects is also necessary to correctly evaluate the complementarity of their effects (Blanc et al., 2008).

In chapter five, we showed the possibility to test for the consistency of the MPP QTL allelic effects between environments. The use of MPP GxE models allows to test if QTL effects have consistent allelic effects within the MPP and between the environments. Such consistent QTLs should be selected and introgressed with priority. Therefore, from a general point of view, the QTLs detected in MPPs should show more generalizable effects, which should increase the chance of a successful introgression (Verhoeven et al., 2006).

6.4.4 Transfer the QTL in other genetic backgrounds

Once interesting QTLs have been identified and validated, the last step of MAS is to transfer them into other genetic backgrounds. Traditionally, breeders have used back-crossing introgression but the development of a MPP can also be seen as an interesting way to assemble QTL alleles from different sources. Indeed, compared to within cross MAS, MPPs can use interesting alleles from a larger pool of QTL effects. It might take a longer time to assemble useful QTL alleles from several donors but Blanc et al. (2008) showed that MPP MAS intercrossing lines coming from more than two parents outperformed the selection of lines coming from a single cross only. In Blanc et al. (2008) simulations, the next generation came from crosses with the best average molecular score. In the MPP MAS, the candidate crosses could be formed between individuals coming from all crosses. In the single cross MAS, next generation crosses could only be formed using individuals from the same family. The success of MPP MAS was larger when the superior alleles were evenly distributed among the parents (Blanc et al., 2008).

As mentioned in the previous section, the MPPs allow to estimate jointly the QTL allelic effects. This is very important to determine the complementarity of these effects and make a selection using more than two parents. Concerning MPP MAS, Blanc et al. (2008) advice to detect and estimate the QTL effects using an MPP design and then perform several rounds of selection in off-season nurseries using the QTL effect estimates to select the best lines. Such a strategy will be conditioned on the precision of the QTL effect estimates and the stability of the QTL effects over the next generations.

Another advantage of the MPPs to develop elite lines is the wide diversity of available designs. Breeders can use different MPP designs to achieve different goals (Cobb et al., 2019). NAM populations allow to test a wide collection of useful donors into the central (elite) parent. Diallel designs maximise the interconnection of the MPP, they guarantee an even contribution of each parent, and they allow to check the QTL deployment in several genetic backgrounds (Cobb et al., 2019). The factorial designs can be used to contrast a set of resistant and susceptible parents. Therefore, the breeder can choose the most appropriate design given the breeding objectives. For example, a NAM design allows to compare a large number of exotic alleles to the alleles of a reference line. Then interesting QTL alleles from the exotic lines can be introgressed in the central parent to increase its allelic diversity. The selection of a particular design could also be determined by evaluating its QTL detection power using the simulation tools developed in chapter four.

Finally, we could imagine that different types of populations correspond to different phases of the breeding process. In an early phase, we want to maximise the allelic diversity to find interesting lines. Therefore, we can use an association panel. For the next generations, we want to keep some diversity but we would like to limit the disadvantages of the association panels in term of population structure. Using MPPs with an interesting set of parents allows to keep some genetic diversity and to introduce some structure in the population at the same time. The use of MPPs also allows to increase the QTL allele frequencies of the selected parents. Finally, at the end of the process, we focus on final elite lines that could be better characterized in bi-parental crosses. According to Cobb et al. (2019), MPP MAS is an interesting strategy to enrich lines in the first stages of the breeding process before using GP to select elite lines.

References

Alimi, N., Bink, M., Dieleman, J., Magán, J., Wubs, A., Palloix, A., & Van Eeuwijk, F. (2013a). Multi-trait and multi-environment qtl analyses of yield and a set of physiological traits in pepper. *Theoretical and Applied Genetics*, *126*, 2597–2625.

Alimi, N., Bink, M., Dieleman, J., Nicolaï, M., Wubs, M., Heuvelink, E., Magan, J., Voorrips, R., Jansen, J., Rodrigues, P. et al. (2013b). Genetic and qtl analyses of yield and a set of physiological traits in pepper. *Euphytica*, *190*, 181–201.

Alimi, N. A. (2016). *Statistical methods for QTL mapping and genomic prediction of multiple traits and environments: case studies in pepper*. Ph.D. thesis Wageningen University.

Aristotle, Duminil, M.-P., & Jaulin, A. (2008). Métaphysique.

Astle, W., & Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, *24*, 451–471.

Aulchenko, Y. S., Ripke, S., Isaacs, A., & Van Duijn, C. M. (2007). Genabel: An r library for genome-wide association analysis. *Bioinformatics*, *23*, 1294–1296.

Bardol, N., Ventelon, M., Mangin, B., Jasson, S., Loywick, V., Couton, F., Derue, C., Blanchard, P., Charcosset, A., & Moreau, L. (2013). Combined linkage and linkage disequilibrium qtl mapping in multiple families of maize (zea mays l.) line crosses highlights complementarities between models based on parental haplotype and single locus polymorphism. *Theoretical and Applied Genetics*, *126*, 2717–2736.

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2004). Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, *21*, 263–265.

Bauer, E., Falque, M., Walter, H., Bauland, C., Camisan, C., Campo, L., Meyer, N., Ranc, N., Rincent, R., Schipprack, W. et al. (2013). Intraspecific variation of recombination rate in maize. *Genome Biology*, *14*, R103.

Beavis, W. D. (1998). Qtl analyses: power, precision, and accuracy. *Molecular dissection of complex traits*, *1998*, 145–162.

Bernardo, R. (2016). Bandwagons i, too, have known. *Theoretical and Applied Genetics*, *129*, 2323–2332.

Bink, M., Boer, M., Ter Braak, C., Jansen, J., Voorrips, R., & Van de Weg, W. (2008). Bayesian analysis of complex traits in pedigreed plant populations. *Euphytica*, *161*, 85–96.

Bink, M., Uimari, P., Sillanpää, M., Janss, L., & Jansen, R. (2002). Multiple qtl mapping in related plant populations via a pedigree-analysis approach. *Theoretical and Applied Genetics*, *104*, 751–762.

Bink, M. C., Totir, L. R., ter Braak, C. J., Winkler, C. R., Boer, M. P., & Smith, O. S. (2012). Qtl linkage analysis of connected populations using ancestral marker and pedigree information. *Theoretical and Applied Genetics*, *124*, 1097–1113.

Blanc, G., Charcosset, A., Mangin, B., Gallais, A., & Moreau, L. (2006). Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. *Theoretical and Applied Genetics*, *113*, 206–224.

Blanc, G., Charcosset, A., Veyrieras, J.-B., Gallais, A., & Moreau, L. (2008). Marker-assisted selection efficiency in multiple connected populations: a simulation study based on the results of a qtl detection experiment in maize. *Euphytica*, *161*, 71–84.

Boer, M. P., Wright, D., Feng, L., Podlich, D. W., Luo, L., Cooper, M., & van Eeuwijk, F. A. (2007). A mixed-model quantitative trait loci (qtl) analysis for multiple-environment trial data using environmental covariables for qtl-by-environment interactions, with an example in maize. *Genetics*, *177*, 1801–1813.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). Tassel: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*, 2633–2635.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*, 199–231.

Broman, K. W., Wu, H., Sen, Ś., & Churchill, G. A. (2003). R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*, *19*, 889–890.

Browning, B. L., & Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, *194*, 459–471.

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J. C. et al. (2009). The genetic architecture of maize flowering time. *Science*, *325*, 714–718.

Bustos-Korts, D., Malosetti, M., Chapman, S., Biddulph, B., & van Eeuwijk, F. (2016). Improvement of predictive ability by uniform coverage of the target genetic space. *G3: Genes, Genomes, Genetics*, *6*, 3733–3747.

Butler, D., Cullis, B. R., Gilmour, A., & Gogel, B. (2009). Asreml-r reference manual. *The State of Queensland, Department of Primary Industries and Fisheries, Brisbane*, .

Cavanagh, C., Morell, M., Mackay, I., & Powell, W. (2008). From mutations to magic:

resources for gene discovery, validation and delivery in crop plants. *Current opinion in plant biology*, *11*, 215–221.

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

Cobb, J. N., Biswas, P. S., & Platten, J. D. (2019). Back to the future: revisiting mas as a tool for modern plant breeding. *Theoretical and Applied Genetics*, *132*, 647–667.

Cockerham, C. C. (1963). Estimation of genetic variances. *Statistical genetics and plant breeding*, *982*, 53–94.

Coles, N. D., McMullen, M. D., Balint-Kurti, P. J., Pratt, R. C., & Holland, J. B. (2010). Genetic control of photoperiod sensitivity in maize revealed by joint multiple population analysis. *Genetics*, *184*, 799–812.

Collard, B. C., Jahufer, M., Brouwer, J., & Pang, E. (2005). An introduction to markers, quantitative trait loci (qtl) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica*, *142*, 169–196.

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the r package sommer. *PLoS ONE*, *11*, 1–15.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* volume 32. CRC Press.

Crepieux, S., Lebreton, C., Flament, P., & Charmet, G. (2005). Application of a new ibd-based qtl mapping method to common wheat breeding population: analysis of kernel hardness and dough strength. *Theoretical and Applied Genetics*, *111*, 1409–1419.

Crepieux, S., Lebreton, C., Servin, B., & Charmet, G. (2004). Quantitative trait loci (qtl) detection in multicross inbred designs recovering qtl identical-by-descent status information from marker data. *Genetics*, *168*, 1737–1749.

Crossa, J., Perez, P., Hickey, J., Burgueno, J., Ornella, L., Cerón-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R., Li, Y. et al. (2014). Genomic prediction in cimmyt maize and wheat breeding programs. *Heredity*, *112*, 48.

Darvasi, A., & Soller, M. (1997). A simple method to calculate resolving power and confidence interval of qtl map location. *Behavior genetics*, *27*, 125–132.

Darwin, C. (1859). *On the Origin of Species*.

Descartes, R. (1637). *Discours de la méthode*.

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

Durkheim, E. (1894). *De la Division du Travail Social*.

van Eeuwijk, F. A., Bink, M. C., Chenu, K., & Chapman, S. C. (2010a). Detection and use of qtl for complex traits in multiple environments. *Current opinion in plant biology*, *13*, 193–205.

van Eeuwijk, F. A., Boer, M., Totir, L. R., Bink, M., Wright, D., Winkler, C. R., Podlich, D., Boldman, K., Baumgarten, A., Smalley, M. et al. (2010b). Mixed model approaches for the identification of qtls within a maize hybrid breeding program. *Theoretical and Applied Genetics*, *120*, 429–440.

van Eeuwijk, F. A., Malosetti, M., & Boer, M. P. (2007). Modelling the genetic basis of response curves underlying genotype x environment interaction. In *Scale and Complexity in Plant Systems Research: Gene Plant-Crop Relations* 21 (pp. 115–126). Springer.

Ellul, J. (1954). *La technique ou l'enjeu du siècle*. A. Colin Paris.

Fisher, R. (1958). *The genetical theory of natural selection 2*. New York, NY (USA) Dover Pub.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, *52*, 399–433.

Flint-Garcia, S. A., Thornsberry, J. M., & Buckler IV, E. S. (2003). Structure of linkage disequilibrium in plants. *Annual review of plant biology*, *54*, 357–374.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning* volume 1. Springer series in statistics Springer-Verlag, Berlin.

Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., Clarke, J. D., Graner, E.-M., Hansen, M., Joets, J. et al. (2011). A large maize (zea mays l.) snp genotyping array: development and germplasm genotyping, and genetic mapping to compare with the b73 reference genome. *PloS one*, *6*, e28334.

Garin, V., Wimmer, V., Borchardt, D., van Eeuwijk, F., & Malosetti, M. (2018). *mppR: Multi-Parent Population QTL Analysis*. R package version 1.1.10.

Garin, V., Wimmer, V., Mezmouk, S., Malosetti, M., & van Eeuwijk, F. (2017). How do the type of qtl effect and the form of the residual term influence qtl detection in multi-parent populations? a case study in the maize eu-nam population. *Theoretical and Applied Genetics*, *130*, 1753–1764.

Gilmour, A. R., Cullis, B. R., & Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, (pp. 269–293).

Giraud, H., Lehermeier, C., Bauer, E., Falque, M., Segura, V., Bauland, C., Camisan, C., Campo, L., Meyer, N., Ranc, N. et al. (2014). Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic qtl for hybrid performance in the flint and dent heterotic groups of maize. *Genetics*, *198*, 1717–1734.

Guo, B., Sleper, D., Sun, J., Nguyen, H., Arelli, P., & Shannon, J. (2006). Pooled analysis of data from multiple quantitative trait locus mapping populations. *Theoretical and Applied Genetics*, *113*, 39–48.

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

Haley, C. S., & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, *69*, 315–324.

Han, S., Utz, H. F., Liu, W., Schrag, T. A., Stange, M., Würschum, T., Miedaner, T., Bauer, E., Schön, C.-C., & Melchinger, A. E. (2016). Choice of models for qtl mapping with multiple families and design of the training set for prediction of fusarium resistance traits in maize. *Theoretical and Applied Genetics*, *129*, 431–444.

Haseman, J., & Elston, R. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior genetics*, *2*, 3–19.

Huang, X., Paulo, M.-J., Boer, M., Effgen, S., Keizer, P., Koornneef, M., & van Eeuwijk, F. A. (2011). Analysis of natural allelic variation in arabidopsis using a multiparent recombinant inbred line population. *Proceedings of the National Academy of Sciences*, *108*, 4488–4493.

Hung, H., Browne, C., Guill, K., Coles, N., Eller, M., Garcia, A., Lepak, N., Melia-Hancock, S., Oropeza-Rosas, M., Salvo, S. et al. (2012). The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity*, *108*, 490–499.

Jannink, J.-l., & Wu, X.-L. (2003). Estimating allelic number and identity in state of qtls in interconnected families. *Genetical research*, *81*, 133–144.

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

Jansen, R. C., Jannink, J.-L., & Beavis, W. D. (2003). Mapping quantitative trait loci in plant breeding populations. *Crop Science*, *43*, 829–834.

Jourjon, M.-F., Jasson, S., Marcel, J., Ngom, B., & Mangin, B. (2005). Mcqtl: multi-allelic qtl mapping in multi-cross design. *Bioinformatics*, *21*, 128–130.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*, 1709–1723.

Kessner, D., & Novembre, J. (2015). Power analysis of artificial selection experiments using efficient whole genome simulation of quantitative traits. *Genetics*, *199*, 991–1005.

Klasen, J., Piepho, H., & Stich, B. (2012). Qtl detection power of multi-parental ril populations in arabidopsis thaliana. *Heredity*, *108*, 626–632.

Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., & Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics*, *44*, 1066.

Ladejobi, O., Elderfield, J., Gardner, K. A., Gaynor, R. C., Hickey, J., Hibberd, J. M., Mackay, I. J., & Bentley, A. R. (2016). Maximizing the potential of multi-parental crop populations. *Applied & translational genomics*, *11*, 9–17.

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

Lehermeier, C., Krämer, N., Bauer, E., Bauland, C., Camisan, C., Campo, L., Flament, P., Melchinger, A. E., Menz, M., Meyer, N. et al. (2014). Usefulness of multiparental populations of maize (zea mays l.) for genome-based prediction. *Genetics*, *198*, 3–16.

Lenth, R. (2018). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.1.2.

Leroux, D., Rahmani, A., Jasson, S., Ventelon, M., Louis, F., Moreau, L., & Mangin, B. (2014). Clusthaplo: a plug-in for mcqtl to enhance qtl detection using ancestral alleles in multi-cross design. *Theoretical and Applied Genetics*, *127*, 921–933.

Leung, H., Raghavan, C., Zhou, B., Oliva, R., Choi, I. R., Lacorte, V., Jubay, M. L., Cruz, C. V., Gregorio, G., Singh, R. K. et al. (2015). Allele mining and enhanced genetic recombination for rice breeding. *Rice*, *8*, 34.

Li, H., Bradbury, P., Ersoz, E., Buckler, E. S., & Wang, J. (2011). Joint qtl linkage mapping for multiple-cross mating design sharing one common parent. *PLoS One*, *6*, e17573.

Li, J., Bus, A., Spamer, V., & Stich, B. (2016). Comparison of statistical models for nested association mapping in rapeseed (brassica napus l.) through computer simulations. *BMC plant biology*, *16*, 26.

Li, J., & Jiang, T. (2005). Haplotype-based linkage disequilibrium mapping via direct data mining. *Bioinformatics*, *21*, 4384–4393.

Li, R., Lyons, M. A., Wittenburg, H., Paigen, B., & Churchill, G. A. (2005). Combining data from multiple inbred line crosses improves the power and resolution of quantitative trait loci mapping. *Genetics*, *169*, 1699–1709.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., & Zhang, Z. (2012). Gapit: genome association and prediction integrated tool. *Bioinformatics*, *28*, 2397–2399.

Lippert, C., Quon, G., Kang, E. Y., Kadie, C. M., Listgarten, J., & Heckerman, D. (2013). The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific reports*, *3*.

Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., & Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. *Nature methods*, *9*, 525–526.

Liu, W., Maurer, H., Reif, J., Melchinger, A., Utz, H., Tucker, M., Ranc, N., Della Porta, G., & Würschum, T. (2013). Optimum design of family structure and allocation of resources in association mapping with lines from multiple crosses. *Heredity*, *110*, 71–79.

Liu, W., Reif, J. C., Ranc, N., Della Porta, G., & Würschum, T. (2012). Comparison of biometrical approaches for qtl detection in multiple segregating families. *Theoretical and Applied Genetics*, *125*, 987–998.

Luo, Z. et al. (1992). Computing inbreeding coefficients in large populations. *Genetics Selection Evolution*, *24*, 305–313.

Macgregor, S., Knott, S. A., White, I., & Visscher, P. M. (2005). Qtl analysis of longitudinal quantitative trait data in complex pedigrees. *Genetics*, .

Mackay, T. F. (1996). The nature of quantittative genetic variation revisited: Lessons from drosophila bristles. *BioEssays*, *18*, 113–121.

Malosetti, M., van der Linden, C. G., Vosman, B., & van Eeuwijk, F. A. (2007). A mixed-model approach to association mapping using pedigree information with an illustration of resistance to phytophthora infestans in potato. *Genetics*, *175*, 879–889.

Malosetti, M., Ribaut, J.-M., & van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Frontiers in physiology*, *4*, 44.

Malosetti, M., Voltas, J., Romagosa, I., Ullrich, S., & Van Eeuwijk, F. (2004). Mixed models including environmental covariables for studying qtl by environment interaction. *Euphytica*, *137*, 139–145.

Maurer, A., Draba, V., Jiang, Y., Schnaithmann, F., Sharma, R., Schumann, E., Kilian, B., Reif, J. C., & Pillen, K. (2015). Modelling the genetic architecture of flowering time control in barley through nested association mapping. *BMC Genomics*, *16*, 290.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. Wiley Online Library.

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C. et al. (2009). Genetic properties of the maize nested association mapping population. *Science*, *325*, 737–740.

Mendel, G. (1866). Versuche über pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brunn 4: 3*, *44*.

Mendel, G., Corcos, A. F., & Monaghan, F. V. (1993). *Gregor Mendel's Experiments on plant hybrids: a guided study*. Rutgers University Press.

Meuwissen, T., Hayes, B., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*, 1819–1829.

Meuwissen, T. H., Karlsen, A., Lien, S., Olsaker, I., & Goddard, M. E. (2002). Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics*, *161*, 373–379.

Muchero, W., Diop, N. N., Bhat, P. R., Fenton, R. D., Wanamaker, S., Pottorff, M., Hearne, S., Cisse, N., Fatokun, C., Ehlers, J. D. et al. (2009). A consensus genetic map of cowpea [vigna unguiculata (l) walp.] and synteny based on est-derived snps. *Proceedings of the national academy of sciences*, *106*, 18159–18164.

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

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., & Buckler, E. S. (2009). Association mapping: critical considerations shift from genotyping to experimental design. *The Plant Cell*, *21*, 2194–2202.

Nei, M., & Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, *76*, 5269–5273.

Nicholson, G., Smith, A. V., Jónsson, F., Gústafsson, Ó., Stefánsson, K., & Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 695–715.

Ogut, F., Bian, Y., Bradbury, P. J., & Holland, J. B. (2015). Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. *Heredity*, *114*, 552–563.

Orel, V., & Wood, R. J. (2000). Essence and origin of mendel's discovery. *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie*, *323*, 1037–1041.

Parisseaux, B., & Bernardo, R. (2004). In silico mapping of quantitative trait loci in maize. *Theoretical and Applied Genetics*, *109*, 508–514.

Pascal, B. (1651). *Préface sur le traité du vide*.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, *2*, e190.

Piepho, H.-P., & Pillen, K. (2004). Mixed modelling for qtl× environment interaction analysis. *Euphytica*, *137*, 147–153.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2017). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131.

Poland, J. A., Bradbury, P. J., Buckler, E. S., & Nelson, R. J. (2011). Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proceedings of the National Academy of Sciences*, *108*, 6893–6898.

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson.

Powell, J. E., Visscher, P. M., & Goddard, M. E. (2010). Reconciling the analysis of ibd and ibs in complex trait studies. *Nature Reviews Genetics*, *11*, 800–805.

Rakshit, S., Rakshit, A., & Patil, J. (2012). Multiparent intercross populations in analysis of quantitative traits. *Journal of genetics*, *91*, 111–117.

Rao, C. R., Toutenburg, H., Shalabh, & Heumann, C. (2008). *Linear Models and Generalizations: Least Squares and Alternatives*. Springer.

Rebaï, A., & Goffinet, B. (1993). Power of tests for qtl detection using replicated progenies derived from a diallel cross. *Theoretical and Applied Genetics*, *86*, 1014–1022.

Rebaï, A., & Goffinet, B. (2000). More about quantitative trait locus mapping with diallel designs. *Genetical research*, *75*, 243–247.

Rincent, R., Moreau, L., Monod, H., Kuhn, E., Melchinger, A. E., Malvar, R. A., Moreno-Gonzalez, J., Nicolas, S., Madur, D., Combes, V. et al. (2014). Recovering power in association mapping panels with variable levels of linkage disequilibrium. *Genetics*, *197*, 375–387.

Rosenberg, N. A., & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, *3*, 380–390.

Saade, S., Maurer, A., Shahid, M., Oakey, H., Schmöckel, S. M., Negrão, S., Pillen, K., & Tester, M. (2016). Yield-related salinity tolerance traits identified in a nested association mapping (nam) population of wild barley. *Scientific Reports*, *6*, 32586.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, *74*, 5463–5467.

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

Schwegler, D. D., Liu, W., Gowda, M., Würschum, T., Schulz, B., & Reif, J. C. (2013). Multiple-line cross quantitative trait locus mapping in sugar beet (beta vulgaris l.). *Molecular breeding*, *31*, 279–287.

Soller, M., Brody, T., & Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics*, *47*, 35–39.

Speed, D., & Balding, D. J. (2014). Multiblup: improved snp-based prediction for complex traits. *Genome research*, *24*, 1550–1557.

Speed, D., & Balding, D. J. (2015). Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, *16*, 33–44.

Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012). Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, *91*, 1011–1021.

Steinhoff, J., Liu, W., Maurer, H. P., Würschum, T., Friedrich, C., Longin, H., Ranc, N., & Reif, J. C. (2011). Multiple-line cross quantitative trait locus mapping in european elite maize. *Crop science*, *51*, 2505–2516.

Steinhoff, J., Liu, W., Reif, J. C., Della Porta, G., Ranc, N., & Würschum, T. (2012). Detection of qtl for flowering time in multiple families of elite maize. *Theoretical and Applied Genetics*, *125*, 1539–1551.

Sun, G., Zhu, C., Kramer, M., Yang, S., Song, W., Piepho, H., & Yu, J. (2010). Variation explained in mixed-model association mapping. *Heredity*, *105*, 333–340.

Team, R. C. (2016). R: A language and environment for statistical computing. vienna: R foundation for statistical computing; 2014.

Ter Braak, C. J., Boer, M. P., Totir, L. R., Winkler, C. R., Smith, O. S., & Bink, M. C. (2010). Identity-by-descent matrix decomposition using latent ancestral allele models. *Genetics*, *185*, 1045–1057.

Toro, M. Á., García-Cortés, L. A., & Legarra, A. (2011). A note on the rationale for estimating genealogical coancestry from molecular markers. *Genetics Selection Evolution*, *43*, 1.

Utz, H. F., Melchinger, A. E., & Schön, C. C. (2000). Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics*, *154*, 1839–1849.

Valdar, W., Flint, J., & Mott, R. (2006). Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics*, *172*, 1783–1797.

VanRaden, P. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, *91*, 4414–4423.

Varshney, R. K., Singh, V. K., Hickey, J. M., Xun, X., Marshall, D. F., Wang, J., Edwards, D., & Ribaut, J.-M. (2016). Analytical and decision support tools for genomics-assisted breeding. *Trends in plant science*, *21*, 354–363.

Verbyla, A. P., Cavanagh, C. R., & Verbyla, K. L. (2014a). Whole-genome analysis of multienvironment or multitrait qtl in magic. *G3: Genes, Genomes, Genetics*, *4*, 1569–1584.

Verbyla, A. P., Cullis, B. R., & Thompson, R. (2007). The analysis of qtl by simultaneous use of the full linkage map. *Theoretical and Applied Genetics*, *116*, 95.

Verbyla, A. P., George, A. W., Cavanagh, C. R., & Verbyla, K. L. (2014b). Whole-genome qtl analysis for magic. *Theoretical and Applied Genetics*, *127*, 1753–1770.

Verhoeven, K., Jannink, J., & McIntyre, L. (2006). Using mating designs to uncover qtl and the genetic architecture of complex traits. *Heredity*, *96*, 139–149.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, *54*, 426–482.

Walling, G. A., Visscher, P. M., Andersson, L., Rothschild, M. F., Wang, L., Moser, G., Groenen, M. A., Bidanel, J.-P., Cepica, S., Archibald, A. L. et al. (2000). Combined analyses of data from quantitative trait loci mapping studies: chromosome 4 effects on porcine growth and fatness. *Genetics*, *155*, 1369–1378.

Watson, J. D., Crick, F. H. et al. (1953). Molecular structure of nucleic acids. *Nature*, *171*, 737–738.

Weeks, D. L., & Williams, D. R. (1964). A note on the determination of connectedness in an n-way cross classification. *Technometrics*, *6*, 319–324.

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

Wimmer, V., Albrecht, T., Auinger, H.-J., & Schön, C.-C. (2012). synbreed: a framework for the analysis of genomic prediction data using r. *Bioinformatics*, *28*, 2086–2087.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, *89*, 82–93.

Wu, R., Ma, C., & Casella, G. (2007). *Statistical genetics of quantitative traits: linkage, maps and QTL*. Springer Science & Business Media.

Wu, X.-L., & Jannink, J.-L. (2004). Optimal sampling of a population to determine qtl location, variance, and allelic number. *Theoretical and Applied Genetics*, *108*, 1434–1442.

Würschum, T. (2012). Mapping qtl for agronomic traits in breeding populations. *Theoretical and Applied Genetics*, *125*, 201–210.

Würschum, T., Liu, W., Gowda, M., Maurer, H., Fischer, S., Schechert, A., & Reif, J. (2012). Comparison of biometrical models for joint linkage association mapping. *Heredity*, *108*, 332–340.

Xavier, A., Xu, S., Muir, W. M., & Rainey, K. M. (2015). Nam: association studies in multiple populations. *Bioinformatics*, *31*, 3862–3864.

Xie, C., Gessler, D. D., & Xu, S. (1998). Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. *Genetics*, *149*, 1139–1146.

Xu, P., Wu, X., Wang, B., Liu, Y., Ehlers, J. D., Close, T. J., Roberts, P. A., Diop, N.-N., Qin, D., Hu, T. et al. (2011). A snp and ssr based genetic map of asparagus bean (vigna. unguiculata ssp. sesquipedialis) and comparison with the broader species. *PloS one*, *6*, e15952.

Xu, S. (1996). Computation of the full likelihood function for estimating variance at a quantitative trait locus. *Genetics*, *144*, 1951–1960.

Xu, S. (1998). Mapping quantitative trait loci using multiple families of line crosses. *Genetics*, *148*, 517–524.

Xu, S., & Atchley, W. R. (1995). A random model approach to interval mapping of quantitative trait loci. *Genetics*, *141*, 1189–1197.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W. et al. (2010). Common snps explain a large proportion of the heritability for human height. *Nature genetics*, *42*, 565–569.

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

Yi, N., & Xu, S. (2001). Bayesian mapping of quantitative trait loci under complicated mating designs. *Genetics*, *157*, 1759–1771.

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

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B. et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, *38*, 203–208.

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.

Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics*, *136*, 1457–1468.

Zheng, C., Boer, M. P., & van Eeuwijk, F. A. (2014). A general modeling framework for genome ancestral origins in multiparental populations. *Genetics*, *198*, 87–101.

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

Summary

Since Mendel and the birth of (plant) genetic as a science, the use of controlled plant populations is central in plant science. Different plant populations have been developed to investigate biological questions like the detection of quantitative trait loci (QTLs). Historically, QTL detection has been performed in bi-parental crosses and association panels but these populations are limited by their reduced genetic diversity or their unknown population structure. Multi-parent populations (MPPs) composed of crosses are structured populations using genotypes coming from several crosses between a set of parents. Therefore, MPPs address the limitations of both bi-parental crosses and association panels. Populations like the nested association mapping, the diallels, and the factorial designs used in breeding programs are important examples of MPPs composed of crosses. An important element for QTL detection is the necessity to reflect the biological properties of the studied population in the statistical model. Thus, the central question of this thesis was to develop a statistical QTL detection methodology adapted to the MPPs composed of crosses.

In the second chapter, we presented a collection of QTL detection models for MPPs with different biological assumptions about the type of QTL effect. We defined models where the QTL effects could be: a) specific to a particular cross (cross-specific), b) where the effects of a common parent or ancestor were consistently defined in the whole MPP (parental and ancestral), and c) with QTL allelic effects attached to the SNP alleles (bi-allelic). We also let the error term variance of our models being cross-specific to reflect differences of residual polygenic effect between crosses.

In the third chapter, we tested our methodology on real data and we evaluated some biological assumptions behind our models. We were not able to verify the assumption that models with a reduced number of QTL terms perform better in MPPs with a narrow genetic basis. We could also not verify the usefulness of cross-specific error term variances to handle the heterogeneity of the genetic distance between the MPPs parents. However, we showed that models with QTLs described by different type of effects could improve the modelling of the phenotypic variation.

In chapter four, we evaluated our methodology by simulation. We simulated traits with a genetic architecture composed of QTLs showing more or less allelic diversity to investigate

the effect of genetic diversity on the QTL detection power. We also tried to determine some guidelines to build MPPs maximising the QTL detection power. We noticed that MPPs with large cross sizes were the most powerful designs. Increasing the number of parents was only useful to detect QTLs with a reduced allele frequency.

In the fifth chapter, we extended our methodology to analyse QTL experiments made on MPPs characterized in multiple environments (MPP-ME). We analysed jointly MPP-ME data taking into consideration the correlation due to the same genotype measured in different environments. We showed that our method could estimate the QTL by environment (QTLxE) effects, which was not the case for methods generally used to analyse MPP-ME data. We also showed that our models could integrate environmental information to get more insight about the underlying mechanisms behind the QTLxE effects.

To conclude, we would like to emphasize that the methodology we developed in this thesis allows to exploit the full potential of MPP-ME QTL experiments. Our models allow to estimate QTL variations: a) within the MPP between different cross genetic backgrounds, and b) between the environments. Therefore, our models can detect QTLs that potentially have a consistent effect within the population and between the environments. Such QTLs are certainly the most valuable for marker assisted selection.

About the author

Vincent Garin was born in Geneva Switzerland in 1985. After graduating from high school in economics and law, he earned a bachelor and a master degree in forensic science and criminology at the University of Lausanne. He spent his last semester at Sam Houston state University in Texas. After a reorientation, Vincent earned a master in statistics at the University of Geneva after an Erasmus year at Leiden University. There, he had the opportunity to meet Professor Fred van Eeuwijk and to establish a connection between statistics and agriculture. Vincent is passionate about agriculture and he tries to explore the multiple dimensions of this fundamental activity. During this PhD project, he could get a good understanding of the scientific and technical aspect of agriculture but he is also highly interested in more concrete aspects of farming that he could experience in volunteer projects in Swiss mountain farms. The connections between agriculture and ethics or agriculture and art are two other aspects that Vincent would like to develop in future projects.

Peer-reviewed journal publications

Garin, V. (2012). Social instability and reaction to deviance: A multilevel analysis of the Swiss lifelong detention initiative. *Punishment & Society*, 14(3), 289-314.

Garin, V., Wimmer, V., Mezmouk, S., Malosetti, M., & van Eeuwijk, F. (2017). How do the type of QTL effect and the form of the residual term influence QTL detection in multi-parent populations? A case study in the maize EU-NAM population. *Theoretical and Applied Genetics*, 1-12.

Other scientific publications

Garin V., Wimmer V., Borchardt D., van Eeuwijk F., Malosetti M. (2018). mppR: Multi-Parent Population QTL Analysis. R package version 1.2.0.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

The C.T. De Wit
Graduate School
PE&RC
PRODUCTION
ECOLOGY
& RESOURCE
CONSERVATION

Review of literature (4.5 ECTS)

- Detection and use of QTL in multi-parent population

Writing of project proposal (1 ECTS)

- Detection and use of QTL in multi-parent population

Post-graduate courses (2.1 ECTS)

- Bayesian statistics; PE&RC (2014)
- Introduction to Bayesian methods for quantitative geneticists; Synbreed summer school TUM Institute for advanced study (2015)

Laboratory training and working visits (4.5 ECTS)

- Construction of an analytical pipeline for the detection and the use of QTL in multi-parent populations; KWS SAAT SE (2014-2019)

Competence strengthening / skills courses (3 ECTS)

- The essential of scientific writing and presenting; Wageningen In'to languages (2014)
- Scientific writing; Wageningen In'to languages (2017)

PE&RC Annual meetings, seminars and the PE&RC weekend (1.5 ECTS)

- PE&RC First years weekend (2014)
- PE&RC Last years weekend (2018)

Discussion groups / local seminars / other scientific meetings (4.8 ECTS)

- Modelling and statistics network meeting (2014)
- Biometris statistical genetics meeting (2014-2019)
- R users meeting (2016)

International symposia, workshops and conferences (9.1 ECTS)

- 20th EUCARPIA general meeting; Zurich, Switzerland (2016)
- International conference on statistics and big data bioinformatics in agricultural research; Hyderabad, India (2016)
- 4th International symposium on genomics of plant genetic resources; Giessen, Germany (2017)
- 17th EUCARPIA meeting of the biometrics section; Gent, Belgium (2018)

Lecturing / supervision of practical's / tutorials (4.2 ECTS)

- Modern statistics for life science (2015)
- Modern statistics for life science (2016)
- R For statistics (2018)

Supervision of MSc students

- Comparison of models for QTL detection in a NAM population

Acknowledgements

Five years and a few months ago, I started this rich and beautiful journey looking at an empty black board as an alpinist that is facing a steep north face. From the first hesitant steps in a dense forest to the intense satisfaction to solve complex problems and understand what seemed beyond my skills, this trip has been full of great discoveries, intense emotions, and nice meetings. I had the huge privilege to learn and develop myself to ultimately get an idea about science and technology. The trip was more complicated than I expected but the difficulty also made it more interesting, more rewarding, and more beautiful. One of the nicest things about this PhD was the human experience that literally transformed me. The chance to meet so many incredible people was the most amazing part of this experience. All of you enriched my spirit by your originality and your uniqueness, you definitively all stay somewhere in my heart. To thank you all, I would try to remember some of the great moments we lived together.

First of all I want to mention **Fred** who gave me this great opportunity to make a connection between statistics and agriculture. Fred who was not afraid to hire me after I brought him to this semi-legal Ethiopian restaurant in Geneva. Even if things were sometimes a bit rock 'n' roll (which should not be a problem for a music lover like you), you gave me an incredible opportunity to learn a lot of things in this project at the interface between academic research and industry. I appreciated your availability, the great freedom you gave me to develop my ideas and also sometimes to fail. My vision of the PhD is a precious time to explore challenging roads, try to find our way but also reach our limits. With you as supervisor, I could definitively confront myself to the world and progress a lot in life. I also appreciated the freedom you gave me to organise the next part of my career trying to work with the CGIAR centres. From that perspective, working at Biometris was a great platform to reach further exotic destinations. I hope that in the future we will be able to stay connected and collaborate in new exciting projects.

The second very important person I want to warmly thank is **Marcos**. It was a great chance to have you as daily supervisor. I did not immediately notice the benefit to confront my work to a critical look but looking back I have to admit that with you I learned that constructive confrontation is the best way to improve my work. Thanks a lot for your availability and for the patience you showed to let me develop my point of view and for explaining me many concepts that were new to me. During all these moments spent working with you, I could see your deep humanity and notice how important it is to have on my side trustable people like you.

Another very important person for me during this project was **Dietrich**. I have really enjoyed all the opportunities I had to interact with you. I will never forget that every time I came to Einbeck, you always took some time to invite me at your place and find some entertainment and social activities for me. It was always a pleasure to meet your very nice wife **Dorothea** and your sons **Roland**, **Konrad** and **Heinrich**. With you as host, I had the feeling that Einbeck was the most culturally vibrant and exciting place in the world. There, I even had the chance to assist to a representation of Metropolis with a live piano player. Not even the most underground-alternative-sub-culture place in Amsterdam offers this kind of performance.

During my thesis, I also had the chance to have **Valentin** in my supervision team. Your support has been very precious during this process. I really appreciated the way you helped me by always being very concrete and useful in the solutions you proposed. I also want to thank you for the extremely valuable assistance you gave me to develop my programming skills, which was one of the most important I developed during the project. Thank you also to offer me your time when I was in Einbeck. It was a great pleasure to have lunch with you and discover Einbeck surroundings with your family.

I really appreciated the time I spent at KWS because it was always the opportunity to get valuable feed-backs and to look at my project from a different perspective. Thanks a lot **Sofiane** and **Thomas** for your contribution to the scientific richness of this experience by sharing your experience and perspective on the problems I had to solve. I also appreciated the efforts you made to come for working visits in Wageningen. These moments contributed a lot to the quality of our collaboration.

My PhD trip would not have been as beautiful if I would not have had the amazing chance to meet all my fellow colleagues. To be surrounded by people with such a wide spectrum of qualities, experience and personalities was one of the most beautiful things that happen to me. I remember **Daniela** who was the first person to introduce herself to me. After that, I progressively started to feel at home, especially after her offer to organise a fondue party at Bornsesteeg. Thanks a lot Daniela for all the great moments and all the fun we had together. A special thanks for forgetting your hotel keys in Munich and giving me the opportunity to use my great tenor voice to make sure you do not have to sleep outside. Thank you also for being there in more difficult and troubled moments. Thank you to give me the opportunity to meet your wonderful husband **Marcelo (Marcelito)**. I remember many great moments spend in your company like when we were biking in Ameland. I always listen carefully to your predictions concerning my future. I will try to do my best to realize them but I cannot guarantee that next time we meet I will be surrounded by six children. Since we are among Latinos, it is my pleasure to acknowledge **Julio**. I was happy to contribute to the nice atmosphere of Biometris by preparing Daniela and Guus defence parties with you. If I get the opportunity, I would love to visit you in Argentina and explore the pampa in a gaucho mode. I will not forget to bring you as a gift your favourite vegetarian tofu sausage (a small treat from the Netherlands).

A very special thanks to **Antoine** who gave me the honour to be my paranymph despite the uncertainty about his location at that moment. However, I was certain that even if you had to come by bike from Hendaye you would be present. Like between the fox and the Petit prince, there was a progressive familiarization since the first times you met me in a meditative judo mood at the sport centre. Since then, I found a great friend to share quite specific humour like the "Tango corse" or "Fout le camps ou met des housses" but also more personal thoughts about our respective futures. I have the pretension to think that we are a bit "fait du même bois". A unique combination of energy and sensitivity. However, I will probably never be able to compete with your communicative "joie de vivre". It is simple, since you come less often in the morning to give us a "Bisou, bisou!!!", the level of vitamin C in Biometris dropped by at least 30 percent. A big thanks to **Pariya** that also accepted the challenge to be my paranymph. I will try to make sure that your future carrier of Professor will not be compromised due to embarrassing events happening at my defence party. It was a great pleasure to meet you and I will miss this small pleasure to tease you with crappy Iranian clichés. Therefore, I hope that the next time I visit you, the minaret tower will be installed on your roof, the kebab machine will be turning in the kitchen and that we will be able to do some belly dancing on your flying carpet.

It is also the occasion to thank **Nadia**, another person who brought a lot of joy and energy at Biometris. Thanks a lot for bringing us to the sport class, for motivating us to go to the WeDay and for making sure that we respect as much as possible the Dutch tradition of the Borrel. Far away, I will also keep a great memory of **Dominique** and of her "Baudelaurienne" distance to life that gave her this very subtle sense of irony that made us laugh a lot. Efcharisto poly **George** and **Katarina** for offering me this touch of Mediterranean taste. I loved the evenings we spent together around some of the tastiest food that I could get here in the Netherlands. Thank you to make me discover Tsipuro, the secret of Greek good mood. Hartelijk bedankt **Frederik**. It was a great pleasure to live with you and share all these tasty bi-lingual meals. Un grand Merci to **Emilie** and **Grégoire** who made a substantial contribution to the tastiness of my life with some cheese and of course with the delicious Quentovic. After you build the pipeline between Le Toucquet and Wageningen, you should consider exporting in India.

I would like to express a deep thought full of emotion to my father **Christian**. Papa, without your continuous and precious support, I would not have been able to achieve what I did. Real achievements are not the results of a few stunts, as admirable and impressive they are. Achieving something is the consequence of a long and patient dedication, the repetition of small and concrete gestures. You are this kind of humble person who put their effort in basics but essential things without looking for spotlights. Your patient and concrete help contributed to truly make a difference in my life. As I already told you few times, I miss the words to express all the love and gratitude I feel for you. It is also with emotion that I remember my mother **Béatrice**. Beyond your death, you stay more than ever living in my heart. The values and ideals you gave me such as the duty to have compassion for people in need and the obligation to not look away from problems, have deeply contributed to make the person I am. I also would like to associate **Irène** in these thoughts. I appreciate the love and the care you give to my father. Coming back to Switzerland was always a great source of rest especially because of the warmth and cosiness you put in the house.

In the closest circle of my family, I think about my brother **Maxence** and his wife **Ana**. Despite we now live far from each other, the occasions to meet are always full of colourful experiences and incredible stories. I am extremely proud of the things you realize in Mexico. The beautiful bread and pastries that you make with your wonderful team are representative of the effort and the dedication you put in your work.

Many thanks also to my aunts **Francine** and **Odette**. I am very grateful for all the years when you looked after me and Maxence. You were present in sensitive and difficult moments for us and you had enough patience and understanding to let us express our personalities. From you, I particularly appreciate a high idea of culture that you transmitted to us through great books, classical music, and a very deep knowledge of the human soul.

Thank you **Dominique** my uncle and godfather, the one who understands that

"Sur l'oreiller du mal c'est Satan Trismégiste
Qui berce longuement notre esprit enchanté,
Et le riche métal de notre volonté
Est tout vaporisé par ce savant chimiste."

(Charles Baudelaire, *Au lecteur*)

As well as **Gérald** another poet who knows this kind of state where

"Et dès lors, je me suis baigné dans le Poème
De la Mer, infusé d'astres, et lactescent,
Dévorant les azurs verts ; où, flottaison blême
Et ravie, un noyé pensif parfois descend"

(Arthur Rimbaud, *Le bateau ivre*)

I also think about my aunts **Catherine** and **Romaine**. The beautiful moments we spent together, your attentions and kind messages always represented precious source of support and constitutes exceptional memories.

It is also the occasion and a great pleasure to acknowledge my friends from Switzerland. **Julia** and **Aude** you are not only the best party mates but the most precious confidants. You were always there for me during the good but also the bad moments. All those small projects, those costumes, those dances but also more simple times where we shared our feelings have greatly contributed to my happiness. I was happy to live many of these moments with **Camille**, **Jennifer** and **Julien** who were always there to share our deliriums. We have still so many things to live together.

An enormous thanks to **Adrien** who by is unique sense of humour and irony always brought me joy and sunshine in the monotony of a day in front of the computer. With you, I can speak in this very particular type of humour made of local and realistic references. You have this love of good words and this subtlety that allows you to catch the fun and the nonsense in the most common situations. We had the chance to sometimes share that with **Matthieu** another artist who understands the "Génie populaire".

Some thoughts also for **Steve**, my birthday brother, my best partner in crime. It is not even possible for me to remember all the small adventures and plans we made, like trying to smoke in underwear with a shower cap to not be caught by our parents. With you, I could find someone who was ready to follow me in the most crappy ideas. This definitively contributed to my creativity and my imagination. Through you, I also had the chance to meet **Vincent (Gros)** and **Grégoire**, our exploration of the sense and the pleasure of life in memorable trips in Amsterdam and Brussels are very precious memories to me.

The friendship of **Vincent (Corsat)** was also a very stable and precious element during all those years spent abroad. Vincent you are for sure one of the most trustable friends. You will be the one I call if I have to cover up a murder. After that we will for sure have a drink with **Marcos (Cosi)**. We definitively have some experience in crime since middle school and the big plans we made for some of our teachers. Hopefully, we became a bit more wise but next time I promise we will go to the end of this story with the Princess of Denmark.

It was also a great pleasure and an honour to meet all the beautiful people from the ICC and the choir. Thanks a lot **Tammy, Sam, Frico, Tristin, Rennie, Stan, Giacomo, Calvin, Tom, Rosio, Monica, Dominica, Father Henry** and all the others. For me you represent one of the best things I could find in Wageningen: the combination between an incredible palette of culture with people coming from all around the world and a nice village atmosphere where people can get to know each other. Thanks for all the beautiful moments and the great emotions we shared. You learned me to be "pelan-pelan", you learned me the Ghanaian way to do business, and I could finally find, at Father Henry's table, a place to drink good wines in the Netherlands.

I also would like to thank the Voorhaar-Lim family. Thank you so much to **Mayke** and **Frits** for all the gifts and attention you gave me. I have immediately felt being part of the family. Thank you to make me discover a bit of the Indonesian culture too. Many thanks to **Wenda** and **Maurice**, I appreciated the nice moments we spent together and the way you made me feel welcome in your family. Even if we will live far away, I hope I could still spend many birthdays in your company. I was very touched by all the presents you offered me for my 34th anniversary. As my father said, I have not received as many gifts since I was two.

Now that it is time to leave the Netherlands, I am also thinking with emotion to the family Faber-van Amstel-Kuiper. You were my first family when I arrived in the Netherlands. I immediately felt at home in a new world full of taste and beauty. Thanks a lot **Hylda**, **André** and **Henriette** for welcoming me in your lovely house and for giving me an introduction in many aspects of the Dutch culture. I also remember all the great moments spent with **Antoinette**. Know that I will never forget your sweet personality made of kindness and of determination. I am very proud of you and the things you have accomplished. I also would like to thank **Thérêse** for the path we walked together. It was beautiful and sunny. From you, I will always keep in my heart a very unique sense of elegance and beauty.

To conclude these acknowledgements, I would like to say few words about the farmers that I had the chance to meet in my life. These people are at the origin of my willingness to work in agriculture. All my gratitude goes to **Hans, Brigitte, Pius, Judith, Daniel, Colette, Bernard (Beno), Mick, Mickael, Pia, and Pierrette**. You and all the other farmers, your predecessors and the ones that today continue to shape our world, have been for me a continuous and powerful source of motivation. Agriculture is not only one of the most fundamental activity that provides food for humanity. Agriculture is our relationship to Nature, an introduction to the essence and the mystery of this world. By their dedicated works, by their self-sacrifice and the constant repetition of small gestures, farmers extracted the most beautiful gift Nature provides us.

"Tu m'as donné ta boue et j'en ai fait de l'or"

(Charles Baudelaire, *Ebauche d'un épilogue pour la deuxième édition des Fleurs du Mal*)

Many thanks to **Marta de Menezes** who graciously let me used her art work to design the cover of my thesis. For me, this specific piece represents a certain idea of creativity when a vision becomes concrete, a space where agriculture meets art through its capacity to shape the world and produce, in cooperation with Nature, some of the most beautiful landscapes.

The richness and the diversity of our food, the colour and the warmth of our culture and traditions are the reflection of this intimate relationship made of love, suffering, and passion. For those who take the time to look at the farmers' noble work, they could see a universe in constant interaction with the beauty and the poetry of the world.

"En toi je tomberai, végétale ambroisie,
Grain précieux jeté par l'éternel Semeur,
Pour que de notre amour naisse la poésie
Qui jaillira vers Dieu comme une rare fleur !"

(Charles Baudelaire, *L'âme du vin*)

The final and the most special words go to **Marijn**. You are the most subtle and delicate flower in the joyful chaos of my life. This oasis of rest, warmth and peace where I can get some fresh energy and new inspiration. Through all your attentions, you progressively showed me the infinite beauty of your soul. You contribute to make the world a place full of flavour, colours, and rich emotions. A place where order, beauty and taste shine. I love you with all my heart and I am deeply grateful for finding you on my road. I am looking forward for the next part of this life trip with you on my side. I promise you the warmest sun in the deepest sky.

Thank you again to all of you for making from those years in Wageningen the most beautiful of my life.

Funding

The research described in this thesis was financially supported by KWS SAAT SE.

Thesis cover

Cover designed by Vincent Garin. Photomontage with 'In the beginning there was the word' work by Marta de Menezes, 2010. Wheat growing on a bible.

Printing

This thesis was printed by Digiforce, Vianen, the Netherlands