Statistical methods for QTL mapping and genomic prediction of multiple traits and environments: case studies in pepper

Nurudeen Adeniyi ALIMI

Thesis committee

Promotor

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

Co-promotor

Dr M.C.A.M. Bink
Senior Quantitative Geneticist
Hendrix Genetics - Research & Technology Centre, Boxmeer

Other members

Prof. Dr R.F. Veerkamp, Wageningen University
Prof. Dr P.C. Struik, Wageningen University
Dr R. Rincent – INRA Clermont Ferrand, France
Dr E.J. Gutteling, Rijk Zwaan, Fijnaart

This research was conducted under the auspices of the Graduate School for Production Ecology & Resource Conservation (PE-RC).

Statistical methods for QTL mapping and genomic prediction of multiple traits and environments: case studies in pepper

Nurudeen Adeniyi ALIMI

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 Tuesday 1 November 2016

at 1.30 p.m. in the Aula.

Nurudeen Adeniyi ALIMI

Statistical methods for QTL mapping and genomic prediction of multiple traits and environments: case studies in pepper

165 pages.

PhD thesis, Wageningen University, Wageningen, NL (2016)

With references, with summary in English

ISBN: 978-94-6257-936-1
DOI: http://dx.doi.org/10.18174/390205

To the past - *Maami and Magaji*

The present – *'Mobola and Kiyaan*

And the future

Abstract

In this thesis we describe the results of a number of quantitative techniques that were used to understand the genetics of yield in pepper as an example of complex trait measured in a number of environments. Main objectives were; i) to propose a number of mixed models to detect QTLs for multiple traits and multiple environments, ii) to extend the multi-trait QTL models to a multi-trait genomic prediction model, iii) to study how well the complex trait yield can be indirectly predicted from its component traits, and iv) to understand the 'causal' relationships between the target trait yield and its component traits.

The thesis is part of an EU-FP7 project "***S***mart tools for ***P***rediction and ***I***mprovements of ***C***rop ***Y***ield" (SPICY- http://www.spicyweb.eu/). This project generated phenotypic data from four environments using 149 individuals from the sixth generation of recombinant inbred lines obtained from intraspecific cross between large – fruited inbred pepper cultivar 'Yolo Wonder' (YW) and the hot pepper cultivar 'Criollo de Morelos 334' (CM 334). A total of 16 physiological traits were evaluated across the four trials and various types of genetic parameters were estimated. In a first analysis, the traits were univariately analyzed using linear mixed model. Trait heritabilities were generally large (ranging between 0.43 – 0.96 with an average of 0.86) and mostly comparable across trials while many of the traits displayed heterosis and transgression. The same QTLs were detected across the four trials, though QTL magnitude differed for many of the traits. We also found that some QTLs affected more than one trait, suggesting QTL pleiotropy (a QTL region affecting more than one trait). We discussed our results in the light of previously reported QTLs for these and similar traits in pepper.

We addressed the presence of genotype-by-environment interaction (GEI) in yield and the other traits through a multi-environment (ME) mixed model methodology with terms for QTL-by-environment interaction (QEI). We opined that yield would benefit from joint analysis with other traits and so deployed two other mixed model based multi-response QTL approaches: a multi-trait approach (MT) and a multi-trait multi-environment approach (MTME). For yield as well as the other traits, MTME was superior to ME and MT in the number of QTLs, the explained variance and accuracy of predictions. Many of the detected QTLs were pleiotropic and showed quantitative QEI. The results confirmed the feasibility and strengths of novel mixed model QTL methodology to study the architecture of complex traits.

The QTL methods considered thus far are not well suited for prediction purposes as only a limited set of QTL-related markers are used. Since the main interest of this research includes improvement of yield prediction, we explored both single-trait and multi-trait versions of genomic prediction (GP) models as alternatives to the QTL-based prediction (QP) models. This was termed direct prediction. The methods differed in their predictive accuracies with GP methods outperforming QP methods in both single and multi-traits situations. We borrowed ideas from crop growth model (CGM) to dissect complex trait yield into a number of its component traits. Here, we integrated QTL/genomic prediction and CGM approaches and showed that the target trait yield can be predicted via its component traits together with environmental covariables. This was termed indirect prediction. The CGM approach seemed

to work well at first sight, but this is especially due to the fact that yield appeared to be strongly driven by just one of its components, the partitioning to fruit.

An alternative representation of the biological knowledge of a complex target trait such as yield is provided by network type models. We constructed both conditional and unconditional networks across the four environments to understand the 'causal' relationships between target trait yield and its component traits. The final networks for each environment from both conditional and unconditional methods were used in a structural equation model to assess the causal relationships. Conditioning QTL mapping on network structure improved detection of refined genetic architecture by distinguishing between QTL with direct and indirect effects, thereby removing non-significant effects found in the unconditional network and resolving QTL pleiotropy. Similar to the CGM topology, yield was established to be downstream to its component traits, indicating that yield can be studied and predicted from its component traits. Thus, the genetic improvements of yield would benefit from improvements on the component traits.

Finally, complex trait prediction can be enhanced by a full integration of the methods described in the different chapters. Recent research efforts have been channelled to incorporating both multivariate whole genome prediction models and crop growth models. Further research is required, but we hope that the present thesis presents useful steps towards better prediction models for complex traits exhibiting genotype by environment interaction.

Table of Contents

Abstract vi
Table of Contents viii
1. General Introduction 1
2. General Introduction 3
1.1. Background 3
1.2. Mapping Populations 4
1.3. Mapping Techniques 4
1.3.1. ANOVA model 5
1.3.2. Interval Mapping/Mixture model 5
1.3.3. Multiple Linked QTLs 5
1.3.4. (Multivariate) Multiple Regression 6
1.3.5. Mixed Model 6
1.3.6. Probabilistic Models Based on Bayesian Theory 7
1.4. Genomic Prediction Models 8
1.5. Phenotype Network Models 9
1.6. QTL by Environment Interaction 10
1.7. QTL mapping in Pepper 10
1.8. Overview/Outline of this thesis 11
Genetic and QTL analyses of yield and a set of physiological traits in pepper 13
3. Genetic and QTL analyses of yield and a set of physiological traits in pepper 15
2.1. Abstract 15
2.2. Introduction 16
2.3. Materials and Methods 17
2.3.1. Plant materials 17
2.3.2. Phenotyping experiments and designs 17
2.3.3. Trait evaluation 18
2.3.4. Phenotypic analysis 19
2.3.5. Marker data and Linkage map 20
2.3.6. QTL estimation 21
2.4. Results 22
2.4.1. Traits evaluation 22
2.4.2. Heritability & genetic correlation 23
2.4.3. QTL results 25
2.5. Discussion 29

2.6. Conclusion 32
Appendix 2A: EU-SPICY Experimental set-up 33
Appendix 2B: Phenotypic Mean comparison 35
4. Multi-Trait and Multi-Environment QTL Analyses of Yield and A Set of Physiological Traits in Pepper 37
5. Multi-Trait and Multi-Environment QTL Analyses of Yield and A Set of Physiological Traits in Pepper 39
3.1. Abstract 39
3.2. Introduction 40
3.3. Materials and Methods 41
3.3.1. Plant materials, marker data and phenotypic evaluation 41
3.3.2. Multi-environment phenotypic and QTL analysis 42
3.3.3. Multi-trait QTL estimation 43
3.3.4. Multi-Traits Multi-Environments QTL estimation 44
3.3.5. MTME final QTL selection and window size 44
3.3.6. Comparisons of ME, MT and MTME approaches 45
3.4. Results 45
3.4.1. Genetic correlations between traits (within and between trials) 45
3.4.2. Multi-Environment analyses 45
3.4.3. Multi-trait analyses 46
3.4.4. Multi-trait multi-environment analysis 48
3.4.5. Comparison of MT, ME, and MTME results 49
3.5. Discussion 55
Appendix 3A: Traits Description 59
Appendix 3B: Biplots for BLUEs and fitted trait values 60
Appendix 3C: QTL by Environment Results from the ME analyses 60
Appendix 3D: QTL Effects from the MT analyses 64
Appendix 3E: QTL Effects from the MTME analyses: Chromosomes 3-12 66
4. Predicting complex traits in multiple environments by a combination of genomic prediction and crop growth modelling: an example in pepper 71
8. Predicting complex traits in multiple environments by a combination of genomic prediction and crop growth modelling: an example in pepper 73
4.1. Abstract 73
4.2. Introduction 74
4.3. Materials and Methods 76
4.3.1. Genotypic and Phenotypic data 76
4.3.2. Univariate and Multivariate QTL Prediction Models 76
4.3.3. Bayesian Genomic Prediction Models 77

4.3.4. Univariate Genomic Prediction Model: Bayesian LASSO Regression 78
4.3.5. Multivariate Genomic Prediction Model: Bayesian Latent Variable 78
4.3.6. Yield Indirect Prediction through Crop Growth Model 79
4.3.7. Model Validation and Accuracy 80
4.3.8. Yield Prediction Strategies 81
4.4. Results 82
4.4.1. Trait Descriptions 82
4.4.2. Predictive Ability and Bias of the Four Prediction Models 82
4.4.3. Accuracies of Yield Prediction from Crop Growth Model 85
4.5. Discussion 88
5. A network analysis of yield and yield components across environments: an example in pepper 93
6. A network analysis of yield and yield components across environments: an example in pepper 95
5.1. Abstract 95
5.2. Introduction 96
5.3. Materials and Methods 97
5.3.1. Genotypic and Phenotypic data 97
5.3.2. Traits Relationships from Crop Growth Model 98
5.3.3. Unconditional Network (MTM) 98
5.3.4. Conditional Network (QTLnet) 99
5.4. Results 101
5.5. Discussion 107
Appendix 5A 110
Appendix 5B 111
6. GENERAL DISCUSSION 113
7. GENERAL DISCUSSION 115
6.1. Introduction 115
6.2. Mapping Population 116
6.3. QTL mapping resolution 118
6.4. Manual and Automated Phenotyping 119
6.5. Complex Traits Analyses 120
6.5.1. QTL methods based on linear mixed model 121
6.5.2. Genomic Prediction and Integrated Crop Growth Models 126
6.5.3. Causal network model 129
6.6. Concluding Remarks 131
References 133
Summary 146

Acknowledgement 149
Curriculum Vitae 151
Publication List 152
PE&RC Training and Education Statement 153

CHAPTER 1

General Introduction

CHAPTER 1

General Introduction

1.1. Background

The production of genetically improved crop cultivars capable of satisfying human requirements such as yield, quality, tolerance to certain environmental conditions and disease resistance has always been the main challenge for plant breeders. The breeder must identify and select superior genotypes capable of conferring the desired requirement(s) on the plant. This is a result of often complex genetics that underlie the expression of most of the economically important traits. Since most of the observable phenotypic variations between individual plants from the same species are quantitative, the development and application of quantitative genetics theory in the last century has greatly improved the understanding of the genetic basis of complex traits. In quantitative genetics, the genetic architecture of a quantitative trait is described with the aid of an underlying genetic model. Variation in (complex) quantitative traits is caused by segregation at multiple loci with individually small effects (polygenic) that may be sensitive to the environment (Mackay, 2001). For such complex traits, the quantitative trait loci (QTL) genotypes cannot be determined from segregation of phenotypes in controlled crosses or pedigrees because the relationship between genotype and phenotype is not a simple ratio.

Until recently, the understanding of complex traits has been developed without having direct access to the DNA, the place where the QTLs responsible for genetic variation ultimately reside. The availability of marker genotyping that provide information directly related to the DNA opened new possibilities for the further development of genetic models that included explicit representations of the hereditary material. With molecular genetics, it is expected that information at the DNA level will lead to faster genetic gain than that achieved based on phenotypic data only. The application of molecular markers has enabled the dissection of complex traits into the underlying QTLs. These molecular markers do exhibit Mendelian segregation (Uptmoor et al., 2008). The increasing knowledge on QTLs for important agronomic traits gives new opportunities in marker-assisted selection (MAS) (Ribaut and Hoisington, 1998; Uptmoor et al., 2008). The use of these molecular breeding techniques has considerably contributed to unravel crop traits affecting quality and yield of plant products and to gain insight into their genetic basis. The basic principle behind the use of MAS in the context of QTL mapping can be expressed as: *If a QTL is linked to a marker locus, there will be a difference in mean values of the quantitative trait among individuals with different genotypes at the marker locus* (Sax, 1923). Among the most popular types of molecular markers employed in QTL mapping are Restriction Fragment Length Polymorphisms (RFLP) (Beckmann and Soller, 1986; Tanksley et al., 1989), Simple Sequence Repeats (SSR) (Powell et al., 1996), Amplified Fragment Length Polymorphisms (AFLP) (Vos et al., 1995) and Single-Nucleotide Polymorphism (SNP) (Syvänen, 2005). Each molecular marker system has

its own advantages and disadvantages (Semagn et al., 2006), the focus of which is not within the scope of this work. Nowadays genomes have been sequenced and commercial SNP arrays are available for many field and horticultural crops which makes genome-wide genotyping affordable.

1.2. Mapping Populations

The types of populations most commonly used for QTL mapping are segregating populations originating from crosses between inbred lines such as F_2, backcross, or recombinant inbred line (RIL). This is mainly due to the possibility of producing relatively large populations with known genetic structure, as there are only two founder genotypes (Robertson, 1967). These population types have been in use long before the advent of molecular markers. Decisions on selection of parents and mating design for development of mapping population and the type of markers used depend upon the objectives of experiments, availability of markers and the molecular map. The parents of mapping populations must have sufficient variation for the traits of interest at both the DNA sequence and the phenotype level. The variation at DNA level is essential to trace the recombination events. The more DNA sequence variation exists, the easier it is to find polymorphic informative markers. When the objective is to search for genes controlling a particular trait, genetic variation of trait between parents is important. If the parents are greatly different at phenotypic level for a trait, there is a reasonable chance that genetic variation exists between the parents, although uncontrolled environmental effects could create large phenotypic variation without any genetic basis for the effects. However, lack of phenotypic variation between parents does not mean that there is no genetic variation, as different sets of genes could result in same phenotype (Mackay, 2001; Ribaut and Hoisington, 1998). Other types of population used in plant breeding include pedigree population (Bink et al., 2002; Rosyara et al., 2009), association panels (Jannink and Walsh, 2002) and Multi-parent Advanced Generation Inter-Cross (MAGIC) population (Cavanagh et al., 2008; Huang et al., 2012).

1.3. Mapping Techniques

Before the development of mapping techniques, the knowledge about the genetic architecture of quantitative traits was limited to estimates of trait heritability and other variance components derived from correlations between relatives and response to selection, estimates of average degree of dominance from changes of mean on inbreeding, estimates of net pleiotropic effects from genetic correlations, and estimates of the total mutation rate from phenotypic divergence between inbred lines. There was the need to go beyond these mere statistical descriptions in order to more effectively select domestic crop species for improved production traits, and understand the genetic basis of adaptation (Mackay, 2001). The need to identify and determine the properties of the individual genes underlying variation in complex traits (Jannink et al., 2001) led to increasing improvements in statistical techniques for QTL mapping, and experimental design.

Within the last two decades, many QTL mapping methods have been developed either based on least square (LS) or maximum likelihood (ML) estimation and recently based on Bayesian paradigm. LS methods test for differences in means between marker class using either

ANOVA or regression (Soller et al., 1976), while ML uses full information from the marker-trait distribution, and explicitly accounts for QTL data being mixtures of normal distributions.

Each of these estimation techniques come with their advantages and disadvantages. However there is, in general, little difference in power between the two techniques (Haley and Knott, 1992; Lander and Botstein, 1989a) and ML interval mapping can be approximated using regressions (Haley and Knott, 1992; Martinez and Curnow, 1992).

1.3.1. ANOVA model

This is the most basic form of QTL mapping comparing the means of the marker genotypes for individual marker loci, under the hypothesis that the marker loci coincide with a QTL (Soller et al., 1976). The marker genotypes define the levels of a treatment factor, an analysis of variance is then performed and marker-trait associations tested using standard F-tests. This model can be easily extended to accommodate QTL interactions and fixed effects. A major drawback is that assumptions of homogeneity (perfect linkage disequilibrium) may be violated. QTL effect and QTL location may be confounded in terms of distance to the marker i.e. a closely linked QTL with a moderate allele effect and a major QTL that is loosely linked will produce a comparable test statistic for marker-trait association.

1.3.2. Interval Mapping/Mixture model

In 1989, Lander and Botstein (Lander and Botstein, 1989a) proposed the use of genetic map information to overcome the limitation of the individual marker approach in a strategy called interval mapping (IM) using ML estimation. ML estimates for the model parameters are obtained with the assumption that observations are from a mixture of normal distributions (one distribution per QTL genotype class). Though the QTL genotypes are unknown in between markers, the flanking markers can be used to infer conditional probabilities for the QTL genotypes given the flanking marker genotypes and the recombination frequencies between the QTL and the markers. The conditional probabilities for the QTL genotypes are then used as mixing proportions in the calculation of the likelihood for the mixture model. Likelihood ratio (LR) test is performed to determine whether the phenotypic data support a mixture distribution, i.e., the presence of a QTL at the evaluation position in the genome. Typically, the log-likelihood ratio (LLR), or LOD score (= LLR/4.61) is plotted along the genome as a profile.

1.3.3. Multiple Linked QTLs

The presence of multiple linked QTLs biases both single marker and interval mapping analysis (Knott and Haley, 1992), and segregation of unlinked QTLs inflates the within-marker class phenotypic variance, thus reducing the power of QTL detection. This led to further improvements to the IM approach through composite interval mapping (CIM) (Jansen and Stam, 1994; Zeng, 1994) and multiple interval mapping (MIM) (Kao et al., 1999). Composite interval mapping (CIM) combines ML interval mapping with multiple regression, using marker cofactors to reduce the bias in estimates of position and effects of QTLs introduced by multiple linked QTLs. CIM also leads to an increase in QTL detection power since the within marker-class phenotypic variation is decreased. Strictly speaking, CIM

methods are not multiple QTLs methods, in that the model for evaluating the effects of each interval depends on the marker cofactors included, which varies across intervals (Mackay, 2001). Multiple interval mapping (MIM) is a true multiple QTLs method. It converges to a stable model providing estimates of positions and main and interaction effects of multiple QTLs (Kao et al., 1999). It should be noted that in CIM and MIM methods, estimates of QTL positions and effects are highly model dependent, and can vary given different numbers of marker cofactors and window sizes (the region to either side of the test interval within which no marker cofactors are fitted) (Pasyukova et al., 2000; Zeng et al., 1999; Zeng, 1994). These factors are under the control of the researcher, who must bear in mind that the model with the best fit and the the highest number of identified QTLs, is not necessarily the closest approximation to reality (Mackay, 2001).

1.3.4. (Multivariate) Multiple Regression

With more than one QTL, the use of mixture models becomes computationally intensive and less versatile. The use of multiple regressions QTL mapping approach was therefore proposed as a more efficient and computationally less intensive alternative. The regression approach has been shown to produce very similar results to the mixture model strategy (Haley and Knott, 1992; Kao, 2000) and can be implemented within standard statistical packages. Regression can also be employed for a complex data structure having multi-QTL and multi-environment, with possibility to model QTL by environment interaction (Jiang and Zeng, 1995; Sari-Gorla et al., 1997).

Extensions to the multi-trait case have been proposed for both mixture and regression based models (Hackett et al., 2001; Jiang and Zeng, 1995; Knott and Haley, 2000). CIM was extended to multiple traits, enabling the evaluation of the main QTL effects as well as pleiotropy and QTL by environment interactions (Jiang and Zeng, 1995). The multi-trait extension of regression based framework was also proposed and implemented (Hackett et al., 2001; Knott and Haley, 2000). Multivariate multiple regression approaches do show greater flexibility than mixture models in extensions to account for additional treatment and block structure, they are not yet robust enough to account for commonly encountered complications as imbalance and complex error structures (Malosetti, 2006).

1.3.5. Mixed Model

Some of the issues in QTL detection and analyses involve the underlying design of the phenotypic experiment which may induce unequal replication of genotypes (unbalanced) and/or measurements over time (repeatedness). Also, most of the experiments involve collections of genotypes evaluated for multiple traits across multiple environments. It is also possible that the relationship between measured traits and explanatory variables such as genotype and environment characterizations is not well captured by a linear assumption. Mixed models (Verbeke and Molenberghs, 2000) offer a suitable framework to jointly analyse such data without imposing unrealistic assumptions, such as zero genetic correlations between environments and traits, and constant variance across environments. Mixed model is also capable of accounting for possible unbalanced design setting and repeatedness. They can account for both intra- and inter-trial variability in the estimation of QTL effects and trait

values prediction and facilitate the representation of genetic relationship among related lines thereby offering a condition for valid inference on QTLs (Van Eeuwijk et al., 2010). The linear assumption can be relaxed and the relationship modelled by non-linear functions with inclusion of growth parameters thereby mimicking eco-physiological models. Mixed model can be applied to several settings commonly found in plant breeding experiments. The simplest of such settings is single-trait-single-environment which can be extended up to the most complex setting of multi-trait (possibly correlated) multi-environment setting with various interactions (traits, environments and/or environmental characterization). Most of these settings with several structures for expectations, correlations and variance-covariance have been analysed in literature.

1.3.6. Probabilistic Models Based on Bayesian Theory

Most genetic properties of plants and animals, individuals, populations or species are a product of processes that are inherently stochastic and are mostly interdependent. Therefore, they are better studied using probabilistic models. In the Frequentist approach, probability is viewed from the framework of hypothetically repeating an experiment many times under identical circumstances. In the Bayesian approach, a probability is a direct measure of uncertainty, and might or might not represent a long-term frequency. Bayesian and Frequentist statistics aim at making inferences about a fixed, but unknown, parameter value but they differ in approach and in interpretation of the results. Bayesian analysis incorporates background (prior) information into the specification of the model. This prior information is combined with information from the data (likelihood) to generate the posterior distribution over the parameter values, according to Bayes' rule. The choice of prior information can be based on previous experiments, experts input, theoretical or other considerations. Bayesian methods can be especially valuable in complex problems or in situations that do not conform naturally to a classical setting. Many genetics problems fall into one of these categories (Shoemaker et al., 1999). In addition, Bayesian approaches can be easier to interpret. The paper of Beaumont and Rannala (2004) reviewed the application of Bayesian inference in some areas of genetics including population genetics, genomics and human genetics with specific reference to analysis of complex trait, linkage mapping and QTL mapping. In QTL analyses, inference is typically concerned with identifying those loci on the genome that contribute significantly to the quantitative trait of interest. Through Bayesian approach, the probability that a locus positioned near a known molecular marker has a genotype directly associated with the trait can be calculated and the QTLs which directly influence the trait can be identified.

The Bayesian approach has been successfully applied in a wide range of applications. Bayesian analysis based on QTL intensity has been proposed for obtaining posterior modes and credibility intervals for the QTLs (Sillanpää and Arjas, 1998). Various Bayesian techniques for handling complex plant and animal pedigreed populations have been suggested and implemented. Sisson and Hurn (2004) discussed existing approaches to the use of Bayesian model in making inference on QTLs and suggested a modification to the loss function for estimating both the number of QTL and their location. Bauer et al. (2009) developed a Bayesian multi-locus multi-environmental method of QTL analysis. Through a

real life data and simulation study, the strategy was compared to (a) Bayesian multi-locus mapping, where each environment is analyzed separately, (b) Restricted Maximum Likelihood (REML) single-locus method, using a mixed hierarchical model, and (c) REML forward selection applying a mixed hierarchical model.

Just as in regression analysis, model selection can also be handled within the Bayesian framework. Here, the model selection problem is transformed to the form of parameter estimation. Several Bayesian model selection methods have been developed among which are Kuo & Mallick, Gibbs Variable Selection (GVS), Stochastic Search Variable Selection (SSVS), adaptive shrinkage and model space approach (reversible jump MCMC and composite model space) (O'Hara and Sillanpää, 2009). Yi et al., (2007) extended the Bayesian model selection framework they earlier proposed for mapping epistatic QTLs in experimental crosses to include environmental effects and gene–environment interactions. A new fast Markov chain Monte Carlo algorithm was proposed to explore the posterior distribution of unknowns. In addition, this takes advantage of any prior knowledge about genetic architecture to increase posterior probability on more probable models.

1.4. Genomic Prediction Models

The availability of genome-wide dense marker maps at affordable cost have made the use of genomic selection (GS) models an interesting alternative to marker-assisted models. GS models predict the genetic value of selection candidates based on the genomic estimated breeding value (GEBV) predicted from high-density markers positioned throughout the genome. Unlike marker-assisted selection, the GEBV is based on all markers including both minor and major marker effects. Thus, the GEBV may capture more of the genetic variation for the particular trait under selection. The GS models have become the standard methods for predicting genetic values in animals (De Los Campos et al., 2009; Goddard and Hayes, 2009) and also recently in plants (Crossa et al., 2010; Heslot et al., 2012; Jannink et al., 2010). GS models can be based on either Frequentist or Bayesian paradigm. Unlike QTL-based models where selected markers are used, in the GS models, all markers are used in a penalized regression context for prediction.

The key principle of GS is to simultaneously estimate the effects of all genome-wide markers in a training population consisting of genotyped and phenotyped individuals and then predict the genomic estimated breeding value (GEBV) of genotyped but not-phenotyped individuals in test/future generations (Meuwissen et al., 2001). GEBVs are calculated as the sum of estimated marker effects for genotyped individuals in a training population. Fitting all markers simultaneously ensures that marker-effect estimates are unbiased, small effects are captured, and there is no multiple testing issue (Jia and Jannink, 2012). Due to the usually large number of markers relative to number of individuals, variable selection and shrinkage estimation methods are employed to tackle the problem of high dimensionality in the predictors (De Los Campos et al., 2009; Habier et al., 2011; Hayes et al., 2009; Legarra et al., 2011). These estimation methods try to reduce mean squared error (MSE) by reducing the variance of the estimator. This may however introduce bias in the estimate. The obtained penalized estimates are the solution to an optimization problem that balances model fit and

model complexity. Both parametric and semi-parametric GS methods have been proposed to handle the problem of high dimensionality and other peculiar issues including markers colinearity. Some example of GS methods include ridge regression (Hoerl and Kennard, 1970), Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996), Bayes A and Bayes B of Meuwissen et al. (2001) and the Bayesian LASSO (Park and Casella, 2008), reproducing kernel Hilbert spaces (RKHS) regression (Gianola and van Kaam, 2008). Evaluations and comparisons of performances of a number of GS models in plant breeding is presented in Heslot et al. (2012).

1.5. Phenotype Network Models

Understanding the interconnectedness among plant phenotypes has become a key objective in QTL mapping. The vast opinions in recent literature advocate for the need to study how plant phenotypes are interconnected in networks of dependencies and the stability of the relationships across environments due to genotype-by-environment interactions (Granier and Vile, 2014; Li et al., 2010; Valente et al., 2013). Complex traits are often associated with multiple correlated traits referred to as component traits. Physiological interactions among target and component traits, together with shared genetic factors may be responsible for observed associations among these traits (Li et al., 2006). The genetic improvements of a complex target trait would benefit from improvements on the component traits, especially when the mechanism of (causal) association between the target and component traits is known. Although traditional multi-trait models are able to account for covariations among traits and establish QTLs with pleiotropic effects, they are not able to disentangle the paths for such effects neither are they able to provide insight into the (causal) relationships among the traits. Properly studying the interconnectedness will reveal causal relationship among phenotypes.

Causal inference methodology was introduced as early as the 1921 (Wright, 1921). The methods have been further developed and applied since then in genetics and other fields (Spirtes et al., 2000). Incorporating QTLs in network models has been shown to facilitate causal inference (Li et al., 2006; Neto et al., 2008), enabling differentiation of QTL effects on phenotypes into direct and indirect effects. QTLs in network models also provide an intuitive explanation for pleiotropic QTLs and possible QTL hotspots region where a QTL influences many traits. Graphical models with arrows pointing in the direction of causality are often used to depict the inferred relationship (Neto et al., 2010). However, causality claim cannot be established from data alone. Some assumptions about the causal relationships among the variables being modelled are needed. In genomic network studies, causality claim stems from two facts. First is the analogy between randomized experimental design and genetic randomization that occurs during meiosis. Second is the intuition that phenotypic variation is caused by genetic factors (Li et al., 2006; Neto et al., 2010). Relying solely on correlation between traits to claim causality is not enough even when the traits share a common QTL. Understanding biological reasoning governing the relationship is crucial (Li et al., 2010).

1.6. QTL by Environment Interaction

Yield as an example of complex quantitative traits of plants measured on collections of genotypes across multiple environments is the result of processes that depend simultaneously on genotype and environment in intricate ways (Boer et al., 2007). For complex traits that exhibit considerable genotype by environment interaction, these QTLs have to be analyzed by considering the combination of the QTLs under different environment using QTL x E analysis (Ribaut and Hoisington, 1998; Slafer, 2003). An overview of the state of the art in QTL analyses and crop performance under environmental conditions is provided in the review by Collins et al. (2008). The improvement of crop yield has been possible through the indirect manipulation of QTLs that control heritable variability of the traits and physiological mechanisms that determine biomass production and its partitioning. Also, most QTLs are not stable across environments. QTLs can therefore be categorized according to the stability of their effects across environmental conditions. A '*constitutive*' QTL is consistently detected across most environments, while an '*adaptive*' QTL is detected only in specific environmental conditions or increases in expression with the level of an environmental factor e.g. a QTL that is expressed more strongly with increasing temperature (Vargas et al., 2006). The magnitudes of these adaptive QTLs, therefore, vary greatly between experiments.

Further, complex agronomic traits such as yield have low heritability, are strongly dependent on environmental changes and show high genotype by environment interactions (GEI) (Tardieu, 2003). The genetic analysis of such highly variable traits needs a strategy to cope with the temporal variability of phenotypes. Physiological models could help in understanding GEI interactions and speed up crop improvement for targeted environments (Boote et al., 2001; MAYES et al., 2005; Slafer, 2003). One strategy involves interpreting networks of field trials using a statistical method that calculates QTL x E interactions (Malosetti et al., 2006). Another strategy known as eco-physiological model involves modelling the measured traits by an underlying physiological model of which several non-genetic input variables closely describe the characteristics of the environments (Marcelis et al., 2006; Tardieu, 2003).

1.7. QTL mapping in Pepper

Pepper, a member of the *Solanaceae* family, is a naturally self-pollinating warm season perennial with expected lifespan of about 20 years. It is diploid with 12 chromosome pairs. Most pepper species originated from South America (DeWitt and Bosland, 1996). *Capsicum annuum*, which is the most commercially important and most widely cultivated species worldwide, is used in this study. A number of studies have reported on genetic parameters of a series of pepper traits and performed QTL mapping for these traits. Among pepper traits already studied are those related to disease resistance (Lefebvre and Palloix, 1996; Caranta *et al.*, 1997a), and sensory traits such as pungency (Blum *et al.*, 2003; Ben Chaim *et al.*, 2006a). Other studies have also looked into fruit-related traits (Lefebvre *et al.*, 1998; Ben Chaim *et al.*, 2001b; Rao *et al.*, 2003; Barchi *et al.*, 2009). Results from these studies revealed clusters of QTLs on chromosomes 2, 3 and 4 for fruit traits such as fruit weight/yield, diameter, length and shape. Only few studies have reported on QTLs influencing vegetative-related traits such as stem length and number of internodes (Ben Chaim *et al.*, 2001b; Barchi *et al.*, 2009; Mimura *et al.*, 2010). In these studies, major QTLs influencing primary vegetative

components like length and number of internodes on the primary stem were found on chromosomes 2, 3 and 12. Only one study reported QTLs responsible for leaf related components such as leaf area and weight (De Swart *et al.*, 2007). Also the majorities of these studies were conducted in a single environment and hence could not compare performances of a particular mapping population under different environmental conditions.

1.8. Overview/Outline of this thesis

The work in this thesis aims at increasing our knowledge and understanding of genetic control of complex traits by exploring various statistical methods capable of properly accounting for general and specific features of experimental designs being employed in predicting phenotypic performances of genotypes. Statistical approaches capable of evaluating multiple, correlated and time dependent traits simultaneously as functions of genes (QTLs) and environmental inputs are considered. The main objective is to evaluate and apply QTL models for multiple correlated physiological traits across a range of environmental conditions. Also, we wish to dissect complex trait yield into a number of component traits by defining ecophysiological relationship among yield and its component traits in combination with environment characterizations, and perform QTL analyses on the defined relationship. Finally, we wish to study causal relationships among yield related traits using QTL information to define such relationships.

Chapter 2 presents the first steps in the genetic and QTL analyses of the four big trials in the European Union sponsored FP7 project tagged 'Smart tools for Prediction and Improvement of Crop Yield' (EU-SPICY) (see www.spicyweb.eu and Voorrips *et al.* (2010)). Sixteen physiological pepper traits are univariately analyzed for a population of 149 recombinant inbred lines, obtained from a cross between the large-fruited pepper cultivar 'Yolo Wonder' (YW) and the small fruited pepper 'Criollo de Morelos 334' (CM334). We start with description and phenotypic analyses of the four large phenotyping experiments and obtained genetic parameters for all traits using linear mixed model. For all environments, we use a multiple-QTL mapping (MQM) method to estimate location, heritability and direction of the QTLs. We investigate QTL pleiotropy and we discuss our results in the light of previously reported QTLs for these and similar traits in pepper.

Chapter 3 compares the performance of three multi-response QTL approaches based on linear mixed models: a multi-trait approach (MT), a multi-environment approach (ME), and a multi-trait multi-environment approach (MTME). We model genetic correlations within (between traits in a given environment) and between environments, and explicitly test the presence of QEI and pleiotropic QTLs. The approaches are compared in terms of number of QTLs detected for each trait, the explained variance, and the accuracy of prediction for the final QTL model. In pepper, GEI and QEI approaches have not been used previously to map multiple quantitative traits in multiple environments. Earlier studies focused mostly on univariate analyses of traits in single environments. Many of the QTLs from all the approaches are pleiotropic and show quantitative QTL by environment interactions. MTME is superior to ME and MT in the number of QTLs, the explained variance and accuracy of predictions. A number of guidelines are proposed to obtain a stable final QTL model in the

MTME approach. The results confirm the feasibility and strengths of novel mixed model QTL methodology to study the architecture of complex traits. These results confirm that multivariate analyses of traits have better capabilities to unravel complex traits than single trait approach.

Chapter 4 sets out to satisfy two main research objectives. The first objective relates to comparing performances of QTL prediction (QP) and genomic prediction (GP) methods as predictive models. Both single-trait and multi-trait versions of the QP and GP methods were explored resulting into four prediction models. The predictive performances of the models were characterized using five yield related pepper traits measured across the four environments in the EU-SPICY project. The second objective relates to prediction of the complex trait yield as a function of breeding values of its component traits and environmental variables. This approach was termed indirect prediction in contrast to predicting yield directly from its own breeding values. A LINTUL type (Light INTerception and Utilization) (Spitters and Schapendonk, 1990; Van Ittersum et al., 2003) crop growth model (CGM) was employed to relate yield to three component traits namely light use efficiency (LUE), partitioning into the fruits (PF) and growth rate of leaf area index (LAI_{rate}). This strategy is implemented as within-environment and across-environment (GEI) analyses. We show that yield in an environment can be successfully predicted from its component traits, provided a suitable function relating yield to the component traits is developed. Also, the GEI CGM indicates that in situations where similarities exist among environments, we may use component traits and environmental information from one environment to predict yield in another environment. The results further show that trait's prediction accuracy depends not only on prediction model of choice and traits genetic architecture but also on the environment.

Chapter 5 focuses on exploring correlation networks models in the study of yield related traits using pepper as a case study. Both conditional and unconditional networks are constructed for four yield related traits across a number of environments. The unconditional networks are based on standard multi trait model (MTM) (Jiang and Zeng, 1995) while the conditional networks are based on the QTL-driven phenotype network method (QTLnet) (Neto et al., 2010). The final networks for each environment from both conditional and unconditional methods are used in a structural equation model (SEM) to quantify and compare the relationships depicted.

Chapter 6 presents some reflections on several important aspects of the EU-SPICY phenotyping experiments including the choice of parents, type and size of the population, type and size of marker data, phenotype measurement protocols etc. The chapter also summarizes and discusses the most important results from this thesis as regards prediction of complex trait yield. The results are discussed in the light of recent developments in quantitative genetics.

CHAPTER 2

Genetic and QTL analyses of yield and a set of physiological traits in pepper

CHAPTER 2

Genetic and QTL analyses of yield and a set of physiological traits in pepper

2.1. Abstract

An interesting strategy for improvement of a complex trait dissects the complex trait in a number of physiological component traits, with the latter having hopefully a simple genetic basis. The complex trait is then improved via improvement of its component traits. As first part of such a strategy to improve yield in pepper, we present genetic and QTL analyses for four pepper experiments. Sixteen traits were analyzed for a population of 149 recombinant inbred lines, obtained from a cross between the large-fruited pepper cultivar 'Yolo Wonder' (YW) and the small fruited pepper 'Criollo de Morelos 334' (CM334). The marker data consisted of 493 markers assembled into 17 linkage groups covering 1775 cM. The trait distributions were unimodal, although sometimes skewed. Many traits displayed heterosis and transgression. Heritabilities were high (mean 0.86, with a range between 0.43 and 0.96). A multiple QTL mapping approach per trait and environment yielded 24 QTLs. The average numbers of QTLs per trait was two, ranging between zero and six. The total explained trait variance by QTLs varied between 9% and 61%. QTL effects differed quantitatively between environments, but not qualitatively. For stem-related traits, the trait-increasing QTL alleles came from parent CM334, while for leaf and fruit related traits the increasing QTL alleles came from parent YW. The QTLs on linkage groups 1b, 2, 3a, 4, 6 and 12 showed pleiotropic effects with patterns that were consistent with the genetic correlations. These results contribute to a better understanding of the genetics of yield-related physiological traits in pepper and represent a first step in the improvement of the target trait yield.

Keywords

Capsicum annuum; Complex trait; Component trait; Dissection; Genetic Correlation; Pleiotropy

2.2. Introduction

Complex traits are traits determined by a relatively large number of QTLs that are environment sensitive, i.e., show QTL by environment interaction, and that are prone to show epistatic interactions. Hence, direct improvement of a complex trait by selection on that trait itself may be difficult. An attractive alternative to selection on the complex trait may be selection on underlying physiological component traits, where the most difficult task consists in finding a model for the complex trait as a function of a number of component traits. The latter should be biologically meaningful and easily measurable, and they should have a relatively simple genetic basis, i.e., one or a few additive QTLs without interactions. Recent reviews on this improvement by dissection approach were given by Hammer *et al.* (2006); Chapman (2008); and van Eeuwijk *et al.* (2010).

The FP7 European Union research project 'Smart tools for Prediction and Improvement of Crop Yield' (EU-SPICY) had as its starting point this dissection approach to complex trait improvement and aimed at the development of a suite of tools for molecular breeding of crop plants for sustainable and competitive horticulture. An introduction to the EU-SPICY project can be found at www.spicyweb.eu and in Voorrips *et al.* (2010). Within the EU-SPICY project, pepper (*Capsicum annuum*) was chosen as a model crop. In pepper, several studies have reported on genetic parameters of a series of traits and their QTL mapping.

Among pepper traits already studied are those related to disease resistance (Lefebvre and Palloix, 1996; Caranta *et al.*, 1997a; Caranta *et al.*, 1997b; Ben Chaim *et al.*, 2001a; Chaim *et al.*, 2003; Lefebvre *et al.*, 2003; Thabuis *et al.*, 2003; Kim *et al.*, 2004; Voorrips *et al.*, 2004; Sugita *et al.*, 2006; Minamiyama *et al.*, 2007; Mimura *et al.*, 2009; Kim *et al.*, 2011), and sensory traits such as pungency (Blum *et al.*, 2003; Ben Chaim *et al.*, 2006a). Other studies have also looked into fruit-related traits (Lefebvre *et al.*, 1998; Ben Chaim *et al.*, 2001b; Chaim *et al.*, 2003; Rao *et al.*, 2003; Wang *et al.*, 2004; Zygier *et al.*, 2005; Ben Chaim *et al.*, 2006b; Lee *et al.*, 2008; Barchi *et al.*, 2009). Results from these studies revealed clusters of QTLs on chromosomes 2, 3 and 4 for fruit traits such as fruit weight/yield, diameter, length and shape. Only few studies have reported on QTLs influencing vegetative-related traits such as stem length and number of internodes (Ben Chaim *et al.*, 2001b; Barchi *et al.*, 2009; Alimi *et al.*, 2010; Mimura *et al.*, 2010). In these studies, major QTLs influencing primary vegetative components like length and number of internodes on the primary stem were found on chromosomes 2, 3 and 12. Only one study reported QTLs responsible for leaf related components such as leaf area and weight (De Swart *et al.*, 2007). Also the majorities of these studies were conducted in a single environment and hence could not compare performances of a particular mapping population under different environmental conditions.

In this study, as part of the EU-SPICY project, we evaluated 16 physiological traits across four environments using a mapping population of recombinant inbred lines (RIL) obtained from the cross between large – fruited 'Yolo Wonder' (YW) and the pungent 'Criollo de Morelos 334' (CM 334) pepper cultivars (Barchi *et al.*, 2007). We started with description and phenotypic analyses of the four large phenotyping experiments (=environments) and obtained genetic parameters for all traits. For all environments, we used a multiple-QTL mapping (MQM) method (Jansen, 1993; Arends *et al.*, 2010) to estimate location, heritability

and direction of the QTLs. We qualitatively investigated QTL pleiotropy and we discuss our results in the light of previously reported QTLs for these and similar traits in pepper.

2.3. Materials and Methods

2.3.1. Plant materials

The bi-parental pepper population comprised 149 individuals from the sixth generation of the recombinant inbred lines (RIL) of an intraspecific cross between the large – fruited inbred cultivar 'Yolo Wonder' (YW) and the small-fruited cultivar 'Criollo de Morelos 334' (CM 334) of *Capsicum annuum* (Barchi *et al.*, 2007). These 149 individuals were selected from a total of 297 RILs as being the most informative subset for selective phenotyping (Vision *et al.*, 2000).

2.3.2. Phenotyping experiments and designs

The phenotyping experiments of the SPICY project were carried out at two locations, i.e., Wageningen in the Netherlands (NL) and El-Ejido in Spain (SP), representing temperate and Mediterranean growing conditions respectively. At both locations, experiments were done during two time periods: December – May (1) and June – December (2). This generated four experiments denoted as NL1, NL2, SP1 and SP2. Border rows and dummy plots were used to minimize the effects of competition between neighbouring plants and genotypes.

The Dutch (NL) experiments were performed in four Venlo-type greenhouse compartments (12m x 12m) with glass cover. A single compartment was too small to grow all genotypes, so the experiments were set up in an incomplete block design (Williams and John, 1999) with subsets of genotypes differentially replicated within and across compartments as explained below. Each compartment (block) consisted of two single border rows and six double rows of 9.6 m length at a distance of 1.50 m. On each row, 46 plants were placed on rockwool slabs. The two plants at the outsides of the double rows were also considered to be border plants. The remaining 2x44=88 plants in each double row were allocated to 11 plots, say columns, of 2x4 plants. This gave a total of 264 (6 double rows x 11 columns x 4 blocks) plots. Only the inner four plants of each plot were used for phenotypic measurements. The 152 genotypes (149 RILs + 2 parents + 1 F_1) were randomly allocated to plots in the following manner. Four non-overlapping subsets of genotypes were defined (Appendix 2A, Table 2A1). A so-called *common set* consisted of the two parents and the F_1. These three genotypes occurred once in each of the four blocks. Four so-called *ladder sets* were defined, each consisted of three RILs (= 4 x 3 = 12 genotypes). Each genotype in the ladder sets appeared in two out of the four blocks and was in addition replicated once in each block. Hence these genotypes were present on two plots inside a block and appeared in two blocks. The genotypes in the ladder sets were selected as being representative from the population of 149 RILs on the basis of an ordination based on a similarity matrix created from a set of markers (Johnson and Wichern, 2002). The genotypes in the ladder sets occupied a total of 12 genotypes x 2 blocks x 2 plots per block = 48 plots. The ladder sets connected the blocks, i.e., block 1 and 2, 2 and 3, 3 and 4, and 4 and 1, respectively. Out of the remaining 137 (149 - 12) RILs, 67 RILs were replicated once in two blocks, the so-called *double set,* giving 134 plots. The remaining 70 RILs were placed in

only one block and were thus not replicated, forming the so-called *single set,* yielding 70 plots.

In the Dutch trials, during cultivation, side shoots were pruned in order to keep a single stem per plant. Plant density was approximately 6.4 plants per m^2 (i.e. about 6 stems per m^2). Side shoots were pruned at the second internode, i.e. they carried three leaves and three flowers. Fruits were harvested when they were at least 50% red. Set points for greenhouse climate were 18/21 °C night/day with 16 °C from sunset to midnight. CO_2 was supplied up to approximately 600 ppm when windows opening were less than five percent. At higher ventilation rates, CO_2 was supplied to a level of 400 ppm. Throughout the experiment, climate data were registered every five minutes: greenhouse air temperature, pipe temperature, humidity, CO_2 concentration, outside temperature, global radiation outside, screen opening and window opening.

The Spanish (SP) experiments were performed in a greenhouse of 40 m x 60 m with plastic cover. The multi-span tunnel was oriented East-West; row orientation was North-South. The greenhouse was divided into two blocks. Each block consisted of 8 experimental rows and 24 experimental columns giving a total of 192 plots per block. The 152 genotypes were randomly allocated to any of the 192 plots in a block, thereby leaving 40 plots. These 40 plots were filled with dummy genotypes (Appendix 2A, Table 2A2). From each plot of five plants, three plants were sampled for phenotypic measurements. During cultivation, two stems per plant were kept. Plant density was approximately three plants per m^2 (i.e. 6 stems per m^2). In the first experiment, side shoots were pruned at the second internode and the two topmost flowers were removed leaving three leaves and one flower. In the second experiment, side shoots were pruned at the second internode as well. However, since no flowers were removed from the side shoots, they bore three leaves and three flowers. Fruits were harvested when they were at least 50% red. In the first experiment, the heating system was used when the outside temperature was lower than 14 ºC, whereas in the second experiment the heating system was not used. No CO_2 was supplied. Climate data were registered every five minutes: greenhouse air temperature, humidity, CO_2 concentration, outside temperature and inside radiation.

Plant measurements: During the experiment, plant development was recorded via counting or measuring the number of fruits per stem (weekly), number of internodes per stem (fortnightly), stem length (monthly), number and fresh weights of harvested fruits (when fruits were 50% red) and dry weights of harvested fruits (periodically). Only fruits larger than 1 cm were counted as fruits. For the fruits for which only fresh weight was measured, fruit dry weight was estimated by calculating the fraction fruit dry matter per genotype, and multiplying this fraction with the measured fruit fresh weight. At the end of the experiment three plants per plot were harvested destructively to record leaf area and dry weights of leaves, stems and fruits, as well as the number of fruits.

2.3.3. Trait evaluation

In this study, we analysed 9 physiological traits representing vegetative and generative development of pepper plants and 7 derived traits, being functions of the original traits (Table 2.1). The vegetative traits were measured via destructive harvesting at the end of the trials. The vegetative traits recorded include leaf area (LA), dry weights of leaves (DWL), stem

(DWS) and vegetative plant parts (DWV=DWL+DWS), specific leaf area (SLA=LA/DWL), the primary axis length (Axl) given as length of primary axis from cotyledons to first branching, number of leaves on primary axis (NLE), mean internode length of primary axis (INL=Axl/NLE). Other vegetative traits included stem length (SL) and number of internodes (NI) measured 6-8 weeks after transplanting. Recorded fruit traits included total number of fruit (NF) and total fruit dry weight (DWF) given as the sum of dry weight of all the fruits harvested during the growing season and the fruits on the plant at the final destructive harvest. DWF was taken to represent yield. Also calculated were the total plant biomass (DWP=DWV+DWF) and the proportion of total biomass in the leaves (pt_leaf), stem (pt_stem) and fruits (pt_frt). For each environment, trait distributions, correlations and pattern of variation across genotypes and blocks were obtained by using statistical and visualization tools after removing outliers.

Table 3.1 Traits measured in each of the four SPICY environments (experiments).

Abbreviation	Trait
LA	Leaf Area (cm^2)
DWL	Dry weight of leaf (g)
DWS	Dry weight of stem (g)
DWV	Dry weight of vegetative part (g) = DWL + DWS
SLA	Specific Leaf Area = LA/DWL/10
NF	Total number of fruits
DWF[1]	Total fruit dry weights from each plant (g)
DWP	Dry weight of plant at end harvest (g) = DWV + DWF
pt_frt	Proportion of the total biomass due to fruit= DWF/DWP
pt_leaf	Proportion of the total biomass due to leaf = DWL/DWP
pt_stem	Proportion of the total biomass due to stem = DWS/DWP
Axl	Primary Axis length (Stem length before first branching) (cm)
SL	Stem length measured 6-8 weeks after transplanting
NLE	Number of Leaves on the primary axis
NI	Number of Internodes at time 3-4 weeks after transplanting
INL	Internode length for the primary axis (Axl/NLE)

[1]representative for yield

2.3.4. Phenotypic analysis

The traits in each environment were univariately analyzed using linear mixed models. We adopted a model specification as proposed in Piepho *et al.* (2006) to analyse the data on all the RILs including the parents and F_1. The linear mixed model used was of the form:

$$Y = \mu + B + R(B) + C(B) + M + (Z^*G) + \varepsilon, \qquad (2.1)$$

where Y represented a phenotypic trait value, μ was the overall mean, B, $R(B)$ and $C(B)$ represented block, row-within-block and column-within-block effects respectively. M was a 4-level factor used to obtain and test phenotypic mean differences among the parents (YW, CM 334), F_1 and RILs. G stood for all the 152 genotypes. We introduced a variable (Z), coded 0 for parents and F_1; and 1 for RILs. This allowed us to handle parents and F_1 as fixed and the 149 RILs as random. Now the random RIL effect was modeled as Z*G. This induced a genetic variance of zero for the parents and a common genetic variance for the RILs. Z was

declared as quantitative in our SAS model statements. Defining Z as quantitative ensured that no genetic effects were induced for the parents (Piepho *et al.*, 2006). All other non-genetic effects were captured in ε term.

In the NL environments, *block* was assumed random for recovery of inter-block information since not all RILs were present in all the blocks (unbalanced). The variance components for the random effects were estimated using restricted maximum likelihood (REML) (Littell *et al.*, 2006). Since RILs were assumed random, they were estimated using best linear unbiased prediction (BLUP) as against the use of best linear unbiased estimation (BLUE) where RILs are treated as fixed. We however investigated the use of both for our data and found that both BLUE and BLUP results were very similar with correlation of about one, despite the shrinkage factor in BLUP. These analyses were performed in SAS (Saxton, 2004; Littell *et al.*, 2006). Trait heritabilities were calculated using the measure based on BLUP as proposed by Cullis *et al.* (2006).

$$H^2 = 1 - \frac{V_{BLUP}}{2\sigma_g^2}, \tag{2.2}$$

where σ_g^2 was the genotypic variance and v_{BLUP} was the mean BLUP variance. The genetic correlations (ρ_g) between traits in each environment were estimated from the estimates of variances and covariances obtained from a multivariate REML under the mixed model procedure in SAS (Littell *et al.*, 2006). The use of multivariate REML was preferred over classical multivariate analysis of variance (MANOVA) as it can handle unbalanced data.

The dominance coefficient (k) for all traits was calculated from the expression: $a(1+k) = d$, where a was the overall additive effect in the CM334 parent $\left(a = \frac{CM334-YW}{2}\right)$ and d was the mean difference between the F_1 and YW (Lynch and Walsh, 1998). When -1<k<0, the YW contribution was dominant over that of CM334 and when 0<k<1, CM334 was dominant. If k < -1 or k > 1 and the phenotypic mean of the F_1 exceeded that of the parent considered to represent the desirable parent, we talk of heterosis. Transgressive segregation means that the phenotypic values of some of the RIL offspring were outside the range of parental phenotypic means. Transgressive segregation was declared substantial when the proportion of RILs with phenotypes lower than the lower parent (denoted Q_{min}) - or higher than the higher parent (denoted Q_{max}), was 50 percent or higher. We also compared phenotypic means of F_1 and RILs using expression D_{RIL}. The statistic D_{RIL} expressed difference in the means of RIL and F_1 (D_{RIL} = RIL-F_1) where RILs are phenotypically superior to F_1 if $D_{RIL} > 0$.

2.3.5. Marker data and Linkage map

The first genetic linkage map (Figure 2.1) used the map similar to that of Barchi et al. (2007). A final set of 493 markers were assembled into 17 linkage groups (LG) covering 1775 cM. These were assigned to the 12 pepper chromosomes based on known positions of SSR markers. The list of publicly-owned markers used for the map construction is available as a supplementary material (Supplemental A1). Five chromosomes had two linkage groups that could not be joined due to insufficient linkage. The percentage of missing genotype information across the full set of markers was low (6.8%). The quality controls conducted

included checks on segregation distortion, recombination fraction and number of crossover events.

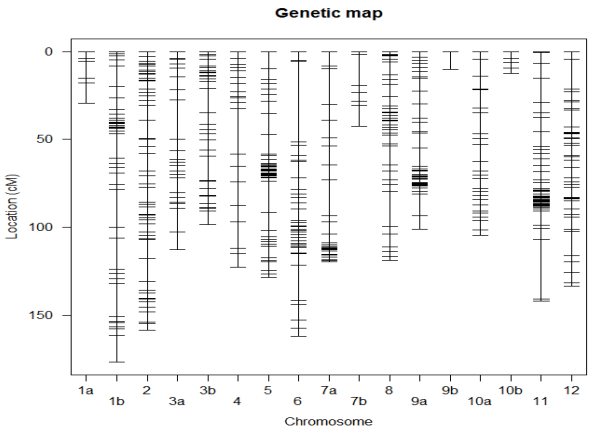


Figure 3.1 Initial Genetic linkage map used in the EU-SPICY experiments

2.3.6. QTL estimation

Many of the traits under investigation were assumed to be controlled by multiple QTLs which made a single QTL model inappropriate. We used a multiple-QTL mapping procedure (MQM) (Jansen, 1993; Arends *et al.*, 2010) for each trait in each environment.

$$Y = \mu + \sum_{q=1}^{Q} x_q \alpha_q + e, \qquad (2.3)$$

where Y was the phenotypic response, μ the population mean, α_q was the additive effect of QTL q, x_q was a marker-genotype indicator variables (0-1) and e was the residual term. The package qtl (Broman and Sen, 2009; Arends *et al.*, 2010) of the software R (R-Development-Core-Team, 2011) was used to deploy the MQM approach in five steps. Firstly, the missing marker genotypes were imputed with their probabilities conditional on neighbouring marker information. Secondly, an initial single QTL scan equivalent to simple interval mapping was performed and a global significance threshold for QTL selection across all traits was determined via a permutation test of 1000 replicates. The obtained significance threshold was equal to a LOD score of 2.9. Thirdly, the MQM model was fitted by forward selection. Fourthly, a backward elimination strategy was applied to the full model with all earlier selected QTLs included to remove the non-significant QTLs and to arrive at the final QTL model. The QTL location confidence intervals were estimated from the final QTL using a Bayes credible interval with the assumption that there was one and only one QTL on the LG of interest for a given trait (Broman and Sen, 2009). This was mostly true for our data except for one trait (INL) on LG 1b in NL1. Lastly, the resulting final QTL model was evaluated to obtain size and direction of QTL effects. Also, the QTL heritability H_Q^2 (proportion of phenotypic variance due to a QTL) was estimated directly from the difference of the log-likelihood (LOD) scores using the relationship: $H_Q^2 = 1 - 10^{-\frac{2}{n}LOD}$ (Broman and Sen, 2009). Pleiotropic QTLs were evaluated via visual inspection of the estimated QTL positions for different traits: QTLs with overlapping confidence intervals were declared to be the same QTL, i.e., a QTL with pleiotropic effects.

2.4. Results

2.4.1. Traits evaluation

The phenotypic means for the parents, F_1 and RILs for the SP2 environment (as an appropriate representative) are presented in Table 2.2 and the means pertaining to the other environments are given in Appendix 2A. The parental means were clearly different for most of the traits. CM334 showed longer stem and internode lengths (Axl, SL and INL), heavier stem and total vegetative dry weight (DWS & DWV) and higher number of internodes (NLE and NI). In contrast, YW had higher values for fruit and leaf related traits. This parent showed bigger and heavier fruits, higher partitioning into fruit (pt_frt), higher leaf area (LA) and leaf dry weight (DWL). These results were consistent with previously reported results for these pepper cultivars (Barchi *et al.*, 2009). These contrasts between parental values were consistent across the environments. For many traits, the F_1 had higher mean values than the averaged parental means especially for vegetative traits (e.g. DWL, DWS, DWV, DWP, Axl, SL and INL). The RILs showed substantial transgressive segregation for some traits (e.g. DWS, DWV, NF and DWP), where the transgression was in the direction of the parent with higher phenotypic values.

Table 3.2 Phenotypic Mean comparison for environment SP2

Traits	YW	CM334	F_1	RIL	k	D_{RIL}	Q_{min}	Q_{max}
LA	8198.46	5985.77	10372.52	9980.89	-2.97$^+$	-391.63	0.04	0.73
DWL	36.25	32.95	46.23	52.26	-7.05$^+$	6.03	0.04	0.91
DWS	29.02	87.28	88.44	95.55	1.04	7.11	0.00	0.62
DWV	72.27	120.22	134.68	147.62	1.60$^+$	12.94	0.01	0.80
SLA	22.85	18.34	22.60	19.25	-0.89	-3.35$^+$	0.34	0.05
NF	10.83	23.00	40.17	37.41	3.82$^+$	-2.76$^+$	0.01	0.90
DWF	104.62	12.51	89.27	87.13	-0.67$^+$	-2.14	0.01	0.31
DWP	176.90	132.73	223.94	234.94	-3.13$^+$	11.00	0.01	0.90
pt_frt	0.58	0.10	0.40	0.36	-0.25	-0.04	0.01	0.00
pt_leaf	0.25	0.25	0.21	0.23		0.02	0.80	0.20
pt_stem	0.16	0.66	0.39	0.41	-0.08	0.02	0.00	0.01
Axl	20.83	28.83	34.00	25.28	2.29$^+$	-8.72$^+$	0.17	0.22
SL	23.83	76.33	83.33	67.38	1.27$^+$	-15.95$^+$	0.00	0.27
NLE	9.33	13.50	11.33	9.87	-0.04	-1.46$^+$	0.44	0.01
NI	8.50	11.67	10.83	10.12	0.47	-0.71	0.09	0.08
INL	2.23	2.15	3.01	2.60	-20.5$^+$	-0.41$^+$	0.14	0.79

k = Dominance coefficient
D_{RIL} = Difference in the means of RIL and F_1
Q_{min} = Proportion of RILs with phenotypes lower than the lower parent
Q_{max} = proportion of RILs with phenotypes higher than the higher parent
$^+$Significant at 0.05 level of significance

In the SP2 environment (Table 2.2) the YW contribution was dominant over CM334 ($-1<k<0$) for SLA, DWF, pt_frt, pt_stem and NLE while the CM334 contribution was dominant for NI ($0<k<1$). Heterosis in the direction of YW ($k < -1$) was observed for LA, DWL, DWP and INL while heterosis in the direction of CM334 ($k > 1$) was observed for DWS, DWV, NF, Axl and SL. The result for NF in SP2 is different from other environments in that the dominance for NF was derived from YW in all the other environments except SP2. Many

traits that showed heterosis also displayed substantial transgressive segregation and often in the same direction (Table 2.3). In all the environments, SLA and NLE for example showed consistently higher value of Q_{min} than Q_{max}. Pt_leaf also showed higher Q_{min} than Q_{max} in NL2 and SP2.While LA, DWL, DWS, DWV, DWP and NF showed higher Q_{max} in all environments. For some of the traits (e.g. NF, DWF, Axl and SL) RILs were phenotypically inferior to F_1 as F_1 displayed higher phenotypic values consistently in the four environments. For other traits, the sign of D_{RIL} varied across environments. As examples, for LA a positive D_{RIL} was only found in SP1 and D_{RIL} was positive in the SP environments for DWL, DWS, DWV and negative in NL environments.

Table 3.3 Traits showing heterosis and substantial transgressive segregation

	Heterosis					Transgressive segregation			
	NL1	NL2	SP1	SP2		NL1	NL2	SP1	SP2
LA				Y-					T2
DWL	Y-	Y-	Y-	Y-				T2	T2
DWS	Y+	Y+		Y+		T2	T2	T2	T2
DWV	Y+	Y+	Y+	Y+		T2	T2	T2	T2
SLA	Y+					T1			
NF	Y-	Y-	Y-	Y+		T2	T2	T2	T2
DWF									
DWP	Y-	Y+	Y-	Y-		T2	T2	T2	T2
pt_frt									
pt_leaf							T1	T2	T1
pt_stem							T2		
Axl		Y+		Y+					
SL	Y+	Y+	Y+	Y+					
NLE		Y-					T1		
NI	Y+	Y+	Y+			T2			
INL	Y+	Y+	Y+	Y-			T2	T2	T2

Y- = presence of heterosis in the direction of YW; Y+ = presence of heterosis in the direction of CM334; T1 = presence of substantial transgression in the direction of parent with lower phenotypic mean (i.e. high Q_{min}) and T2 = presence of substantial transgression in the direction of parent with higher phenotypic mean (i.e. high Q_{max}).

2.4.2. Heritability & genetic correlation

The heritability estimates (H_T^2) of traits were consistently high across the environments with average of 0.86 and varied from 0.43 – 0.96 (Table 2.4). H_T^2 were mostly higher in the SP environments than in the NL environments. Genetic correlations as a measure of association between traits within each environment were calculated. The correlations were found to show similar patterns across the environments (Figure 2.2). Fruit-related traits (DWF, NF and pt_frt) were positively correlated. Vegetative-related traits (LA, DWL, DWS, and DWV) were also positively correlated. However, SLA showed no significant correlation to any other trait except LA to which it was moderately correlated.

Generally, fruit-related traits (NF, DWF, pt_frt) showed negative correlations to stem traits (SL, NI, NLE, Axl, pt_stem) but weaker positive correlations with leaf traits (LA, DWL, SLA, pt_leaf), reflecting level of competition between development of organs. This indicates that fruit competes more with stem than with leaves for nutrients. Total plant biomass (DWP),

being an aggregate of fruit, stem and leaf components, showed positive correlations to many fruit, stem and leaf traits (e.g. DWF, NF, DWS, LA, DWL, DWV).

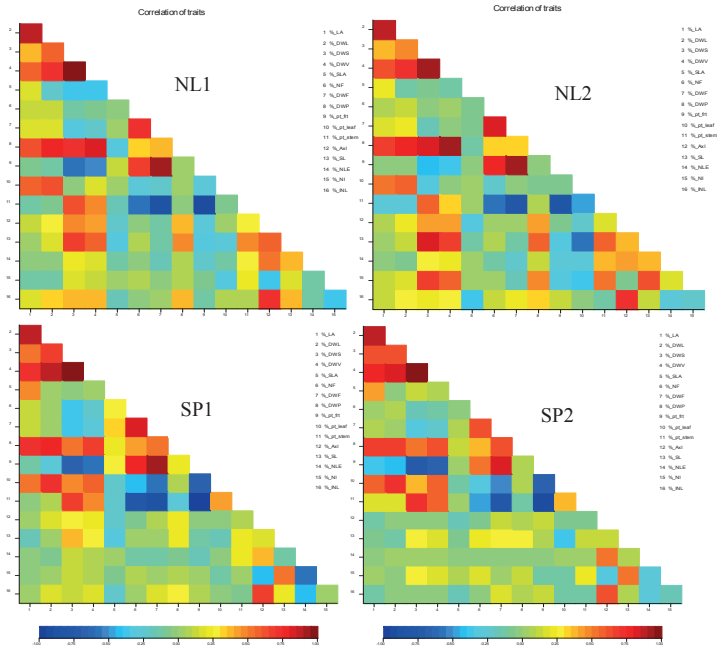


Figure 3.2 Genetic correlations for traits in each of the environments

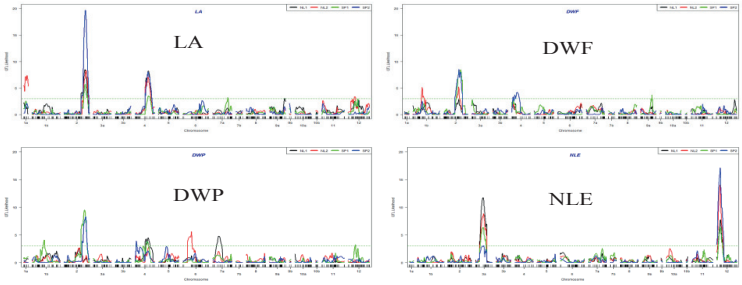


Figure 3.3 QTL likelihood profiles for selected traits
Colour codes: Black = NL1, Red=NL2, Green=SP1 and Blue=SP2

Table 3.4 Number of QTLs (#QTL), QTL explained variation (HQ2) and broad-sense heritability (HT2) in each of the four environments.

	NL1			NL2			SP1			SP2		
Trait	#QTL	H_Q^2	H_T^2	#QTL	H_Q^2	H_T^2	#QTL	H_Q^2	H_T^2	#QTL	H_Q^2	H_T^2
LA	2	0.41	0.71	3	0.40	0.90	2	0.25	0.91	2	0.56	0.94
DWL	3	0.46	0.88	3	0.36	0.86	3	0.31	0.91	2	0.45	0.90
DWS	*	*	0.85	1	0.23	0.85	1	0.18	0.89	1	0.13	0.89
DWV	*	*	0.71	1	0.20	0.90	1	0.12	0.87	1	0.19	0.88
SLA	1	0.12	0.43	*	*	0.66	2	0.28	0.68	2	0.31	0.90
NF	*	*	0.77	1	0.11	0.89	3	0.34	0.82	1	0.16	0.88
DWF	1	0.09	0.82	2	0.27	0.76	2	0.31	0.89	2	0.33	0.87
DWP	1	0.13	0.57	*	*	0.86	3	0.42	0.85	2	0.31	0.81
pt_frt	*	*	0.89	2	0.22	0.89	1	0.15	0.95	2	0.27	0.93
pt_leaf	2	0.26	0.79	3	0.34	0.80	2	0.17	0.93	3	0.28	0.91
pt_stem	*	*	0.91	3	0.30	0.89	1	0.14	0.95	2	0.27	0.94
Axl	2	0.37	0.96	3	0.34	0.93	*	*	0.94	3	0.29	0.86
SL	2	0.23	0.95	3	0.43	0.92	3	0.48	0.94	5	0.61	0.93
NLE	2	0.45	0.92	2	0.51	0.93	2	0.33	0.91	2	0.46	0.79
NI	3	0.39	0.86	3	0.52	0.94	1	0.16	0.90	4	0.48	0.88
INL	6	0.61	0.94	2	0.18	0.93	5	0.48	0.89	*	*	0.75

*No QTL detected

2.4.3. QTL results

Result from our QTL analyses revealed a total of 24 unique regions with QTLs for the 16 traits across the four environments. All QTLs with overlapping confidence intervals were clustered into unique QTL regions with pleiotropic effects. Figure 2.3 shows the likelihood profiles of the QTL models for the traits LA, DWF, DWP and NLE, respectively. For many of the traits, the number of QTLs (#QTL) and their heritabilities (H_Q^2) differed across environments (Table 2.4). The proportions of phenotypic variance explained by each of the significant QTLs ranged between 0.09 and 0.45. For many of the traits, the detected QTLs together explained a considerable amount of the phenotypic variability (≥ 0.25) but always much less than the heritability estimates in the phenotypic analyses (H_T^2) (Table 2.4). The proportion of variance explained by QTLs for fruit related traits were not as high as those of vegetative related traits.

The list of QTLs with their magnitudes across the environments is tabulated in Table 2.5. QTLs with larger signals were usually identified consistently significant across the four environments; however, the magnitude of their LOD scores differed among environments. For example, for LA, two major QTLs were detected on LG 2 and 4. These QTLs were significant in the four environments but differed in magnitude. Similarly, for NLE, two large QTLs were detected on LG 3a and 12 in all environments. For some traits, one or more of their QTLs were only picked up in certain environments. For example, DWF had one QTL on LG 2 that was significantly expressed in all environments, but all other QTLs for DWF were significant in only one environment (e.g. on LG 1b in NL2, 4 in SP2 and 9a in SP1). For total biomass

(DWP), a highly significant QTL was found on LG 2 in the SP environments but not in the NL environments.

The directions of most QTL effects were similar across environments (Figure 2.4 and Table 2.5). QTL effect directions generally followed direction of parental mean differences for each of the traits. For example, YW showed a higher mean value for DWF and at all QTLs influencing DWF, the increasing alleles were inherited from YW. In SP2, where the parental difference for DWF was about 92g, the substitution of the YW alleles for CM334 alleles at the two significant QTLs on LG 2 and 4 would increase fruit yield per plant by about 63g. With parental differences of 48g, 17g and 142g in NL1, NL2 and SP1 respectively, the same allelic substitution would increase yield per plant by about 17g (1 QTL), 16g (2 QTLs) and 98g (2 QTLs) respectively. Conversely, the three QTLs for pt_stem would increase partitioning to stem by 26%, 28%, 28% and 20% in NL1, NL2, SP1 and SP2 respectively if the two CM334 alleles were substituted for YW alleles. This was also in agreement with the parental mean of CM334 showing a higher mean value. Most of the alleles increasing stem related traits such as DWS, pt_stem, SL, NI and INL were derived from CM334 while those increasing leaf and fruit traits such as LA, DWL, SLA, DWF and pt_frt were derived from YW. For some traits such as NLE, NF, pt_leaf and Axl, alleles increasing the traits originated from both parents.

The pleiotropic effect of QTLs (Figure 2.4) revealed the presence of some QTLs influencing many traits simultaneously. As can be expected, many of these pleiotropic QTLs were very consistent with the genetic correlations between traits. Two QTLs affecting many traits were found some distance apart on LG 2 (around 98 cM and 131 cM). One of these two QTLs was specific to fruit related traits (DWF, pt_frt, NF) and also pt_stem. The other QTL on LG 2 affected mainly vegetative traits such as LA, DWL, DWS, DWV, SLA and DWP. On LG 3a (around 50 cM), a QTL that affected the number of internodes (NLE and NI) and stem length (SL and Axl) was detected. On LG 4 was a QTL affecting LA and DWL. A QTL affecting primarily SL, NI, DWS and DWV was located on LG 6. A QTL specific to plant development before primary branching (Axl and NLE) and NI was found on LG 12. Some QTLs with environment specific pleiotropic effect were also detected. For example, in NL1, a QTL influencing INL, Axl and DWL was found on LG 7a. On LG 1b, a QTL governing DWF, pt_frt, pt_stem and NI was picked up only in NL2 and SP2. In SP1, a QTL affecting SL, NI, INL, DWV and DWS was found on LG 11.

Table 3.5 QTL scores and effect sizes

Trait	Markers[$]	LOD Score[$$]				Effect magnitude and direction			
		NL1	NL2	SP1	SP2	NL1	NL2	SP1	SP2
LA	PMD					-4153.8^{+}	-4309.3^{+}	-3886.4^{+}	-2212.7
	1a@15.1	1.75	4.74	2.64	1.91	-277.7	-339.8*	-608.9	-89.3
	2@130.9	6.46	3.39	5.54	13.35	-443.2*	-504.1*	-860.6*	-1396.5*
	4@87.7	7.01	3.41	3.35	4.48	-528.2*	-628.8*	-656.6*	-1075.5*
DWL	PMD					-14.21^{+}	-14.07^{+}	-16.99^{+}	-3.3
	2@130.9	3.13	2.55	2.87	7.57	-2.12*	-1.68	-2.32	-5.08*
	4@87.7	8.20	7.70	4.33	6.63	-3.45*	-1.76*	-5.25*	-5.61*
DWS	PMD					46.43^{+}	33.41^{+}	74.65^{+}	58.26^{+}
	2@130.9	0.14	0.12	0.29	3.60	-2.28	-1.26	-4.09	-11.43*
	6@53.3	2.52	7.65	2.15	1.98	10.12	10.82*	11.27	8.35
	11@88.1	0.46	0.16	5.09	0.09	4.18	-1.40	17.74*	1.74
DWV	PMD					31.68^{+}	19.81^{+}	71.66^{+}	47.95^{+}
	2@130.9	0.54	0.64	0.98	5.54	-5.72	-3.85	-11.17	-20.33*
	6@53.3	2.04	6.45	1.49	1.71	11.17	12.99*	14.00	10.62
	11@88.1	0.28	0.13	3.76	0.19	4.00	-1.73	22.75*	3.52
SLA	PMD					-5.27^{+}	-6.4^{+}	-4.19^{+}	-4.51^{+}
	2@12.7	4.23	2.23	2.64	5.22	-1.15*	-0.85	-0.76	-1.44*
	2@135.9	0.57	0.34	2.35	4.91	-0.43	-0.32	-0.71	-1.39*
	6@121.6	1.26	2.97	1.74	5.24	-0.60	-0.99*	-0.59	-1.43*
NF	PMD					-3.04	-2.5^{+}	-8.5	12.17^{+}
	1b@60.6	0.51	3.11	0.81	1.85	-1.41	-2.40*	-2.80	-4.51
	2@98.3	0.90	2.69	6.68	1.66	-1.90	-2.23	-8.40*	-4.22
	3a@89.3	0.73	1.02	0.44	4.85	1.71	1.35	2.06	7.49*
	8@103.7	0.56	1.78	3.87	1.77	1.51	1.77	6.30*	4.46
DWF	PMD					-48.29^{+}	-17.45^{+}	-141.75^{+}	-92.11^{+}
	1b@60.6	1.57	4.49	1.22	2.57	-6.13	-3.76*	-10.43	-10.09
	2@98.3	3.15	5.29	8.14	7.33	-8.86*	-4.06*	-28.91*	-18.43*
	4@32.6	0.93	1.03	1.24	4.06	-4.72	-1.75	-10.53	-13.18*
	9a@74.4	1.73	0.94	3.76	0.04	-6.41	-1.63	-18.84*	-1.23
DWP	PMD					-17.61	2.36	-84.09^{+}	-44.17
	1b@75.7	1.16	0.62	3.03	0.52	-8.40	-4.18	-20.92*	-5.26
	2@130.9	1.32	0.93	7.31	6.27	-8.77	-4.97	-34.44*	-19.01*
	4@8.9	0.82	2.24	0.66	3.86	-6.82	-7.95	-8.99	-14.82*
pt_frt	PMD					-0.47^{+}	-0.28^{+}	-0.61^{+}	-0.48^{+}
	1b@60.6	1.03	3.72	0.73	3.08	-0.04	-0.04*	-0.04	-0.05*
	2@98.3	1.67	3.59	4.48	5.54	-0.05	-0.04*	-0.09*	-0.06*
pt_leaf	PMD					-0.12^{+}	-0.23^{+}	0	0
	1a@15.1	3.55	3.40	1.59	2.38	-0.02*	-0.02*	-0.02	-0.01
	2@75.2	0.33	0.44	1.99	3.24	0.01	0.01	0.02	0.02*
	4@58.5	3.12	3.82	0.61	2.39	-0.02*	-0.02*	-0.01	-0.01
	10a@34.7	4.93	4.28	1.42	3.77	0.02*	0.02*	0.02	0.02*
pt_stem	PMD					0.59^{+}	0.51^{+}	0.61^{+}	0.5^{+}
	1b@60.6	1.80	3.93	1.15	3.95	0.05	0.05*	0.04	0.04*
	2@98.3	2.13	3.91	4.61	5.74	0.05	0.05*	0.07*	0.05*
Axl	PMD					17^{+}	8.5^{+}	11.17^{+}	8^{+}
	1b@69.3	2.25	1.48	0.87	3.89	-2.69	-2.26	-1.42	-2.56*
	3a@50.0	3.40	0.57	0.37	1.45	3.33*	1.41	0.91	1.51
	7a@39.0	9.17	3.00	1.89	0.39	-5.64*	-3.51*	-2.12	-0.79
	9b@0.0	1.63	4.30	1.33	2.97	2.24	3.92*	1.74	2.16*
	12@23.1	2.27	3.06	1.46	2.27	-2.69	-3.36*	-1.82	-1.93
SL	PMD					49.33^{+}	90.62^{+}	29.15^{+}	52.5^{+}
	3a@50.0	3.24	4.52	0.90	6.95	6.55*	12.71*	2.43	10.70
	6@53.3	3.75	5.84	5.90	5.29	7.16*	15.03*	6.83*	9.16*

	8@114.0		0.45	0.35	3.76	1.09		2.37	3.33	5.17*	3.91
	10a@0.0		3.38	5.08	3.08	4.14		-6.70*	-13.65*	-4.70*	-8.24*
	11@78.2		0.01	0.77	5.02	4.46		-0.38	-5.15	6.02*	-8.43*
NLE	PMD							4.19^{+}	0.81	5^{+}	4.17^{+}
	3a@50.0		8.84	7.26	6.23	2.92		1.71*	1.04*	1.34*	0.58*
	12@23.1		6.25	11.39	5.77	13.94		-1.44*	-1.35*	-1.29*	-1.40*
NI	PMD							0.87^{+}	8.25^{+}	2.34^{+}	3.17^{+}
	1b@69.3		1.82	3.60	0.10	4.81		0.29	0.94*	0.10	0.66*
	2@49.8		3.86	2.74	0.14	1.34		-0.41*	-0.82	-0.12	-0.31
	3a@50.0		0.79	2.64	0.04	7.54		0.16	0.82	0.06	0.81*
	6@53.3		7.44	14.32	2.64	4.17		0.59	2.01*	0.54	0.60*
	11@84.1		0.48	0.08	7.27	0.03		0.14	0.12	0.92*	0.04
	12@23.1		1.26	0.00	3.92	4.36		0.22	0.04	0.65*	0.62*
INL	PMD							0.57^{+}	0.45^{+}	0.08	-0.08
	1b@0.0		4.22	1.75	2.14	0.29		0.21*	0.20	0.15	0.06
	1b@43.7		4.12	1.79	3.53	2.49		-0.20*	-0.21	-0.20*	-0.18
	2@2.8		3.62	1.97	2.05	0.75		0.19*	0.21	0.15	0.10
	3a@68.2		0.53	0.68	3.37	0.02		-0.06	-0.13	-0.18*	-0.01
	4@8.9		4.02	0.07	2.86	0.12		0.20*	0.04	0.17*	0.04
	7a@39.0		12.86	1.95	0.74	0.07		-0.39*	-0.21	-0.09	0.03
	10a@21.8		2.07	3.68	3.97	1.90		0.14	0.31*	0.21*	0.16
	11@88.1		0.18	0.32	5.20	0.75		0.04	0.08	0.24*	-0.10

PMD = mean difference for the parental lines (CM334 - YW)
Marker notation e.g. 1a@15.1 represents a QTL found around position 15.1cM on LG 1a
[$]The list of publicly-owned markers used for the map construction is available as a supplementary material (see: Supplemental A1).
[+] Significant PMD (at 0.05 level of significance)
*Effects corresponding to significant QTL.
[$$] Percentage of explained variation for each QTL can be estimated directly from the difference of log-likelihood (LOD scores). The relationship is given in material and method section.

Table 3.6 Comparison of mapped QTL for related traits among studies

Study	This study	Alimi *et al.* (2010)	Barchi *et al.* (2009)	Mimura *et al.* (2010)	Zygier *et al.* (2005)	Rao *et al.* (2003)	Ben Chaim *et al.* (2001b)
Parents	YW x CM334	YW x CM334	YW x CM334	CW x LS2341	Chinense x Frutescens	Maor x Frutescens	Maor x Perennial
Type	RIL F_6	RIL F_6	RIL F_6	DH F_1		BC2	RIL F3
Size	149	149	297	94		248	180
NF	1b, 2, 3a, 6, 8	*	*	*	*	2, 3, 11	*
DWF	1b, 2, 4, 9a	*	3, 4, 11, 12, LG15, LG24, LG45	*	2, 4	1, 2, 3, 4, 8, 10, 11	2, 3, 4, 8
Axl	1b, 2, 3a, 7a, 9b, 12	1, 2, 3, 12	2, 6, 9, LG24, LG47	3, 12, LG8	*	*	2, 3, 4, 6, 8
SL	3a, 4, 6, 8, 10a, 11	*	*	*	*	*	2, 3, 4, 6, 8
NLE	3a, 12	3, 12, LG45	3, LG38, LG45, LG47	12, LG8	*	*	*
INL	1b, 2, 3a, 4, 7a, 10a, 11	1, 2, 3, 6, LG22, LG28	1,2, LG28	3, LG8	*	*	*

*=Traits not reported. Other traits in our study not reported in any of the earlier studies are omitted.
LG15, LG22, LG24, LG38 and LG45 from Barchi *et al.* (2009) are part of LG 6, 1b, 6, 11 and 3b respectively in the present map, but the markers from LG28 and LG47 were not integrated.

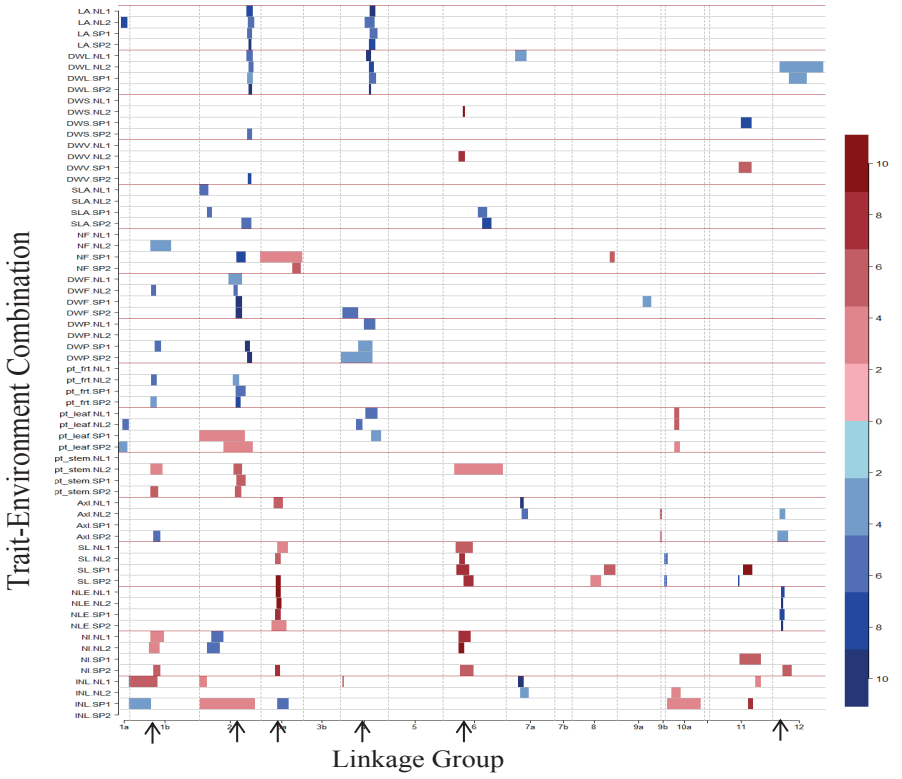


Figure 3.4 QTL LOD size, allelic direction and Pleiotropy.
Blue indicates QTL of which YW allele increases the trait value while Red indicates that CM334 allele increase the trait value. Colour intensity gives QTL LOD score magnitude while width of each line represents QTL location intervals. Trait-environment combinations are on the y-axis and linkage groups (LG) are on the x-axis. The arrows point to QTLs with major pleiotropic effects on LG1b@60cM, LG2@98cM, LG2@131cM, LG3a@50cM, LG4@88cM, LG6@55cM, LG7a@39cM, LG11@88cM and LG12@23cM

2.5. Discussion

This study is the first in a series on breeding for complex traits by dissection of these complex traits into simpler component traits. In this study, we investigated performances of a RIL population in pepper for a set of 16 traits across four environments to detect QTLs. We used the same mapping population derived from an intraspecific cross between Yolo Wonder (YW) and Criollo de Morelos 334 (CM 334) that was used by Barchi *et al.* (2009). We removed markers and individuals with more than 6% missing data and we added markers to improve the quality of the map earlier reported (Barchi *et al.*, 2007). This improved the map by resolving several smaller and unassigned linkage groups (LG) resulting into 17 LG assigned to the 12 pepper chromosomes and covering 1775 cM compared to 23 LG covering 1553cM (70% of mapped genome) in the former map. Future efforts are directed to adding more markers to improve the map towards the 12 chromosomes and fill some larger gaps on for example LG 6 and LG 11.

Trait heritabilities were calculated following Cullis *et al.* (2006). We prefer this measure because our study design involved unbalanced replication of genotypes in NL. This measure has been found to perform optimally for balanced and unbalanced data relative to the usual measure of heritability (Piepho and Moehring, 2007). In all four environments, traits were highly heritable (mean 0.86 with range from 0.43 – 0.96); hence good possibilities for mapping QTLs governing these traits. Our estimates of heritability are generally comparable with earlier results from Barchi *et al*. (2009) and Mimura *et al*. (2010) for related traits such as Axl, NLE and INL. We estimated higher heritabilities for NF and DWF as compared to result of Rao *et al*. (2003). Heritabilities were found to be slightly higher in SP environments than NL environments.

Some of the traits consistently displayed heterosis (e.g. DWL, DWV, NF, DWP, SL and INL) in the four environments ($k < -1$ or $k > 1$). Only four traits (DWF, pt_frt, pt_leaf and pt_stem) did not show heterosis in any of the environments. These four traits however displayed dominance ($-1<k<0$ or $0<k<1$) of one parent over the other. Substantial transgressive segregation ($Q_{min} \geq 0.5$ and/or $Q_{max} \geq 0.5$) was observed in 6, 8, 7 and 8 traits in NL1, NL2, SP1 and SP2 respectively. For any trait, suitability of the RILs for selection would imply that phenotypic values for many of the RILs are beyond that of the desirable/superior parent, where desirability is trait specific and may not necessarily imply higher phenotypic value. In other words, high Q_{min} is preferred for some traits (e.g. SL, DWS, pt_stem) while high Q_{max} is preferred for others (e.g. LA, DWL, DWF, pt_frt). Four of those traits (DWS, DWV, NF and DWP) consistently showed transgressive segregation in the direction of the parent with higher mean value in all environments studied. Among possible explanation for transgressive segregation is complementarity of QTL alleles (parental lines being fixed for sets of alleles having opposite effects). In our study, transgressive segregation in some of the traits could be explained by presence of complementary QTL alleles. For example, of the three detected QTLs conferring increase in dry weight of stem (DWS), two came from CM334, while one came from YW. Most of the alleles increasing stem related traits such as DWS, pt_stem, SL, NI and INL are derived from CM334 while those increasing leaf and fruit traits such as LA, DWL, SLA, DWF and pt_frt are derived from YW. For traits such as NLE, NF, pt_leaf and Axl, alleles increasing the traits originated from both parents. With average parental difference of about 4 leaves for NLE in NL1, one QTL from YW increased NLE by about 3 leaves while another QTL from CM334 increased it by about 4 leaves. This indicates that YW and CM334 were fixed for alternative alleles at major gene loci, resulting in effects that largely neutralized each other.

The genetic correlation patterns were very similar in the four environments. The correlations showed that related traits (fruit, leaf or stem) cluster together. This clustering was further confirmed in our QTL mapping results as QTLs with pleiotropic effects were found for the correlated traits. Leaf area (LA) was found to be correlated with other vegetative traits such as DWL, DWS, DWV, DWP and SLA; hence the major QTL governing LA on LG 2 was also picked up for these traits. The same scenario was observed for DWF, NF, pt_frt and pt_stem with a QTL influencing them also on LG 2 but at a different position from the QTL affecting leaves. The QTLs on LG 3a and 12 are important for early vegetative development as they affected both Axl and NLE.

A number of studies have already assessed performances of many fruit and some vegetative phenotypes of pepper (Ben Chaim *et al.*, 2001b; Rao *et al.*, 2003; Zygier *et al.*, 2005; Ben Chaim *et al.*, 2006a; Barchi *et al.*, 2009; Mimura *et al.*, 2010). In all of these studies, traits were assessed in a single environment; hence comparing the performances of their populations between environments was not possible. Ben Chaim *et al.* (2001b) and Rao *et al.* (2003) investigated a possible year effect on some fruit and yield-related traits but plants were grown only in summer season and at one location. We have gone further by considering two seasons (summer and winter) and two geographical locations (Temperate and Mediterranean). Generally speaking, our population behaved fairly consistent in all the environments. There were however some differences in the performances of the RILs between the two locations (SP and NL) and between seasons. For most traits, the RILs showed higher phenotypic values in the Mediterranean climate (SP) than in the temperate (NL) and also generally higher in autumn than spring (SP2 > SP1 and NL2 > NL1). Consequently, the number, level of expression, fraction of variance explained and effect sizes of QTLs for most of the traits varied among environments. Most of the differences in QTLs found across environments were quantitative and not qualitative, i.e. they showed the same sign in all the environments and only differed in magnitude. The most significant QTL for LA, found on LG 2, was picked up in all environments with highest level of expression in SP2 which also had the highest phenotypic mean for this trait. The same situation occurred for DWF with the most significant QTL on LG 2 showing highest level of expression in SP1 and SP2, the environments in which the highest fruit yields were obtained. Also, the most significant QTL for NLE found on LG 12 showed the highest level of expression in SP2 and NL2. This is an indication that though many of these traits are genetically determined in any given environment, their degree of expression differs from one environment to the other.

As has already been noted (Utz *et al.*, 2000; Hackett, 2002), failure to detect all QTLs modulating a trait in any experiment might be caused by factors including the genetic structure of the trait, the genetic background of the parents, the size of the mapping population, magnitude of the experimental error, the environment and interactions between QTLs. In the same population like we used, Barchi *et al.* (2009) showed that the number of detected QTLs decreased with population size. Although we used a smaller population, our study also revealed many of the highly significant QTLs found in earlier studies on the same traits (Table 2.6). QTLs for fruit yield (DWF) found on LG 2 and 4 were picked up in most of the studies that evaluated fruit yield. The detected QTL for Axl on LG 3 in this study was also found in the study of Mimura *et al.* (2010), Alimi *et al*. (2010), Barchi *et al.* (2009) and Ben Chaim *et al.* (2001b). Some QTLs were only picked up in some of the studies possibly due to any of the factors listed above including lack of segregation in that particular population and/or population size.

While the same genetic material was used in the four environments, we observed differences in the number of detected QTLs, the magnitudes of their effects and their heritabilities, reflecting possible QTL-by-Environment interactions. Combining data from the four environments and performing a multi-environment analysis would be more powerful (Boer *et al.*, 2007; van Eeuwijk *et al.*, 2010). Also pleiotropic effects of many of the QTLs were observed, which most likely result from relationships between the traits they govern.

Pleiotropy may also suggest redundancy between the measured traits, which could be avoided to decrease the cost of experiments. Such pleiotropic effects can be more accurately studied by explicit modelling of the correlation/covariance structure among the traits through a joint multi-trait analysis. Such joint analysis of the traits will improve the power and precision of QTL mapping (Jiang and Zheng, 1995). Our next task is therefore to perform multi-trait and multi-environment analyses of these data following the approach used in Malosetti *et al.* (2008).

2.6. Conclusion

In this study, we established that phenotypic performances of many pepper physiological traits are usually consistent across the four environments with some variations in level of expression. Many of the traits showed heterosis and transgressive segregation in all environments and mostly in the same direction. Also, for these traits the same QTLs were picked up across the four environments. Most QTLs were only quantitatively different between the environments, though some of them were environment-specific. The directions of the QTL effects generally followed the directions of parental mean differences for the traits. CM334 showed higher mean values for DWS, DWV, Pt_stem, Axl, SL, NLE, NI and INL. Most of the alleles increasing these traits were thus from CM334. Conversely, YW showed higher mean values for LA, DWL, SLA, NF, DWF, DWP, Pt_frt and Pt_leaf. Most of the alleles increasing these traits were thus from YW. QTLs showing pleiotropic effects on many traits were found on LG 1b, 2, 3a, 4, 6 and 12. The pleiotropic effects were consistent with physiological correlations among these traits. These results contribute to a better understanding of the genetics of yield-related physiological traits in pepper and represent a first step in the improvement of the target trait yield.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211347. We thank the SPICY Industrial Advisory Board for support and discussions. Rik van Wijk and Syngenta are especially acknowledged for their highly valuable help in making available additional SNP markers that strongly improved the quality of the genetic map.

Appendix 2A: EU-SPICY Experimental set-up

Table 2A1 Allocation of genotypes to subsets, plots and blocks in the NL experiments, including the number of genotypes (#gtp) and number of occupied plots (#plots). The cumulative (cum.) numbers are given, arriving at 152 genotypes and 264 plots, and the number of replicates within compartment (Rep/B).

Subset	Block					Cum.			Cum.
	1	2	3	4	#gtp	#gtp	Rep/B	#plots	#plots
Common	C_0	C_0	C_0	C_0	3	3	1	12	12
Ladder	T_{41}			T_{41}	3	6	2	12	24
	T_{12}	T_{12}			3	9	2	12	36
		T_{23}	T_{23}		3	12	2	12	48
			T_{34}	T_{34}	3	15	2	12	60
Double	D_{12}	D_{12}			11	26	1	22	82
	D_{13}		D_{13}		11	37	1	22	104
	D_{14}			D_{14}	11	48	1	22	126
		D_{23}	D_{23}		11	59	1	22	148
		D_{24}		D_{24}	11	70	1	22	170
			D_{34}	D_{34}	12	82	1	24	194
Single	S_1				18	100	1	18	212
		S_2			18	118	1	18	230
			S_3		17	135	1	17	247
				S_4	17	152	1	17	264

Common refers to a RIL subset consisting of both parents and F_1.
Ladder refers to four RIL subsets, each consisting of three genotypes (= 12 genotypes). Each genotype appeared in two blocks with single replication.
Double refers to RILs replicated once in two blocks (67 genotypes).
Single refers to RILs with no replication (70 genotypes)

Table 2A2 Representation of a sample block in SP trials, xx represents the 152 genotypes replicated once each, dd are the dummy genotypes while b stands for border plants.

		1	2	3	4	5	6	7	8	
	b	b	b	b	b	b	b	b	b	b
1	b	dd	xx	xx	dd	xx	xx	xx	xx	b
2	b	xx	xx	dd	xx	xx	xx	dd	xx	b
3	b	xx	dd	xx	xx	dd	xx	xx	xx	b
4	b	xx	xx	xx	xx	xx	xx	xx	dd	b
5	b	xx	xx	xx	xx	xx	dd	dd	xx	b
6	b	xx	dd	xx	dd	xx	dd	xx	xx	b
7	b	xx	xx	dd	xx	xx	xx	dd	xx	b
8	b	xx	xx	xx	xx	xx	xx	xx	xx	b
9	b	dd	xx	xx	xx	xx	xx	xx	xx	b
10	b	xx	dd	xx	xx	dd	xx	xx	dd	b
11	b	xx	xx	xx	dd	xx	xx	xx	xx	b
12	b	dd	xx	dd	xx	xx	dd	xx	xx	b
13	b	xx	xx	xx	xx	xx	xx	xx	xx	b
14	b	xx	xx	xx	xx	dd	xx	dd	xx	b
15	b	xx	xx	xx	xx	xx	xx	xx	xx	b
16	b	xx	xx	dd	xx	xx	xx	xx	dd	b
17	b	xx	dd	xx	xx	xx	xx	xx	xx	b
18	b	xx	xx	xx	dd	xx	dd	xx	dd	b
19	b	xx	xx	xx	xx	xx	xx	xx	xx	b
20	b	xx	xx	dd	xx	xx	xx	xx	xx	b
21	b	dd	xx	xx	xx	dd	xx	dd	xx	b
21	b	dd	xx	xx	xx	xx	xx	xx	xx	b
23	b	xx	xx	xx	dd	xx	dd	xx	xx	b
24	b	xx	dd	xx	xx	dd	xx	xx	dd	b
	b	b	b	b	b	b	b	b	b	b

Appendix 2B: Phenotypic Mean comparison

Table 2B1 Phenotypic Mean comparison for environment NL1

Traits	YW	CM334	F_1	RIL	k	D_{RIL}	Q_{min}	Q_{max}
LA	6441.82	2288.03	6002.91	4643.69	-0.79	-1359.22	0.02	0.08
DWL	26.43	12.22	32.83	25.57	-1.90$^+$	-7.26$^+$	0.03	0.41
DWS	24.78	71.21	102.18	76.65	2.33$^+$	-25.53$^+$	0.01	0.62
DWV	51.84	83.52	134.45	102.15	4.22$^+$	-32.30$^+$	0.03	0.77
SLA	25.49	20.22	18.73	18.60	1.57$^+$	-0.13	0.76	0.01
NF	6.29	3.25	16.30	12.54	-7.59$^+$	-3.76$^+$	0.09	0.81
DWF	49.03	0.74	37.16	23.12	-0.51$^+$	-14.04$^+$	0.03	0.08
DWP	101.93	84.32	171.04	124.97	-8.85$^+$	-46.07$^+$	0.06	0.80
pt_frt	0.48	0.01	0.21	0.18	0.15	-0.03$^+$	0.03	0.01
pt_leaf	0.27	0.15	0.19	0.21	0.33	0.02$^+$	0.06	0.07
pt_stem	0.25	0.84	0.60	0.61	0.19	0.01$^+$	0.00	0.00
Axl	21.75	38.75	36.25	27.12	0.71$^+$	-9.13$^+$	0.22	0.05
SL	23.94	73.27	78.94	59.83	1.23$^+$	-19.11$^+$	0.00	0.11
NLE	11.56	15.75	13.94	12.94	0.14	-1.00$^+$	0.26	0.10
NI	4.69	5.56	6.50	6.13	3.16$^+$	-0.37$^+$	0.03	0.76
INL	1.89	2.46	2.60	2.12	1.49$^+$	-0.48$^+$	0.32	0.19

$^+$Significant at 0.05 level of significance

Table 2B2 Phenotypic Mean comparison for environment NL2

Traits	YW	CM334	F_1	RIL	k	D_{RIL}	Q_{min}	Q_{max}
LA	6648.09	2338.83	5931.49	4894.72	-0.67	-1036.77$^+$	0.02	0.07
DWL	25.48	11.41	27.24	21.22	-1.25$^+$	-6.02$^+$	0.04	0.23
DWS	19.02	52.43	76.48	52.57	2.44$^+$	-23.91$^+$	0.01	0.53
DWV	43.96	63.77	104.83	73.75	5.15$^+$	-31.08$^+$	0.06	0.74
SLA	27.74	21.34	22.98	23.30	0.49	0.32	0.20	0.05
NF	3.50	1.00	13.18	6.66	-8.74$^+$	-6.52$^+$	0.09	0.69
DWF	17.67	0.22	16.53	7.77	-0.87$^+$	-8.76$^+$	0.01	0.10
DWP	61.63	63.99	121.36	81.08	49.62$^+$	-40.28$^+$	0.14	0.85
pt_frt	0.28	0.00	0.13	0.10	0.07	-0.03	0.01	0.02
pt_leaf	0.41	0.18	0.22	0.26	0.65$^+$	0.04$^+$	0.00	1.00
pt_stem	0.31	0.82	0.64	0.64	0.29	0.00	0.86	0.00
Axl	33.44	41.94	49.69	40.71	2.82$^+$	-8.98$^+$	0.14	0.43
SL	37.69	128.31	133.69	107.37	1.12$^+$	-26.32$^+$	0.00	0.16
NLE	12.63	13.44	11.81	11.97	-3.02$^+$	0.16	0.71	0.14
NI	9.19	17.44	17.56	15.12	1.03$^+$	-2.44$^+$	0.00	0.10
INL	2.67	3.12	4.21	3.42	5.84$^+$	-0.79$^+$	0.07	0.71

$^+$Significant at 0.05 level of significance

Table 2B3 Phenotypic Mean comparison for environment SP1

Traits	YW	CM334	F_1	RIL	k	D_{RIL}	Q_{min}	Q_{max}
LA	7399.86	3513.49	6549.43	6985.59	-0.56	436.16	0.01	0.38
DWL	47.36	30.37	51.56	57.02	-1.49$^+$	5.46	0.03	0.71
DWS	36.66	111.31	108.10	113.54	0.91	5.44	0.01	0.53
DWV	70.01	141.67	159.66	167.35	1.50$^+$	7.69	0.01	0.73
SLA	15.97	11.78	12.83	12.33	0.50	-0.50	0.37	0.01
NF	13.17	4.67	45.50	28.98	-8.61$^+$	-16.52$^+$	0.03	0.92
DWF	144.89	3.14	131.94	66.89	-0.82$^+$	-65.05$^+$	0.01	0.01
DWP	228.90	144.81	291.60	237.45	-2.49$^+$	-54.15$^+$	0.01	0.60
pt_frt	0.63	0.02	0.46	0.28	-0.44	-0.18$^+$	0.01	0.00
pt_leaf	0.21	0.21	0.18	0.24		0.06$^+$	0.27	0.73
pt_stem	0.16	0.77	0.37	0.48	-0.31	0.11$^+$	0.00	0.00
Axl	22.50	33.67	31.00	26.93	0.52	-4.07$^+$	0.16	0.07
SL	23.28	52.43	58.28	41.39	1.40$^+$	-16.89$^+$	0.01	0.06
NLE	11.50	16.50	13.67	12.94	-0.13	-0.73	0.21	0.04
NI	6.83	9.17	9.50	8.19	1.28$^+$	-1.31$^+$	0.11	0.21
INL	1.98	2.06	2.27	2.12	6.25$^+$	-0.15	0.42	0.50

$^+$Significant at 0.05 level of significance

CHAPTER 3

Multi-Trait and Multi-Environment QTL Analyses of Yield and A Set of Physiological Traits in Pepper

CHAPTER 3

Multi-Trait and Multi-Environment QTL Analyses of Yield and A Set of Physiological Traits in Pepper

3.1. Abstract

For many agronomic crops, yield is measured simultaneously with other traits across multiple environments. The study of yield can benefit from joint analysis with other traits and relations between yield and other traits can be exploited to develop indirect selection strategies. We compare the performance of three multi-response QTL approaches based on mixed models: a multi-trait approach (MT), a multi-environment approach (ME), and a multi-trait multi-environment approach (MTME). The data come from a multi-environment experiment in pepper, for which 15 traits were measured in four environments. The approaches were compared in terms of number of QTLs detected for each trait, the explained variance, and the accuracy of prediction for the final QTL model. For the four environments together, the superior MTME approach delivered a total of 47 regions containing putative QTLs. Many of these QTLs were pleiotropic and showed quantitative QTL by environment interaction. MTME was superior to ME and MT in the number of QTLs, the explained variance and accuracy of predictions. The large number of model parameters in the MTME approach was challenging and we propose several guidelines to help obtain a stable final QTL model. The results confirmed the feasibility and strengths of novel mixed model QTL methodology to study the architecture of complex traits.

Keywords

Pepper; Complex trait; Genetic Correlation; Pleiotropy; QTL by Environment Interaction; Quantitative Trait Locus

3.2. Introduction

Yield and other complex traits of agronomic importance are typically measured for collections of genotypes across multiple environments, and genotype by environment interactions is common (GEI)[1] (Van Eeuwijk et al., 2010): superiority of genotypes can change in relation to the environment. The statistical genetic analyses of complex traits showing GEI can effectively be addressed by mixed model methodology with terms for QTL by Environment Interaction (QEI) (Boer et al., 2007). QTLs can then be categorized according to the stability of their effects across different environments. A 'constitutive' QTL is consistently detected across most environments, while an 'adaptive' QTL is detected only in specific environmental conditions, or increases in expression with the level of an environmental factor (Vargas et al., 2006).

For measurements obtained simultaneously for several traits, it is more appropriate to perform statistical analyses multivariately than univariately. This requirement is even stronger when biological processes are interdependent. Traits are genetically correlated and proper QTL mapping helps differentiating whether correlations are due to pleiotropic QTLs or closely linked QTLs. Analyzing correlated traits univariately, leads to higher sampling variances of estimated parameters and lower power for hypothesis tests. The joint analysis of multiple traits has been shown to improve the power and precision of QTL mapping. It has also helped in improving the selection of some primary traits with low heritabilities or that are difficult to measure by exploiting their genetic correlations with other traits (Jiang and Zeng, 1995).

Recent advances in statistical genetics methodology have led to extensions of the traditional QTL mapping techniques and the mixed model is now the approach of choice (Van Eeuwijk et al., 2010; Vilhjalmsson and Nordborg, 2013). This is a result of the suitable framework offered by mixed models in handling many of the challenges present in QTL analysis, including simultaneous observations on many traits and across multiple environments, the possibility of unequal replication of genotypes either due to experimental design and/or missing observation and phenotypic measurements over time (Verbeke and Molenberghs, 2000). Furthermore, mixed models do not rely on unrealistic assumptions, such as zero genetic correlations between environments and traits, and constant variance across environments. It can account for both intra- and inter-trial variability in the estimation of QTL effects and trait values prediction (Van Eeuwijk et al., 2010). Mixed models have been extensively applied in many QTL mapping settings (Anhalt et al., 2009; Boer et al., 2007; Hackett et al., 2001; Klasen et al., 2012; Korte et al., 2012; MacMillan et al., 2006; Malosetti et al., 2008; Malosetti et al., 2006; Malosetti et al., 2004; Panozzo et al., 2007; Piepho, 2000; Verbyla et al., 2003; Xu, 2013), ranging from single trait single environment analysis up to the most complex setting of multi-trait multi-environment (MTME) with various interactions (traits, environments and/or environmental characterizations).

In pepper, GEI and QEI approaches have not been used previously to map multiple quantitative traits in multiple environments. Earlier studies focused mostly on univariate analyses of traits in single environments (Alimi et al., 2012; Alimi et al., 2013a; Barchi et al., 2009; Ben Chaim et al., 2006; Ben Chaim et al., 2001; Kargbo and Wang, 2010; Lee et al.,

[1] The list of all abbreviations is given in Table 3A1 in Appendix 3A.

2008; Lefebvre et al., 2003; Mimura et al., 2010; Rao et al., 2003; Zygier et al., 2005). In MTME analysis, the most challenging aspect often arises from the number of trait by environment combinations (TE's) in relation to computational requirements. This paper contains a large implementation of MTME in QTL analysis with emphasis on how to circumvent some of the computational issues that may arise due to the increase in the number of parameters being estimated. In this paper, we implemented three different multivariate modelling strategies to analyse data on a recombinant inbred line (RIL) pepper population (Alimi et al., 2013a; Voorrips et al., 2010; www.spicyweb.eu). These modelling strategies are multi environment (ME), multi trait (MT) and multi-trait multi-environment (MTME) analyses. We modelled genetic correlations within (between traits in a given environment) and between environments, and explicitly test the presence of QEI and pleiotropic QTLs. In the GEI stage, we performed multi-environment (ME) analysis for each trait to investigate GEI. In the multi-trait (MT) analysis, we combined the 15 traits for each trial in a joint analysis to investigate pleiotropic QTLs. We thereafter created factorial combinations of traits and environments for use in the MTME analysis. We employed unstructured covariance model which allowed each pair of TE combinations to have unique covariance. We then searched for main effect QTLs and QEI effects, by including genome-wide marker data. We investigated accuracy of predictions by the fitted QTL models from each of the three methods and discuss the relative improvements of the final QTL results. We further reduced the TE combinations through principal component analysis. QTL analysis was then performed on the selected components to investigate if QTLs similar to those from ME, MT and MTME analyses would be detected.

3.3. Materials and Methods

3.3.1. Plant materials, marker data and phenotypic evaluation

We summarize the main features of the data here. A detailed description can be found in Alimi et al. (2013). The mapping population consists of sixth generation (F_6) and still segregating recombinant inbred lines (RILs) of an intraspecific pepper cross between the large – fruited inbred cultivar 'Yolo Wonder' (YW) and the pungent small-fruited cultivar 'Criollo de Morelos 334' (CM 334). DNA was extracted from 149 RILs to produce information for 455 markers assembled into 12 pepper chromosomes, covering 1705cM (Figure 3.1). The map used here is an improved version of the map used in Alimi et al. (2013) which had five chromosomes with two linkage groups each. All chromosomes now have only one linkage group each. The majority of markers used in the current map are SNP and SSR markers. Almost all the AFLP markers in the former map were discarded (Nicolaï et al., 2012). The percentage of missing genotype information across the full set of markers was 13.7%. None of the markers showed segregation distortion.

Phenotypic evaluations of the RILs were carried out via designed greenhouse experiments across two locations; Spain (SP) and the Netherlands (NL). The trials were conducted under both spring (1) and autumn (2) weather conditions in 2009. This gave a total of four trials (i.e. environments); Netherlands trial in spring (NL1), Netherlands trial in autumn (NL2), Spain trial in spring (SP1) and Spain trial in autumn (SP2). A total of 15 traits (Table 3.1) were analyzed, 13 of which were already detailed in Alimi et al. (2013). Two additional traits, increase rate of leaf area index (LAI) and light use efficiency (LUE), were added. LAI

expresses mean increase in leaf area index per unit time, where time is expressed in degree-days. LUE is the dry matter production (g) per megajoule (MJ) of intercepted global radiation. LUE was estimated as the slope of a graph in which the increase in total plant biomass was plotted against the cumulative amount of intercepted light.

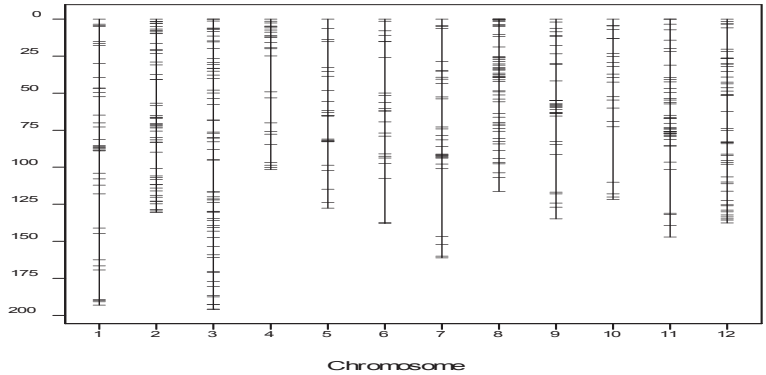


Figure 5.1 The final genetic map showing the 12 pepper chromosomes and positions of markers used in the study

Table 5.1 Traits measured in each of the four SPICY environments (experiments).

Abbreviation	Trait
DWF[1]	Total fruit dry weights from each plant (g)
NF	Total number of fruits
pt_frt	Proportion of the total biomass due to fruit
DWL	Dry weight of leaf (g)
DWS	Dry weight of stem (g)
DWV	Dry weight of vegetative part (g)
LUE	Dry matter production (g) per megajoule (MJ) of intercepted global radiation (g/MJ)
LAI	Mean increase in leaf area index per unit time (m^2 m^{-2} $°Cd^{-1}$)
pt_leaf	Proportion of the total biomass due to leaf
Axl	Primary Axis length (Stem length before first branching) (cm)
SL	Stem length measured 6-8 weeks after transplanting (cm)
NLE	Number of Leaves on the primary axis
NI	Number of Internodes at time 3-4 weeks after transplanting
INL	Internode length for the primary axis (cm)
SLA	Specific Leaf Area (m^2/g)

[1]representative for yield

3.3.2. Multi-environment phenotypic and QTL analysis

Each trait was evaluated over the four trials with the aim of investigating genotype-by-environment interaction (GEI) and QTL-by-environment interaction (QEI). As data for this analysis, for each RIL, we used best linear unbiased estimates (BLUE) per environment from an earlier analysis reported in Alimi et al. (2013). To enhance numerical stability, for each trait scale effects were removed and the BLUE values were standardized such that they form a distribution with mean equal to zero and standard deviation equal to one.

Following Boer et al. (2007), the multi-environment phenotypic analysis and QTL estimation were combined. For QTL detection so-called genetic predictors (functions of conditional QTL genotype probabilities) need to be calculated. The genetic predictors were calculated at all 455 marker positions and 184 intermediate positions for those marker intervals that were larger than 5cM, genomic positions will be indexed by q, with q = 1, 2,…, 639. The genetic predictor for individual i at genomic evaluation point q is denoted by x_{iq}. The genetic predictors for the additive QTL effect had the value x_{iq} = -1 if both alleles at a fully informative marker arose from parent 1 (YW), or x_{iq} = 1 if they arose from parent 2 (CM334). At intermediate positions and marker positions with missing marker genotypes, these integer values were replaced by linear combinations of conditional QTL genotype probabilities given marker information. Starting with fitting single QTL models using simple interval mapping (SIM) (Lander and Botstein, 1989b),

$$y_{ij} = E_j + x_{iq}\alpha_{jq} + g_{ij} + \varepsilon_{ij}. \quad (3.1)$$

Where y_{ij} denotes the standardized phenotype of the ith genotype (i = *1,...,149*) in environment j (j = *1,...,4*), E_j is the environmental mean, g_{ij} represented the genetic effect of genotype i at environment j, and ε_{ij} represented the non-genetic component. We assumed that the vectors $\boldsymbol{g_i} = (g_{i1}, \ldots, g_{ij})$ follow a multivariate normal distribution with zero mean and an unstructured VCOV matrix G i.e. $g_i \sim N(0, \boldsymbol{G})$. α_{jq} was the environment-specific QTL main effect at evaluation point q. Testing for the significance of α_{jq} was done through Wald tests (Verbeke and Molenberghs, 2000) with H_0: $\alpha_{1q} = \alpha_{2q} = \alpha_{3q} = \alpha_{4q} = 0$, where $\alpha_1,\ldots,\alpha_4$ refers to the QTL effect at each of the four environments. From the fit of model (1), the map positions showing significant deviations from H_0 were selected and the corresponding genetic predictors were set as cofactors in subsequent composite interval mapping (CIM) (Zeng, 1994).

$$y_{ij} = E_j + \sum_{c \in C} x_{ic}\alpha_{jc} + x_{iq}\alpha_{jq} + g_{ij} + \varepsilon_{ij}, \quad (3.2)$$

where C was the set of cofactors. The cofactor selection thresholds were determined using an approach described by Li and Ji (Li and Ji, 2005), with genome-wide significance level set at 0.05. CIM was run at least twice consecutively to confirm stability of the test statistic profiles. The full set of significant positions from CIM was subjected to a backward selection procedure to arrive at the final QTL model (Boer et al., 2007). The minimum distance between significant QTLs was assumed to be 20cM for the final QTL model. In the final QTL model significant QEI effects were determined by testing significance of environment-specific deviations from the main environmental effect through a Wald test. In this case, an effect was called significant when its P-value was below the significance level of 0.05, no correction for multiple testing was applied at this stage.

3.3.3. Multi-trait QTL estimation

The specification of multi-trait (MT) model is very similar to the ME model. In the case of MT model, instead of having environment (E) in QTL model (2), we have trait (T). Per environment, there were 15 traits, resulting in four MT analyses. With the inclusion of multiple QTLs as cofactors, the QTL model for CIM is:

$$y_{ip} = T_p + \sum_{c \in C} x_{ic}\alpha_{pc} + x_{iq}\alpha_{pq} + g_{ip} + \varepsilon_{ip}, \quad (3.3)$$

where T_p (p = 1, 2,..., 15) is the trait mean, α_{pq} is the trait-specific QTL main effect at evaluation point q, g_{ip} represents the genetic effect of genotype i for trait p, and ε_{ip} is the residual effect. This model allowed us to explicitly model genetic correlations between traits by specifying an unstructured VCOV matrix among each pair of traits giving a total of 120 parameters. It further allowed us to identify QTLs with pleiotropic effects. Synergistic pleiotropy refers to positive covariance between the effects of a gene or gene substitution on two or more traits, based upon correspondence in expression (sign of effects) with regards to the traits. This implies that the increasing alleles for all the traits being influenced by the pleiotropic QTL are from just one of the parents. In antagonistic pleiotropy, pleiotropic effects of a QTL are opposite in sign, positive in one context of expression and negative in another (West-Eberhard, 2003).

3.3.4. Multi-Traits Multi-Environments QTL estimation

Extension to multi-trait multi-environment (MTME) setting was achieved by combining traits across the four environments in a single mixed model analysis. ME and MT models are extended by allowing the response trait (y) to be a vector of the traits (T) and environments (E) combinations. The mean for the trait by environment combination, TE, is taken as fixed in the QTL analysis. We restricted ourselves to SIM method for the MTME as CIM could not be implemented successfully as a result of increase in the number of parameters after adding cofactors. The model for SIM is:

$$y_{iz} = TE_z + x_{iq}\alpha_{zq} + g_{iz} + \varepsilon_{iz}, \qquad (3.4)$$

where TE_z (z = 1, 2,..., 60) is the TE mean (z is the product of four environments and 15 traits = 60), α_{zq} is the environment-specific and trait-specific QTL main effect at evaluation point q, g_{iz} represents the genetic effect of genotype i for TE z, and ε_{iz} is the residual effect. We specified an unstructured VCOV matrix for all pairs of the TE combinations, giving a total of 1830 parameters. With the MTME model, GEI and genetic correlations between traits were simultaneously modelled.

3.3.5. MTME final QTL selection and window size

We performed the SIM scan and carried out a backward selection on the significant positions. An initial step was taken to determine an optimal QTL peak window size for the final QTL model, that is, what should be the minimum distance between consecutive QTLs at a chromosome. We investigated QTL window sizes ranging from 5cM to 40cM. When QTL window sizes above 20cM were used, some putative QTLs were missed. Using window sizes below 20cM led to selecting some QTLs at very close distance that affected the same set of traits and thus looked as representing a single QTL. A window size of 20cM was found to be optimum for our data and was used in the final QTL modelling step. The final QTLs were selected using a peak window size of 20 cM and taking into account changes in the signs of neighbouring QTLs. If for two QTLs next to each other, the signs for QTL effects remained unchanged over all TEs, the QTLs were interpreted to represent the same QTL and only the position showing the strongest effects was retained in the final QTL model.

The phenotypic and QTL analyses were performed using the QTL facilities in GenStat 15 (VSNi, 2012).

3.3.6. Comparisons of ME, MT and MTME approaches

For the three QTL mapping methods, the number of significant QTLs and their explained variance for each of the TE combinations, e.g. Axl in NL1 (Axl.NL1) were compared. We also investigated whether the same QTL positions were detected for a given TE by the different methods. This enabled us to confirm if QTLs as detected by simpler methods were not lost in the more complex methods. Predictive accuracies of the models were also explored and compared. Predictive accuracy was defined conveniently, although slightly simplistically, as the correlation between BLUE and predicted phenotypic values from the final QTL models in the three approaches. (More in depth treatment of predictive accuracy of various QTL and genomic prediction methods will be submitted in a follow up paper.)

3.4. Results

3.4.1. Genetic correlations between traits (within and between trials)

The genetic correlations of traits among environments are given in Table A2 in Appendix 3A, while the genetic correlations between traits within each trial are presented with the aid of biplots from the first two principal components of the traits (Appendix 3B, Figure 3B1). The correlations between the four environments for individual traits were mostly comparable (uniform correlations) and were generally moderate to high, ranging from 0.30 for NI between NL2 and SP1 to 0.86 for NLE between NL1 and NL2. Overall mean of the genetic correlation was 0.62, with the majority of the correlations above 0.5. Trait variances differed over environments (Appendix 3A, Table 3A3). Within trial correlations were consistent in sign within the trials (Appendix 3B, Figure B1). Many of the correlations were according to physiological expectation, considering the relationships between traits, where one trait was computed from others (e.g. DWV from DWS and DWL), or traits related jointly to a part of the plant, e.g. fruit related traits such as DWF, NF and pt_frt. There were some very high (e.g. between LAI and DWL) and very low (DWF and NLE) correlations, but most correlations between traits within environments were moderate. Some negative correlations were considered remarkable; they depicted resource allocation competitions between plant organs. For example pt_leaf was negatively correlated to fruit related traits such as NF, DWF and pt_frt. These negative correlations were more pronounced in SP trials than in NL trials.

3.4.2. Multi-Environment analyses

The plot of the CIM genome scan for DWF (yield) for the ME approach is given in Figure 3.2. The plots of the CIM genome scans for the other traits are presented in Appendix 3C, Figure 3C1. Table C1 in Appendix 3C presents the QTL positions and effects for all 15 traits. For DWF, three significant QTLs were detected on chromosomes 2, 4 and 7, respectively. Two of these QTLs (C4-35cM and C7-79cM) were constitutive i.e. these showed consistent significant effects across the four environments. The QTL on chromosome 2 showed QEI effects in magnitude, but not in direction (= non-crossovers). Such QEI are regarded as quantitative; i.e., the effects had the same sign in all environments. Generally for most traits, QEI effects were quantitative. However, one QTL on chromosome 11 (~70cM) showed significant crossover interactions (i.e. qualitative QEI) for the traits LUE, Axl, SL and INL in SP1 and SP2 environments. This particular QTL may be categorized as location specific and adaptive as it was significant only in Spanish trials (Appendix 3C).

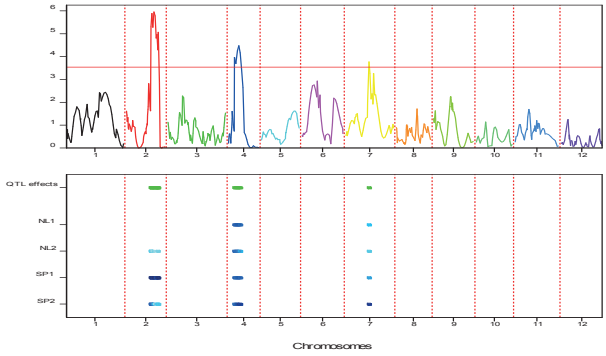


Figure 3.2 CIM Profile plot of the Multi-Environment analyses for yield (DWF)
The top section shows the *P*-values of tests for QTL main effects. The bottom section shows heat maps along the genome for each environment, where blue means that the YW allele had a significant positive effect and red means that the CM334 allele had a significant positive effect in that environment (the darker the colour, the higher the significance level of the QTL). Three QTLs were detected on chromosomes 2, 4 and 7. The QTLs showed no crossovers across environments.

3.4.3. Multi-trait analyses

The plots of CIM genome scans for the MT analysis in the four environments (Figure 3.3) showed many significant QTLs across the genome, influencing different traits to different magnitude and direction. After applying backward selection on the CIM scan, a total of 13, 17, 16 and 15 QTL regions exceeded the significance threshold in NL1, NL2, SP1 and SP2 respectively. All QTLs showed pleiotropic effects, i.e., multiple traits were affected by the same QTL. A few of these pleiotropic QTLs displayed synergistic pleiotropic effects while many of them showed antagonistic pleiotropic effects. Clear examples of synergistic pleiotropic QTLs were found on chromosomes 4@70cM in NL1, 4@11cM in NL2, 7@35cM in NL2 and 3@40cM) in NL1. An example of an antagonistic QTL was present on chromosome 3 (~150cM) in SP2. This QTL showed increasing effects from YW on fruit related traits (DWF and pt_frt) and increasing effects from CM334 on other traits such as SL, NLE, NI, Axl and LUE. Many of these pleiotropic QTLs are consistent with genetic correlations among the traits. As an example, the QTLs on chromosomes 2 and 4, influencing pt_leaf and fruit traits such as DWF showed antagonistic pleiotropy especially in SP trials, which is consistent with the negative correlations that exist between pt_leaf and the fruit traits. For many traits, MT analyses revealed more QTLs than the ME analyses (Table 3.2). These QTLs also explained more genetic variations than those from ME analyses. In SP2, about 10 QTLs were detected for DWF including the three QTLs detected in ME analyses. These QTLs explained about 45% of genetic variation against 29% explained by the three QTLs from ME analyses. The MT QTL positions and effects for each of the environments are presented in Appendix 3D.

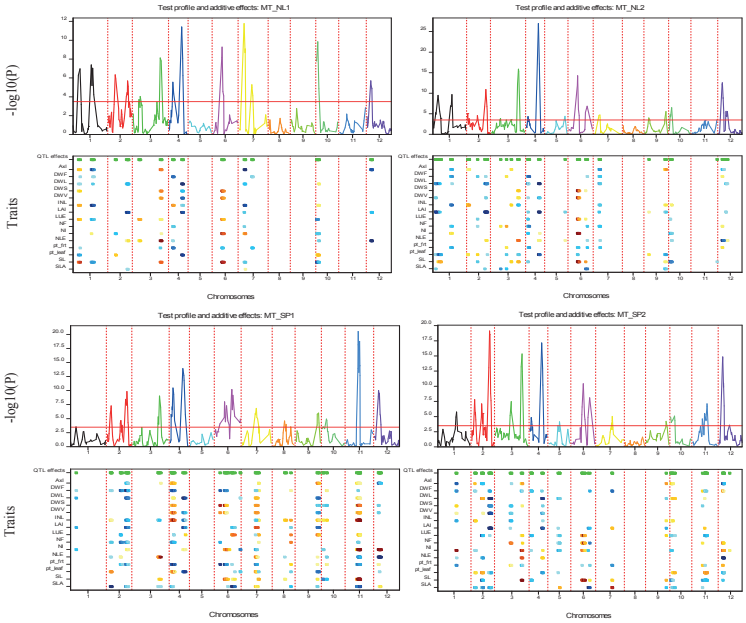


Figure 5.3 CIM Profile plots of the Multi-Trait analyses for the four environments
The top section shows the *P*-values of tests for QTL main effects. The bottom section shows heat maps along the genome for each trait, where blue means that the YW allele had a significant positive effect and red means that the CM334 allele had a significant positive effect on the given trait (the darker the colour, the higher the significance level of the QTL). Most of the QTLs showed pleiotropies which were most times antagonistic.

Table 5.2 Comparison of Number of QTLs (#QTL) and Explained Variance (H2(qtl)) from SE, ME, MT and MTME models

		Number of QTLs (#QTL)				QTL Variance Explained ($H^2_{(qtl)}$)					
Trait	Method	NL1	NL2	SP1	SP2	NL1	NL2	SP1	SP2	Avg. #QTL	Avg. $H^2_{(qtl)}$
DWF	SE	1	2	3	3	18	18	25	37	2.3	25
	ME	2	3	3	3	21.8	24.9	39.6	28.7	2.8	28.8
	MT	4	8	9	10	23.6	46	42.8	44.6	7.8	39.3
	MTME	7	10	10	13	32.1	53.1	44.3	56	10.0	46.4
NF	SE	0	1	4	3	0	10	31	34	2.7	25
	ME	4	2	2	3	20.4	14	15.2	33.5	2.8	20.8
	MT	2	7	3	6	7.7	33.1	17.5	39.5	4.5	24.5
	MTME	6	9	7	4	27.4	40.6	35.3	28.9	6.5	33.1
pt_frt	SE	0	0	4	3	0	0	32	26	3.5	29
	ME	4	3	6	3	42.1	34.1	66.6	36.4	4.0	44.8
	MT	5	7	5	5	25.2	35.3	34.1	27.9	5.5	30.6
	MTME	7	11	10	10	33.2	54	46.3	44.3	9.5	44.5
DWL	SE	2	3	2	3	25	28	18	39	2.5	27.5
	ME	5	5	6	5	43.4	46.1	53.8	44.3	5.3	46.9
	MT	4	4	6	4	33.3	33	32.4	48.5	4.5	36.8
	MTME	7	6	13	7	41.7	34	60.4	52	8.3	47.0
DWS	SE	1	2	3	1	11	18	28	11	1.8	17
	ME	1	1	2	3	10.6	12.4	29.6	26.6	1.8	19.8
	MT	4	2	4	3	25.7	23.5	34.6	21.3	3.3	26.3
	MTME	7	4	7	4	33.6	21.6	48.9	23.2	5.5	31.8
DWV	SE	0	1	1	2	0	16	9	23	1.3	16
	ME	1	2	2	2	6.7	10.9	18.1	23	1.8	14.7
	MT	3	3	6	4	23.3	21.7	32.8	32.6	4.0	27.6
	MTME	6	3	8	6	34.4	18.4	52.9	39.7	5.8	36.4

LUE	SE	2	1	1	4		26	14	17	31	2	22
	ME	3	2	2	3		25.8	19.5	11.6	22.5	2.5	19.9
	MT	6	2	5	4		32.3	15.5	24.3	33.7	4.3	26.5
	MTME	8	8	10	12		44.6	36	49.3	52.4	9.5	45.6
LAI	SE	2	3	2	2		33	48	22	42	2.3	36
	ME	4	4	6	5		37.4	49.7	57.8	42.7	4.8	46.9
	MT	4	4	5	4		30	39.9	31.6	50.9	4.3	38.1
	MTME	5	7	10	7		35.9	44.4	47.8	46.8	7.3	43.7
pt_leaf	SE	3	4	2	2		26	34	19	12	2.8	23
	ME	1	1	3	5		8	12.2	25.3	48.3	2.5	23.5
	MT	3	4	6	5		23.7	28.8	30	28.5	4.5	27.8
	MTME	6	7	10	7		33.4	31.3	45.5	33.6	7.5	36.0
Axl	SE	3	5	0	3		38	31	0	24	3.7	31
	ME	3	3	5	2		40.3	26.2	30.6	14.5	3.3	27.9
	MT	5	5	5	6		39.5	26.5	27.1	33.4	5.3	31.6
	MTME	10	5	10	6		46.5	26.2	42.4	30.1	7.8	36.3
SL	SE	3	4	1	5		22	35	14	30	3.3	25
	ME	3	2	4	3		24.4	22.4	44.6	28.3	3.0	29.9
	MT	5	5	5	6		39.7	34.3	35.3	42.4	5.3	37.9
	MTME	6	6	11	9		35.9	38	61.6	45.7	8.0	45.3
NLE	SE	2	2	3	1		36	42	29	36	2	36
	ME	2	2	2	2		38.8	23.1	31.6	42.1	2.0	33.9
	MT	3	4	4	3		31.3	22.8	35.7	42	3.5	33.0
	MTME	9	4	8	8		61.6	26.8	58.2	63.4	7.3	52.5
NI	SE	3	3	2	4		34	40	26	37	3	34
	ME	1	1	3	3		18.7	29	36.1	38.1	2.0	30.5
	MT	5	2	5	5		26	25.8	45.3	37.8	4.3	33.7
	MTME	7	5	6	10		37.7	26.8	52.6	48.1	7.0	41.3
INL	SE	4	3	3	0		42	24	29	0	3.3	32
	ME	4	3	5	2		42.8	23.7	50.4	17.3	3.5	33.6
	MT	4	7	6	3		34.1	32.6	37.5	13.6	5.0	29.5
	MTME	11	10	11	13		50.1	45.3	59	61.4	11.3	54.0
SLA	SE	1	1	3	5		13	14	36	49	2.5	28
	ME	2	4	5	5		7.1	33.5	39.4	39.9	4.0	30.0
	MT	3	2	4	5		8.4	12.2	38.5	33.7	3.5	23.2
	MTME	4	5	7	8		27.1	31.2	36.8	48.4	6.0	35.9
Means Across Traits	SE	1.8	2.3	2.3	2.7		21.6	24.8	22.3	28.7	2.6	27.1
	ME	2.7	2.5	3.7	3.3		25.9	25.4	36.7	32.4	3.1	30.1
	MT	4.0	4.4	5.2	4.9		26.9	28.7	33.3	35.4	4.6	31.1
	MTME	7.1	6.7	9.2	8.3		38.3	35.2	49.4	44.9	7.8	42.0

3.4.4. Multi-trait multi-environment analysis

The plot of the SIM genome scan for the MTME analysis using an unstructured VCOV is given in Figure 3.4. A total of 47 regions were identified as harbouring putative QTLs. Chromosomes 4 and 10 had the smallest number of QTLs (=2) while chromosomes 1 and 3 had the highest number of QTLs (=6). Similar to the results from MT analyses, pleiotropic QTLs were observed for genetically correlated traits. The majority of the 47 QTLs showed antagonistic pleiotropic effects, i.e., the increasing alleles originated from both parents for different traits. Five QTL with synergistic pleiotropic effects for the YW parent (contributing the increasing allele) were found on chromosomes 2 (31cM), 4 (53cM), 7 (0cM), 11 (20cM), and 12 (75cM). Also for parent CM334, five of these QTL were found on chromosomes 2 (128 cM), 3 (135 cM), 5 (38 cM), 6 (0 cM), and 8 (19 cM). The majority of the pleiotropic QTLs were not constitutive as they were not consistently affecting particular traits across all

environments. This means that many of the QTLs displayed QEI. The QEI were mostly quantitative, but there were some qualitative QEI especially on chromosome 11 for LUE, Axl, SL and INL, similar to the results from ME analyses. Table 3.3 contains the list of QTL positions from chromosomes 1 and 2 as detected from MTME analysis after backward selection. Results for the remaining chromosomes are in appendix 3E, Tables 3E1 to 3E4.

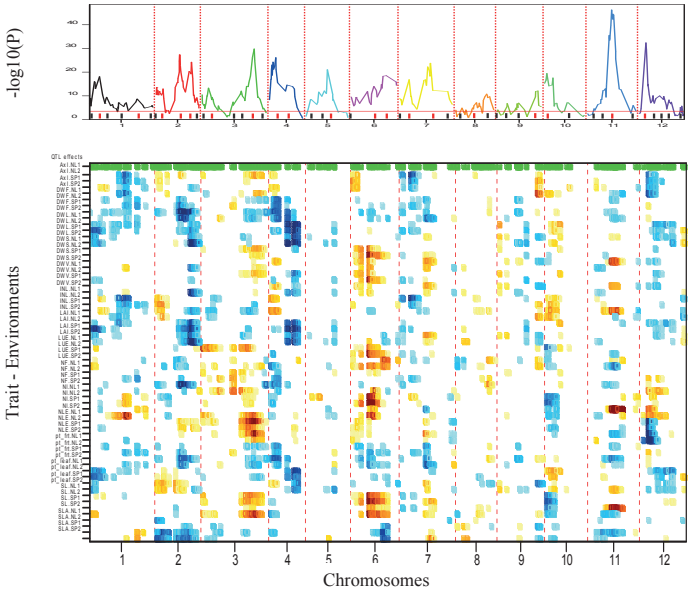


Figure 5.4 SIM Profile Plot for Multi-Trait Multi-Environment Analysis
The top section shows the *P*-values of tests for QTL main effects across all trait-environment combinations with the bars on the x-axis indicating the 47 QTL positions after backward selection. The bars in red indicate QTL positions similar to significant positions from ME and MT analyses while those in black are unique to MTME. The bottom section shows heat maps along the genome for each trait, where blue means that the YW allele had a significant positive effect and red means that the CM334 allele had a significant positive effect on the given trait-environment (the darker the colour, the higher the significance level of the QTL).

3.4.5. Comparison of MT, ME, and MTME results

In environment SP2, a total of 13 QTLs were detected for DWF in the MTME analysis, 3 and 10 more than those from MT and ME analyses respectively. The percentages explained variances by these QTL jointly were 56%, 45% and 29% in the MTME, MT and ME analyses respectively (Table 3.2). QTL effects for DWF on chromosomes 3 and 4 were significant in the four environments. DWF QTLs were in many cases pleiotropic to other yield related traits such as pt_frt and NF. Such pleiotropic QTLs were observed on chromosomes 2, 3, 4, 6 and 12 (Figure 3.5). Pleiotropy with other traits was also observed such as with Axl, NI and INL on chromosome 1; with DWL, DWS, DWV, LAI, LUE and INL on chromosome 2. Others were with LUE, SLA, SL and NI on chromosome 6 and NLE, NI and INL on chromosome 12.

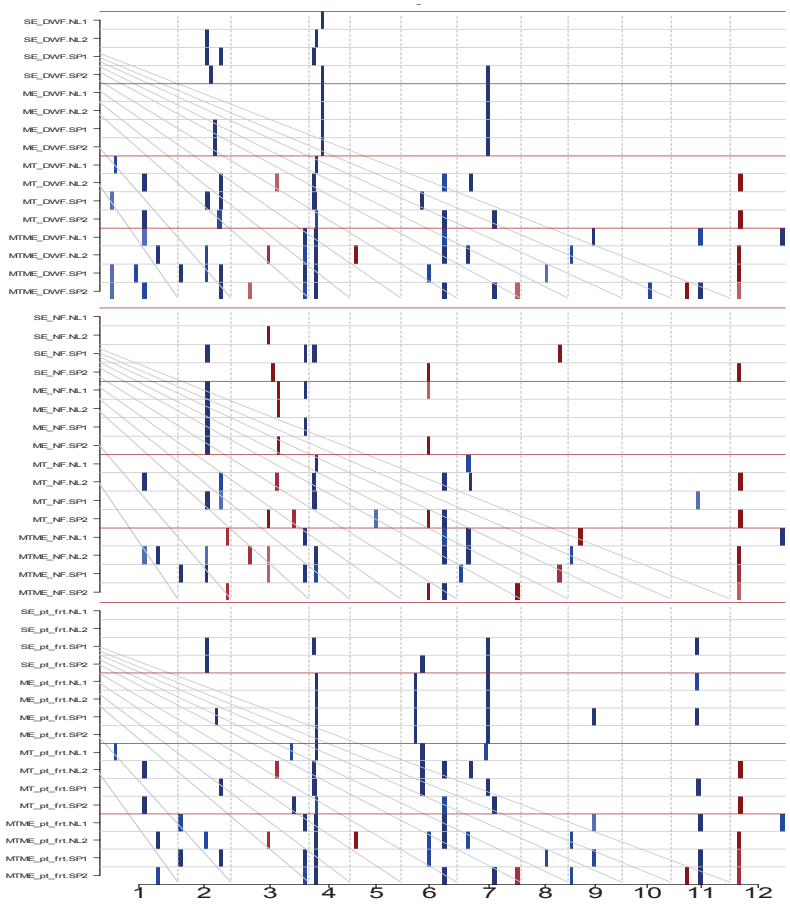


Figure 5.5 Comparing QTL positions from SE, ME, MT and MTME analyses for yield related traits (DWF, NF & pt_frt) across the four environments.
Blue indicates QTLs with significant effect from YW allele while red indicates QTLs with significant effect from CM334 allele. QTLs detected in SE, ME and MT analyses were present in QTLs from MTME with additional QTLs only picked up in MTME

Table 5.3 Detected QTLs and Their Effects for Trait-Environment Combinations from MTME analysis, showing results for chromosomes 1 and 2.
Results for other chromosomes are in Appendix 3E. Negative QTL effects mean that the YW allele gives higher trait values than the CM334 allele, and positive QTL effects mean that the CM334 allele gives higher trait values.

Markers	Pos	Env	DWF	NF	pt_frt	DWL	DWS	DWV	LUE	LAI	pt_leaf	Axl	SL	NLE	NI	INL	SLA
SP887	1@3.4	NL1	-0.11	-0.06	-0.07	-0.12	0.02	-0.01	0.08	-0.11	-0.15	-0.04	0.01	-0.12	0.12	-0.04	-0.05
		NL2	-0.08	-0.03	-0.05	-0.19	-0.01	-0.08	-0.08	-0.17	-0.18	0.04	0.05	-0.04	0.05	0.05	0.01
		SP1	-0.10	0.05	-0.10	-0.07	0.14	0.07	0.05	-0.04	-0.10	-0.01	0.09	0.02	-0.03	-0.03	0.07
		SP2	0.05	0.11	-0.01	-0.02	0.14	0.09	0.18	0.05	-0.17	-0.08	0.18	0.00	0.16	-0.09	0.08
SP310	1@30.1	NL1	0.02	0.01	0.00	-0.14	0.21	0.12	0.21	-0.12	-0.37	0.33	0.28	0.27	-0.21	0.26	-0.08
		NL2	-0.02	0.00	0.03	-0.28	-0.07	-0.15	0.05	-0.35	-0.26	0.10	0.13	0.00	-0.19	0.08	-0.11
		SP1	-0.16	-0.13	-0.07	-0.22	-0.10	-0.16	0.01	-0.21	-0.05	0.21	-0.03	0.18	-0.29	0.04	-0.13
		SP2	-0.17	-0.11	-0.08	-0.28	-0.11	-0.19	0.11	-0.25	-0.09	0.23	0.11	0.09	-0.21	0.16	-0.13
Gpms_178	1@52.5	NL1	-0.09	-0.03	-0.09	0.08	-0.06	-0.02	0.04	0.23	0.20	-0.08	0.08	-0.28	0.05	0.10	0.20
		NL2	0.06	0.03	0.04	0.12	0.10	0.12	0.00	0.18	0.04	-0.04	0.04	-0.13	0.04	0.05	0.16
		SP1	0.05	-0.03	0.02	0.12	0.09	0.12	0.11	0.07	0.02	0.04	0.15	-0.11	0.07	0.12	-0.05
		SP2	0.05	-0.01	0.08	0.03	-0.05	-0.02	-0.08	0.08	0.01	-0.19	-0.01	-0.13	0.17	-0.11	0.16
SP789	1@88.9	NL1	-0.02	0.01	-0.01	-0.04	0.03	0.02	-0.02	-0.17	-0.06	-0.07	-0.06	0.15	-0.03	-0.24	-0.22
		NL2	-0.07	0.01	-0.06	-0.09	0.10	0.03	0.03	-0.07	-0.20	-0.12	0.07	0.04	0.03	-0.19	-0.09
		SP1	-0.18	-0.13	-0.13	-0.05	0.01	-0.02	-0.07	-0.11	0.07	-0.08	-0.04	0.14	-0.05	-0.18	-0.11
		SP2	-0.06	0.08	-0.10	0.01	0.11	0.08	-0.01	0.06	0.02	-0.05	0.13	0.21	-0.03	-0.25	0.08
SP580	1@112.1	NL1	-0.17	0.10	-0.08	-0.21	-0.08	-0.12	-0.20	-0.22	-0.21	-0.26	-0.17	-0.07	0.08	-0.29	-0.24
		NL2	-0.15	-0.17	-0.15	-0.11	-0.11	-0.11	-0.11	-0.14	0.02	-0.11	-0.06	-0.03	0.14	-0.12	-0.02
		SP1	-0.09	-0.07	-0.04	-0.15	-0.06	-0.07	-0.16	-0.05	0.00	-0.19	-0.15	-0.01	0.05	-0.20	0.12
		SP2	-0.18	-0.16	-0.15	-0.10	0.04	-0.01	-0.07	-0.06	0.03	-0.20	0.04	-0.22	0.32	-0.04	0.02
SP576	1@144.8	NL1	-0.15	-0.14	-0.14	-0.13	0.06	0.01	-0.02	-0.12	-0.09	0.02	-0.04	0.05	0.16	-0.04	-0.02
		NL2	-0.29	-0.25	-0.27	-0.09	0.06	0.02	-0.08	-0.04	-0.08	-0.14	0.09	0.04	0.07	-0.20	0.04
		SP1	-0.05	-0.01	-0.04	-0.04	0.06	-0.03	-0.01	0.04	-0.09	0.00	0.07	0.05	0.19	0.00	0.15
		SP2	-0.15	-0.08	-0.20	0.03	0.13	0.12	-0.04	0.15	0.11	-0.01	0.00	0.06	-0.02	-0.03	0.17
Gpms_37	2@1.7	NL1	-0.13	0.04	-0.19	0.06	0.11	0.10	0.13	-0.09	-0.05	0.13	0.14	-0.11	-0.01	0.19	-0.61
		NL2	-0.04	0.03	-0.12	0.09	0.15	0.15	0.14	0.02	0.02	0.18	0.14	-0.02	0.07	0.21	-0.15
		SP1	-0.26	-0.23	-0.24	0.04	0.15	0.14	-0.07	-0.05	0.19	0.11	0.11	-0.01	-0.10	0.11	-0.19
		SP2	-0.06	0.06	-0.09	0.04	0.06	0.04	0.20	-0.05	0.09	0.09	0.21	-0.04	0.00	0.13	-0.24
4293-2	2@31	NL1	0.05	0.01	0.07	0.02	0.01	0.02	-0.06	-0.03	0.06	0.06	-0.08	0.15	-0.22	-0.02	0.09
		NL2	-0.01	-0.07	0.04	-0.05	-0.14	-0.13	-0.10	-0.05	0.11	0.01	-0.07	-0.01	-0.15	0.04	0.03
		SP1	0.04	0.02	0.06	0.02	-0.09	-0.10	0.05	-0.01	0.05	0.03	0.00	-0.05	0.00	0.06	-0.12

		SP2	-0.09	-0.13	-0.07	-0.01	-0.09	-0.05	-0.21	-0.05	0.16	0.04	-0.16	0.04	-0.18	0.03	-0.07
SP412	2@71.9	NL1	-0.14	-0.08	-0.09	-0.11	-0.16	-0.16	-0.10	-0.13	0.11	0.06	-0.04	-0.06	-0.08	0.12	-0.20
		NL2	-0.20	-0.17	-0.19	0.00	0.00	-0.01	-0.02	-0.03	0.13	0.05	-0.03	0.01	-0.02	0.10	-0.06
		SP1	-0.18	-0.26	-0.10	-0.09	-0.11	-0.05	-0.18	-0.13	0.14	-0.05	-0.13	-0.03	-0.07	0.00	-0.10
		SP2	-0.13	-0.08	-0.14	0.07	0.06	0.06	-0.12	-0.04	0.20	-0.04	-0.09	-0.21	0.17	0.11	-0.21
SP474	2@108.4	NL1	-0.10	-0.17	-0.02	-0.23	0.01	-0.06	-0.14	-0.39	-0.15	0.19	0.15	0.32	0.15	0.06	-0.26
		NL2	-0.07	-0.01	0.06	-0.32	-0.24	-0.29	-0.19	-0.37	-0.17	-0.03	-0.05	0.13	-0.02	-0.21	-0.12
		SP1	-0.28	-0.15	-0.21	-0.19	-0.01	-0.06	-0.01	-0.22	0.06	-0.13	0.06	0.07	0.14	-0.19	-0.19
		SP2	-0.27	-0.14	-0.06	-0.43	-0.32	-0.39	-0.21	-0.41	-0.06	-0.11	-0.10	0.11	-0.03	-0.22	-0.11
14828-2	2@128.8	NL1	0.11	0.22	0.07	0.03	-0.02	0.00	0.26	0.08	-0.07	-0.12	-0.11	-0.07	-0.13	-0.10	0.00
		NL2	0.05	0.11	0.00	0.16	0.18	0.20	0.25	0.08	0.05	0.09	0.09	0.08	0.08	0.14	-0.10
		SP1	0.06	0.03	0.06	-0.07	-0.10	-0.09	0.02	-0.09	-0.06	-0.03	0.07	0.05	0.03	-0.05	-0.06
		SP2	0.15	0.19	0.13	-0.01	-0.05	-0.04	0.13	-0.09	-0.13	0.16	0.13	-0.02	0.00	0.18	-0.12

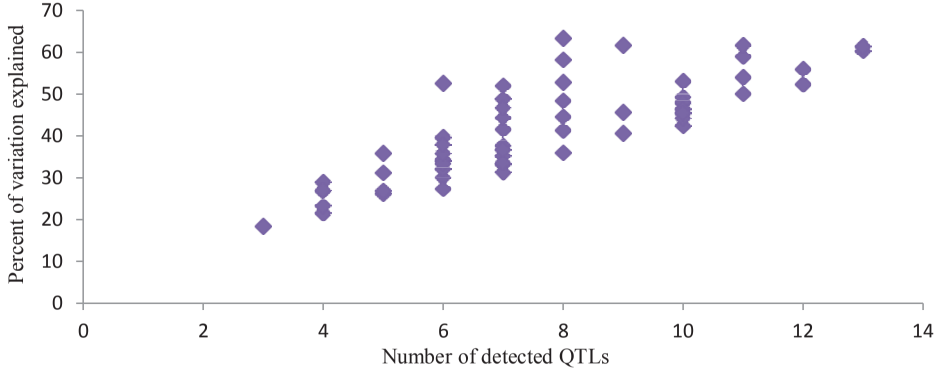


Figure 5.6 Joint distribution of total percent of variation attributable to significant QTLs from MTME
This revealed varying contributions of different QTLs to the total amount of variation explained. In general, the proportions of variation explained were positively correlated to the number of detected QTLs. However, some traits need a smaller number of QTLs to explain as much variation as other traits

Figure 3.6 shows the joint distribution of total percent of variation attributable to QTLs from the MTME model, which ranges from three QTLs explaining about 19% to 13 QTLs explaining 60%. This revealed varying contributions of different QTLs to the total amount of variation explained. In general, the proportions of variation explained were positively correlated to the number of detected QTLs. However, for some traits fewer QTLs explained similar percentages of variation as other traits with more QTLs. For example, eight QTLs for NLE.SP2 explained more variation (63.4%) than 13 QTLs for INL.SP2 (61.4) and DWL.SP1 (60.4%). This was consistent with the presence of a few QTLs with large effects for some traits and many QTLs of smaller effects for other traits. On average over the four environments, INL and NLE had the highest proportion of explained genetic variance (54% and 53%, respectively), this proportion was 46% for DWF while DWS and NF had the lowest proportions of 32% and 33%, respectively (Table 3.2).

Table 3.2 gives the number of QTLs together with their explained variance for each of the 15 traits in the four environments using ME, MT and MTME methods and also results from single trait single environment (SE) QTL analysis for comparison. As we used a different map in this study, the results for the SE analysis here was slightly different from those reported in Alimi et al. (2013). In principle, the QTL approach for SE is similar to other methods explained except that each trait in each environment was handled univariately. CIM was also used to account for multiple QTL. For each trait in each environment, there was a clear increase in the number of QTLs and explained variance going from ME to MT to MTME. There was also a clear gain in going from univariate analysis to multivariate analyses and in modelling correlations among environments and among traits within an environment. As an example, 1, two, four and seven QTLs were identified for DWF in the NL1 trial using SE, ME, MT and MTME methods respectively explaining about 18%, 22%, 24% and 32% of genetic variations respectively. Ten QTLs explaining 44% of the variance were detected for pt_frt in SP2 trials as against five (28%), three (36%) and three (26%) QTLs for MT, ME and SE respectively. The percentages explained variation by individual QTLs from ME, MT and MTME ranged from 3% to 35% (Figure 3.7). The MTME method yielded many QTLs of small effects (between 3% - 8%) that were not detected in both ME and MT methods. Also, MT and ME had more QTLs that explained 10% - 20% variation than MTME. This might be related to the “Beavis effect” (Beavis, 1994, 1997) as simpler models failed to detect some QTLs with small effects and also resulted in overestimation of some effect sizes.

Almost all QTLs detected in simpler methods were also detected in more complex methods. Using fruit-related traits for illustration (Figure 3.5), the three QTLs picked up for DWF in SP2 by SE method were also picked up by ME, MT and MTME methods. The positions of the three QTLs shifted slightly for MT and MTME as a result of their effects on other traits. The directions of their effects were also consistent. The QTL on chromosome 7 was significant in all environments under the ME method, but it disappeared for NL1, NL2 and SP1 trials using any of the other three methods. Many of the extra QTLs detected in MT were also detected in MTME. Similar patterns were observed for NF and pt_frt (Figure 3.5).

The prediction accuracies of the final QTL models for each trait under ME model were largely similar across environments, though prediction accuracies from SP trials were slightly higher in most cases (Table 3.4). Highest prediction accuracy for DWF under the ME model

(0.54) was obtained in SP environments. This agreed well with our earlier findings that the three QTLs found for DWF under the ME model explained far more variation in SP environments than in NL environments. This also indicated the presence of QEI for this trait. There was an improvement of trait predictions going from ME to MT and MTME models. The fitted QTL model from MTME predicted trait phenotypes better than MT and ME models. Prediction accuracies for DWF improved from about 0.54 under the ME model to about 0.7 under MT and 0.83 under MTME. Furthermore, the genetic correlations between predicted traits in each environment were similar to genetic correlations between BLUEs (appendix 3B).

Table 5.4 Predictive accuracy of final QTL models from ME, MT and MTME analyses.
Predictive accuracy is defined here in terms of correlation between BLUE and fitted phenotypic values. These values should be viewed as the upper limit of the predictive accuracies as they are not based on cross validation.

	NL1			NL2			SP1			SP2			Mean		
Trait	ME	MT	MTME	ME	MT	MTME	ME	MT	MTME	ME	MT	MTME	ME	MT	MTME
DWF	0.39	0.51	0.71	0.35	0.60	0.75	0.54	0.70	0.83	0.53	0.64	0.81	0.45	0.61	0.78
NF	0.38	0.36	0.63	0.35	0.54	0.72	0.48	0.57	0.76	0.51	0.63	0.78	0.43	0.53	0.72
pt_frt	0.51	0.50	0.67	0.42	0.60	0.72	0.73	0.71	0.80	0.52	0.56	0.76	0.55	0.59	0.74
DWL	0.65	0.63	0.80	0.64	0.71	0.73	0.63	0.65	0.79	0.71	0.71	0.79	0.66	0.68	0.78
DWS	0.35	0.54	0.65	0.45	0.55	0.62	0.49	0.64	0.74	0.40	0.51	0.67	0.42	0.56	0.67
DWV	0.32	0.53	0.66	0.43	0.58	0.62	0.37	0.60	0.72	0.41	0.57	0.70	0.38	0.57	0.68
LUE	0.47	0.54	0.75	0.45	0.51	0.63	0.40	0.53	0.77	0.51	0.65	0.77	0.46	0.56	0.73
LAI	0.59	0.62	0.75	0.69	0.73	0.78	0.65	0.64	0.76	0.70	0.75	0.81	0.66	0.69	0.78
pt_leaf	0.40	0.55	0.72	0.35	0.63	0.73	0.43	0.62	0.77	0.59	0.59	0.76	0.44	0.60	0.75
Axl	0.58	0.67	0.83	0.46	0.61	0.72	0.50	0.55	0.80	0.48	0.63	0.74	0.51	0.62	0.77
SL	0.50	0.64	0.75	0.51	0.65	0.72	0.54	0.60	0.80	0.59	0.68	0.81	0.54	0.64	0.77
NLE	0.61	0.66	0.81	0.47	0.64	0.68	0.55	0.68	0.77	0.65	0.69	0.77	0.57	0.67	0.76
NI	0.46	0.6	0.72	0.55	0.66	0.67	0.58	0.67	0.77	0.60	0.71	0.82	0.55	0.66	0.75
INL	0.61	0.66	0.83	0.45	0.64	0.76	0.67	0.71	0.88	0.39	0.51	0.72	0.53	0.63	0.80
SLA	0.34	0.32	0.55	0.43	0.45	0.65	0.64	0.65	0.80	0.66	0.67	0.85	0.52	0.52	0.71
Mean	0.48	0.56	0.72	0.47	0.61	0.70	0.55	0.63	0.78	0.55	0.63	0.77	0.51	0.61	0.75

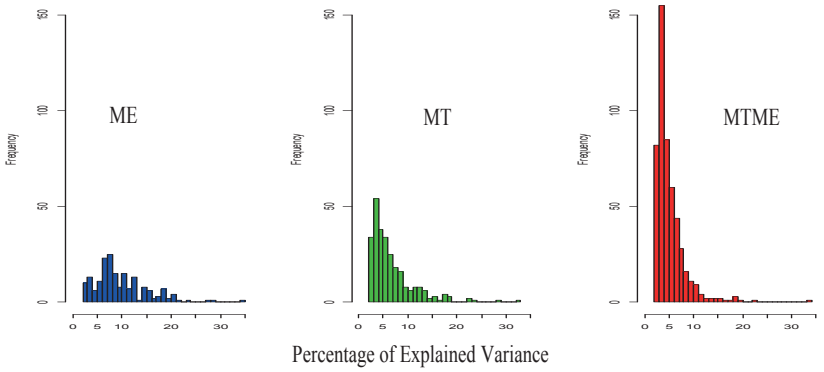


Figure 5.7 Histogram of Explained Variance by individual QTLs as detected by ME, MT and MTME analyses.
MTME produced far more QTLs than ME and MT but many of the extra QTLs from MTME are of small effects

3.5. Discussion

Several studies have shown that multi-trait and/or multi-environment QTL analyses based on linear mixed models are more powerful and effective to map pleiotropic QTL and QTL by environment interactions than performing single trait and single environment analyses (Boer et al., 2007; Korte et al., 2012; Malosetti et al., 2008; Sukhwinder et al., 2012). We also showed that in situations such as the EU-SPICY project (Barócsi, 2012; Nicolaï et al., 2012; Van der Heijden et al., 2012; Voorrips et al., 2010; www.spicyweb.eu), where phenotypic data on a large number of traits have been collected in multiple environments, using QTL methods that properly model underlying VCOV structures among the traits and between environments led to improved power to detect more QTLs than performing individual trait/environment analyses. The joint analysis was especially suitable for complex traits (such as yield) whose genetic variations are usually due to a large number of QTLs of smaller effects which might go undetected with single trait/environment analysis.

We performed and compared three mixed modelling approaches that modelled correlations between environments and/or among traits within an environment. In multi-environment studies, independent analyses without explicit modelling of the correlation structure between environments would not allow to identify GEI and QEI. In multi-traits datasets, univariate analysis that do not account for possible correlations among the traits would not allow us to properly identify QTLs with pleiotropic effects. The probability of finding QEI and/or pleiotropic QTLs is influenced by the magnitude of genetic correlations between environments and between traits within each environment respectively. It was expected that QTLs with identical effect directions will be detected for highly correlated traits while no common QTLs may be detected for non-correlated traits. Equally, high between-trial correlations would reduce the incidence of QEI. Pleiotropic QTLs that showed effects with trait increasing alleles from both parents are more likely to be detected for traits with negative correlations. The pepper traits considered showed positive and mostly uniform correlations between environments. This was also supported by the QEI results as most of the QEI observed were only due to differences in magnitude, and not different in direction. In our multi-trait analysis, synergistic pleiotropic QTLs were picked up for positively correlated traits. The pleiotropy was usually consistent across the four environments. Also, antagonistic pleiotropic QTLs were found for negatively correlated traits. These negative correlations depicted resource allocation competitions that exist between plant organs e.g. leaf and fruit related traits.

Factorial combinations of traits and environments and their joint analysis through the MTME method significantly increased the power of QTL detection with increased precision. This model fully utilizes covariance structures between environments and among traits within environments, and hence is better capable of mimicking biological process for complex traits than fitting ME and MT models separately. Considering yield, the results from SE and ME analyses showed that all the alleles increasing yield originated from the large fruited YW parental line. However, MT and MTME permitted to detect also favourable alleles from the small fruited parent CM334 on chromosomes 3, 5, 7, 11 and 12 (Figure 3.5). All those QTLs displayed pleiotropic effects with number of fruits (NF) and/or proportion of partitioning to fruit (pt_frt). The detection of these QTLs with MTME will permit to take it into account

when selecting recombinant individuals for high yield. This is more generally true since QTLs for vegetative traits were mainly restricted to chromosomes 1, 2 and 9, and to chromosomes 2, 3, 4 and 10 for fruit traits in the previous SE analyses (Alimi et al., 2012; Alimi et al., 2013a; Barchi et al., 2009; Ben Chaim et al., 2006; Rao et al., 2003). Since MTME model and also ME and MT models are based on mixed modelling technique, they are capable of handling unbalanced data in situation where not all traits are measured in all environments.

However, it is not in all situations that an MTME model can be successfully fitted. In situations where linear dependencies exist among some traits in the combination, some of these traits might need to be removed or transformed before an MTME fit can be successful. As an example, the total plant biomass (DWP) was partitioned to fruit (pt_frt), leaf (pt_leaf) and stem (pt_stem) components. We had to remove DWP and one of the partitioned components (pt_stem) before we could successfully fit the MTME model. However, we decided to leave some of the dependent traits such as DWV, DWS and DWL in our model as their presence did not affect the success of the MTME model. Also, this problem is more of combinatorial issue than correlation. As an example, DWF and pt_frt in our model are very correlated (about 0.9). When traits are very correlated, the method can still be successful unless the number of combinations to be handled are big with some linear dependencies among the traits.

MTME models might also prove difficult to fit due to the increase in the number of parameters to be estimated in the REML step as a result of large number of TE combinations. This becomes more laborious if markers (genetic predictors) are specified in the model as cofactors. If the problem occurs after adding cofactors, the result from the simple interval mapping could be subjected to backward selection before applying the final QTL model using appropriate QTL window size to separate the QTL positions. With a large QTL window size, some putative QTLs are lost while a small QTL window size could lead to declaration of duplicate QTLs. Duplicate QTLs could be detected via careful visual inspections of the QTL effect signs. If the signs of two neighbouring QTLs remain unchanged over all the traits, the QTLs can be regarded as one. For example, consider four traits T1, T2, T3 and T4 being influenced by three QTLs Q1, Q2 and Q3 that are very close to each other on a chromosome. If the effects of the three QTLs on the four traits follow these sequences: Q1 = {+,+,-,+}, Q2 = {+,+,-,+} and Q3 = {+,+,-,-}. Then Q1 and Q2 could be regarded as one QTL since the patterns are identical while Q3 is a different QTL from Q1 and Q2 because of the change in effect sign on T4. Furthermore, the appropriate QTL window size can be analytically checked using the Weller and Soller (2004) approach. In our case, the appropriateness of a 20cM peak window size was confirmed by analytically calculating the required confidence intervals for QTL location for a RIL population of our size given the magnitude of QTL effects (Weller and Soller, 2004). For the standardized traits, this was found to be around 15cM assuming (standardized) effect size of 0.25 with sample size of 149 and heritability of 0.25. It should be noted that this calculation was for univariate analysis with no multivariate correction. The actual interval in the multivariate case would even be smaller. So taking the smallest interval across all traits and environments can be seen as the upper bound of the interval in the multivariate sense. In our case the effects from many of the detected QTLs were more than 0.25 with the highest being around 0.6.This means that 15cM is like the upper bound for the interval.

In this study, we successfully applied the MTME approach to a dataset of 60 TE combinations. A simple approximating approach would have been to first apply data reduction techniques such as principal component analysis to reduce the number of variables and then perform a QTL analysis on the new set of variables, the principal component scores, just to identify the major genomic regions where DNA variation affects trait variation. We explored this approach – taking the scores of the first 10 principal components as trait values, and found that it produced most of the important QTLs underlying the original variables. That is, 16 QTLs were detected with high correspondence to significant QTLs from SE, ME, MT and MTME analyses (Figure 3.8). A major drawback of the use of principal components is the biological interpretation of the results, but as method to identify the most interesting genomic regions, it performs well.

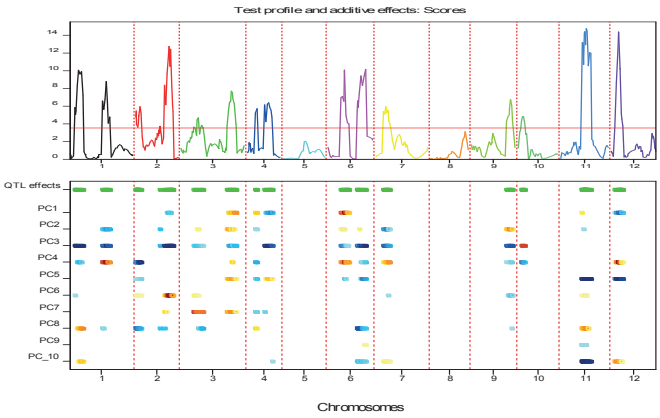


Figure 5.8 CIM Profile plot of the Multi-Trait analyses for scores from 10 PCs.
The top section shows the P-values of tests for QTL main effects. The bottom section shows heat maps along the genome for each PC, where blue means that the YW allele had a significant positive effect and red means that the CM334 allele had a significant positive effect on the given PC.

The QTL identified in this study will be aligned with eQTL results from a gene expression study in the same EU-SPICY project, (M. Vuylsteke, personal communication). The eQTL results will provide a set of candidate genes co-located with the QTL for yield and, hence, being likely involved in growth of pepper. Identifying these candidate genes would increase insight into the functioning of the pepper plant, and also increase efficiency of breeding, since this allows multiple alleles to be found within the gene, accounting for different phenotypes. Successful candidate genes, whose sequence position is related to QTL position, will be used to assess the marker-phenotype association in a core collection of pepper accessions (Nicolaï et al., 2012). Such an association genetics approach will be helpful in further selection of candidate genes, and will provide us with potential allelic values for phenotype prediction.

In conclusion, multivariate QTL mapping methods such as the MTME approach are instrumental to boost the power and accuracy of QTL detection for complex traits by successful identification of QTLs with relatively small effects. It would also lead to better detection of alleles in repulsion phase, differential allele expression according to environments and an increased explained variance for most complex traits. This would lead to improvement in the prediction of phenotype by the genotype and thus the genetic gain in

genome assisted breeding. This will ultimately increase our understanding of complex traits and our ability to use QTL in genome-assisted breeding.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211347.

We thank the EU-SPICY Industrial Advisory Board for support and discussions. Rik van Wijk and Syngenta are especially acknowledged for their highly valuable help in making available additional SNP markers that strongly improved the quality of the genetic map. Roeland Voorrips and other members of the EU-SPICY project are acknowledged for their contributions and helpful comments. We also thank Paul Keizer, Marcos Malosetti and Martin Boer of Biometris for their valuable insights.

Appendix 3A: Traits Description

Table 3A1 Description of abbreviations used in the manuscript

Abbreviation	Description
VCOV	Variance-Covariance
REML	REstricted Maximum Likelihood
QTL(s)	Quantitative Trait Locus (Loci)
BIC	Bayesian Information Criterion
SE/STSE	Single Trait Single Environment
ME	Multi - Environment
MT	Multi - Trait
MTME	Multi - Trait Multi - Environment
SIM	Simple Interval Mapping
CIM	Composite Interval Mapping
QEI	QTL by Environment Interaction
GEI	Genotype by Environment Interaction
BLUE	Best Linear Unbiased Estimation
RIL(s)	Recombinant Inbred Line(s)
TE	Trait-Environment
eQTL	Expression QTL
YW	Yolo Wonder parental line
CM334	Criollo de Morelos 334 parental line
GEI	Genotype by Environment
NL1	Netherlands phenotypic experiment in Spring
NL2	Netherlands phenotypic experiment in Autumn
SP1	Spain phenotypic experiment in Spring
SP2	Spain phenotypic experiment in Autumn

Table 3A2 Trait genetic correlations between environments.

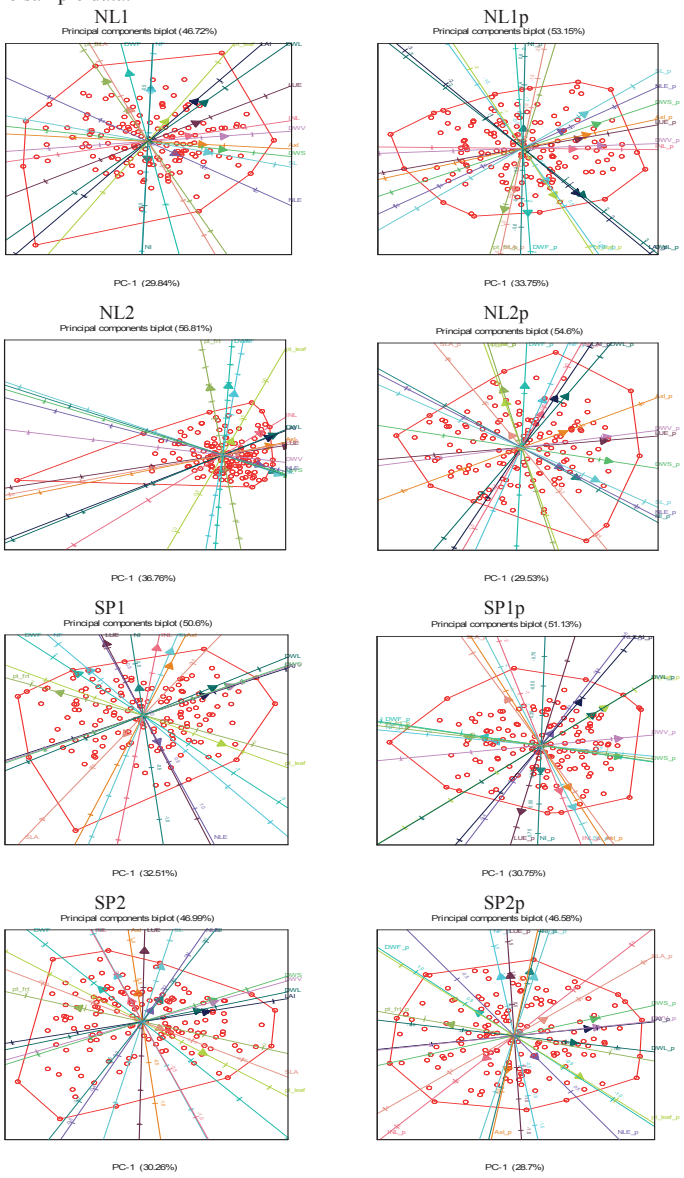
	NL1.NL2	NL1.SP1	NL1.SP2	NL2.SP1	NL2.SP2	SP1.SP2
DWF	0.72	0.60	0.61	0.53	0.62	0.58
NF	0.70	0.55	0.54	0.49	0.65	0.41
pt_frt	0.69	0.65	0.67	0.54	0.72	0.57
DWL	0.74	0.75	0.67	0.61	0.69	0.69
DWS	0.67	0.67	0.56	0.51	0.57	0.53
DWV	0.68	0.64	0.54	0.46	0.55	0.50
LUE	0.64	0.45	0.60	0.34	0.64	0.36
LAI	0.73	0.76	0.70	0.67	0.75	0.79
pt_leaf	0.68	0.61	0.55	0.49	0.54	0.66
Axl	0.74	0.69	0.54	0.74	0.66	0.61
SL	0.78	0.67	0.66	0.60	0.84	0.48
NLE	0.86	0.82	0.69	0.81	0.78	0.67
NI	0.76	0.51	0.65	0.30	0.69	0.40
INL	0.68	0.64	0.33	0.71	0.52	0.48
SLA	0.64	0.52	0.55	0.53	0.62	0.53

Table 3A3 Trait Variances in each environment

Trait	NL1	NL2	SP1	SP2
DWF	281.13	40.96	1345.35	1073.79
NF	49.67	21.52	121.84	130.76
pt_frt	1.4	0.6	1.9	1.3
DWL	60.06	38.74	260.63	145.95
DWS	420.87	196.97	802.33	482.03
DWV	659.18	323.96	1794.20	953.33
LUE	0.06	0.04	0.02	0.02
LAI	0.49	1.01	0.24	0.19
pt_leaf	0.2	0.3	0.2	0.1
Axl	42.82	56.53	23.40	20.98
SL	142.35	552.94	64.88	149.99
NLE	4.24	2.99	3.51	2.01
NI	0.64	5.32	1.46	1.45
INL	0.19	0.36	0.16	0.18
SLA	6.10	9.83	2.99	5.90

Appendix 3B: Biplots for BLUEs and fitted trait values

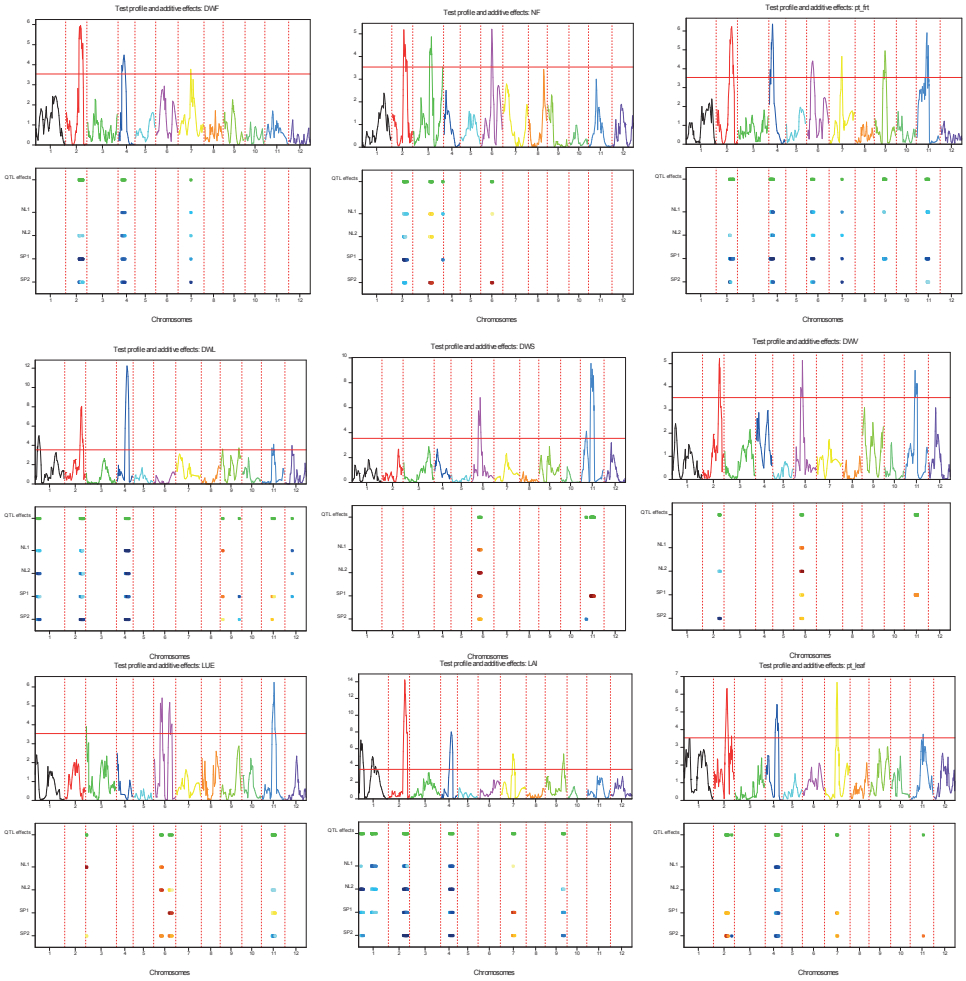
Figure 3B1 Biplots for BLUEs and fitted trait values in each environment.
NL1, NL2, SP1 and SP2 are the biplots of BLUE for the traits in each environment while NL1p, NL2p, SP1p and SP2p are the biplots for fitted values of each trait in each environment from the MTME QTL model. The cosine of the angle between the lines approximates the correlation between the traits they represent. The closer the angles are, the higher the correlations. Angles close to 90 or 270 degrees reflect weaker correlations. In each environment, angles between traits are similar for biplots from BLUEs and fitted values. E.g. the biplot for NL1 and NL1p, show a strong relationship between DWF and NF, and a weak relationship between DWF and NLE. The lines enclosing the sample points in the biplots are known as convex hulls, representing the smallest convex set of the sample data.



Appendix 3C: QTL by Environment Results from the ME analyses

Figure 3C1 CIM Profile plot for all the traits in the Multi-Environment analyses.

The top section shows the *P*-values of tests for QTL main effects. The bottom section shows heat maps along the genome for each environment, where blue means that the YW allele had a significant positive effect and red means that the CM334 allele had a significant positive effect in that environment. Many of the QTLs are constitutive i.e. consistent across environments with no crossover interaction except the QTL on chromosome 11 for LUE, Axl, SL and INL



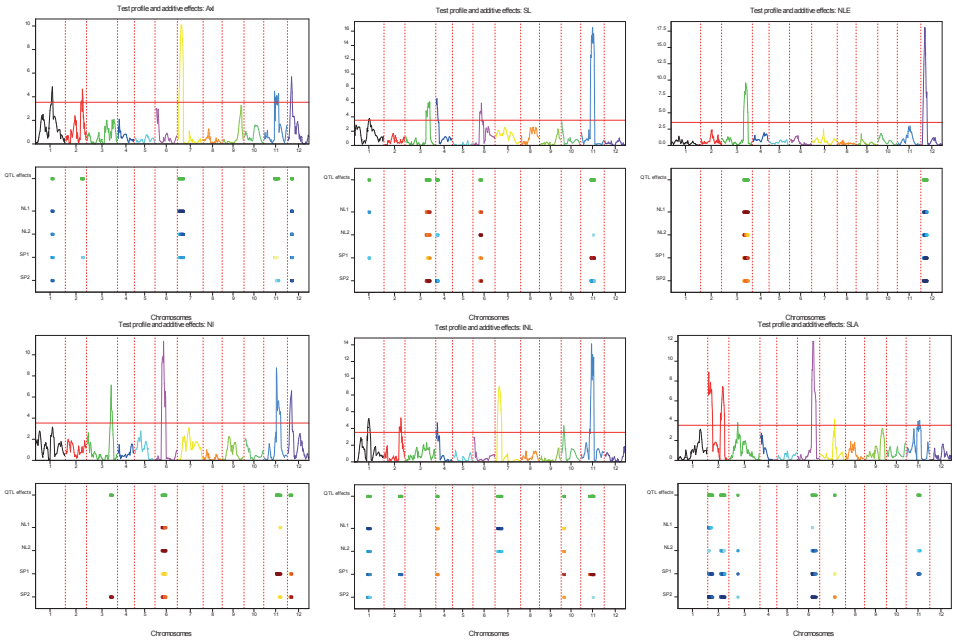


Table 3C1 Environment-specific QTL-effects for all traits in ME analysis
Negative QTL effects mean that the YW allele gives higher trait values than the CM334 allele, and positive QTL effects mean that the CM334 allele gives higher trait values. The underlined values are significant QTL effects from CIM. QEI refers to presence or absence of QTL by environment interaction. Many of the QEI are only in magnitude i.e. quantitative QEI.

Trait	Marker	Pos	NL1	NL2	SP1	SP2	QEI
DWF	C2P93	2@93.4	-0.06	-0.15	-0.42	-0.25	YES
DWF	C4P35	4@34.6	-0.40	-0.40	-0.40	-0.40	NO
DWF	22315	7@78.6	-0.25	-0.25	-0.25	-0.25	NO
NF	SP45	2@75.6	-0.28	-0.28	-0.28	-0.28	NO
NF	HpmsE008	3@119.9	0.29	0.25	0.05	0.36	YES
NF	SP451	3@195.8	-0.23	-0.06	-0.28	0.00	YES
NF	SP745	6@69.4	0.19	0.12	0.09	0.36	YES
pt_frt	C2P97	2@97.1	0.00	-0.08	-0.32	-0.05	YES
pt_frt	9745	4@20	-0.34	-0.34	-0.34	-0.34	NO
pt_frt	C6P35	6@35.5	-0.45	-0.45	-0.45	-0.45	NO
pt_frt	22315	7@78.6	-0.29	-0.29	-0.29	-0.29	NO
pt_frt	SP677	9@65.4	-0.15	0.00	-0.22	0.05	YES
pt_frt	CDKE	11@65	-0.18	-0.06	-0.31	-0.16	YES
DWL	C1P26	1@26	-0.29	-0.29	-0.29	-0.29	NO
DWL	SP474	2@108.4	-0.30	-0.30	-0.30	-0.30	NO
DWL	C4P66	4@65.9	-0.46	-0.46	-0.46	-0.46	NO
DWL	SP890	9@11.6	0.20	0.20	0.20	0.20	NO
DWL	C9P121	9@121.1	-0.03	-0.09	-0.28	-0.17	YES
DWL	SP170	11@76.8	-0.10	-0.09	0.15	0.09	YES
DWL	SP518	12@42.5	-0.27	-0.22	-0.21	-0.06	YES
DWS	SP863	6@51.7	0.35	0.35	0.35	0.35	NO
DWS	C11P35	11@35.3	0.04	-0.11	-0.06	-0.33	YES
DWS	CDKE	11@65	0.08	-0.06	0.42	0.19	YES
DWV	SP595	2@105.8	-0.10	-0.17	-0.09	-0.39	YES
DWV	SP863	6@51.7	0.28	0.28	0.28	0.28	NO
DWV	CDKE	11@65	0.07	-0.07	0.32	0.14	YES

LUE	SP119	3@0	0.44	0.09	0.06	0.15	YES
LUE	SP863	6@51.7	0.36	0.32	0.03	0.24	YES
LUE	C6P103	6@102.5	0.30	0.30	0.30	0.30	NO
LUE	SP935	11@79.3	0.05	-0.16	*0.18*	-0.25	YES
LAI	16310	1@15	-0.22	-0.40	-0.28	-0.18	YES
LAI	C1P93	1@92.7	-0.36	-0.17	-0.20	0.03	YES
LAI	C2P103	2@103.3	-0.43	-0.43	-0.43	-0.43	NO
LAI	C4P66	4@65.9	-0.36	-0.36	-0.36	-0.36	NO
LAI	22315	7@78.6	0.15	0.02	0.30	0.19	YES
LAI	19369	9@118	0.00	-0.15	-0.26	-0.21	YES
pt_leaf	C2P86	2@86.4	0.03	-0.02	0.30	0.39	YES
pt_leaf	23714	2@116	-0.16	-0.11	-0.10	-0.39	YES
pt_leaf	C4P73	4@73	-0.36	-0.36	-0.36	-0.36	NO
pt_leaf	22315	7@78.6	-0.06	-0.08	0.27	0.25	YES
pt_leaf	10035	11@81.4	-0.01	0.05	0.10	0.28	YES
Axl	SP580	1@112.1	-0.26	-0.26	-0.26	-0.26	NO
Axl	SP474	2@108.4	0.09	0.01	-0.24	-0.05	YES
Axl	C7P20	7@19.8	-0.55	-0.34	-0.27	-0.04	YES
Axl	Epms_410	11@66.4	-0.05	-0.05	*0.17*	-0.16	YES
Axl	SP729	12@21.9	-0.28	-0.28	-0.28	-0.28	NO
SL	C1P100	1@100.2	-0.24	-0.01	-0.18	0.05	YES
SL	16929	3@160.8	0.28	0.28	0.28	0.28	NO
SL	E1-1	4@2.7	0.08	-0.08	0.07	-0.24	YES
SL	SP863	6@51.7	0.39	0.39	0.39	0.39	NO
SL	4123-2	11@74	0.05	-0.03	*0.44*	-0.14	YES
NLE	C3P156	3@156.2	0.54	0.30	0.40	0.27	YES
NLE	SP729	12@21.9	-0.45	-0.38	-0.40	-0.59	YES
NI	Epms_386	3@159.1	0.16	0.14	0.01	0.39	YES
NI	SP863	6@51.7	0.51	0.54	0.25	0.35	YES
NI	SP935	11@79.3	0.11	0.01	0.45	0.06	YES
NI	SP729	12@21.9	0.15	0.10	0.32	0.33	YES
INL	C1P96	1@96.5	-0.33	-0.33	-0.33	-0.33	NO
INL	SP474	2@108.4	0.02	-0.11	-0.30	-0.11	YES
INL	5589	4@5.2	0.27	0.04	0.29	0.10	YES
INL	C7P24	7@24.2	-0.50	-0.25	-0.10	0.11	YES
INL	SP118	10@13.1	0.26	0.26	0.26	0.26	NO
INL	Gpms_101	11@67	0.05	0.05	0.39	*-0.09*	YES
SLA	Gpms_37	2@1.7	-0.29	-0.29	-0.29	-0.29	NO
SLA	C2P93	2@93.4	-0.32	-0.32	-0.32	-0.32	NO
SLA	15343	3@53.8	-0.17	-0.17	-0.17	-0.17	NO
SLA	SP991	6@97.6	-0.35	-0.35	-0.35	-0.35	NO
SLA	SP147	7@93.8	-0.25	0.08	0.16	0.25	YES
SLA	5682	11@85.5	-0.01	-0.16	-0.24	0.09	YES

Appendix 3D: QTL Effects from the MT analyses

Table 3D1 Trait-specific QTL-effects for MT analysis

Negative QTL effects mean that the YW allele gives higher trait values than the CM334 allele, and positive QTL effects mean that the CM334 allele gives higher trait values. The underlined values are significant QTL effects. #Traits gives the number of traits influenced by a particular QTL.

Env	Markers	Pos	DWF	NF	pt_fr	DWL	DWS	DWV	LUE	LAI	pt_lf	Axl	SL	NLE	NI	INL	SLA	#Traits
NL1	22154	1@39.5	-0.19	-0.06	-0.18	-0.20	0.20	0.10	0.25	-0.10	-0.32	0.17	0.33	-0.01	0.06	0.21	-0.02	9
NL1	21850-1	1@104.0	-0.10	-0.03	-0.05	-0.10	-0.05	-0.06	-0.27	-0.19	-0.15	-0.28	-0.25	0.01	0.20	-0.37	-0.30	7
NL1	13070-2	2@37.3	0.06	0.13	0.06	0.10	-0.06	-0.02	-0.06	-0.05	0.17	0.07	-0.03	0.13	-0.20	0.05	-0.23	2
NL1	SP474	2@108.4	-0.09	-0.09	-0.02	-0.21	-0.05	-0.09	-0.07	-0.39	-0.12	0.07	0.01	0.14	0.00	0.01	-0.29	4
NL1	SP58	3@40.1	-0.04	0.04	-0.03	0.06	0.06	0.06	0.31	0.05	0.14	0.02	0.16	0.19	-0.01	-0.10	0.04	2
NL1	SP447	3@153.4	-0.12	0.08	-0.19	0.12	0.27	0.26	0.21	0.11	0.06	0.26	0.29	0.44	0.17	-0.01	-0.15	8
NL1	9745	4@20	-0.32	-0.23	-0.31	0.00	0.15	0.12	-0.02	0.14	0.01	0.14	0.11	-0.15	-0.09	0.29	-0.02	5
NL1	19586	4@70.1	-0.11	-0.08	0.03	-0.51	-0.21	-0.32	0.06	-0.40	-0.45	-0.05	-0.05	0.05	0.11	-0.14	0.05	6
NL1	SP863	6@51.7	-0.15	0.05	-0.24	0.12	0.38	0.33	0.38	0.15	-0.02	0.10	0.39	0.07	0.43	0.08	0.02	6
NL1	SP648	7@28.7	-0.13	-0.21	-0.04	-0.20	-0.11	-0.15	-0.11	-0.15	0.00	-0.42	-0.17	-0.10	0.23	-0.37	0.30	8
NL1	SP652	7@72.8	-0.16	0.02	-0.18	0.05	0.03	0.05	-0.08	0.13	-0.01	0.07	0.07	0.16	-0.04	-0.12	-0.28	2
NL1	2249	10@7	-0.07	-0.12	-0.07	0.12	-0.05	0.00	-0.30	0.14	0.28	0.12	-0.25	-0.10	-0.21	0.18	-0.07	7
NL1	SP724	12@20.2	0.08	0.16	0.10	-0.11	-0.16	-0.17	-0.06	-0.20	-0.14	-0.29	-0.09	-0.41	0.09	-0.05	-0.10	3
NL2	SP310	1@30.1	0.01	0.00	0.06	-0.27	0.00	-0.10	0.04	-0.30	-0.32	0.06	0.20	-0.12	-0.12	0.10	-0.02	5
NL2	SP580	1@112.1	-0.27	-0.24	-0.25	-0.14	-0.04	-0.08	-0.15	-0.12	-0.06	-0.20	-0.03	0.00	0.16	-0.25	-0.03	7
NL2	Gpms_37	2@1.6	-0.08	-0.07	-0.12	0.10	0.05	0.08	0.04	0.01	0.16	0.16	0.06	-0.05	-0.03	0.22	-0.16	3
NL2	SP127	2@58.4	-0.21	-0.19	-0.17	-0.05	-0.10	-0.10	-0.17	-0.01	0.16	-0.02	-0.16	-0.03	-0.09	0.03	-0.03	5
NL2	SP474	2@108.4	-0.12	-0.01	-0.03	-0.26	-0.10	-0.16	-0.04	-0.36	-0.11	0.02	0.05	0.16	0.05	-0.08	-0.10	7
NL2	SP496	3@116.7	0.17	0.18	0.21	0.06	-0.01	0.02	0.12	-0.02	0.03	0.20	0.16	0.14	0.03	0.17	-0.14	7
NL2	Epms_386	3@159.1	-0.06	0.02	-0.13	-0.02	0.24	0.18	0.07	-0.05	-0.28	-0.02	0.20	0.22	0.13	-0.29	0.03	6
NL2	SP614	4@11.1	-0.29	-0.23	-0.23	-0.15	0.04	-0.02	-0.08	-0.07	-0.04	0.06	-0.03	0.02	-0.05	0.11	0.15	7
NL2	19586	4@70.1	-0.10	-0.02	-0.02	-0.39	-0.10	-0.22	-0.03	-0.40	-0.31	-0.02	0.06	0.16	0.12	-0.13	0.01	5
NL2	SP510	5@114.8	0.06	0.08	0.11	-0.13	-0.06	-0.09	0.02	-0.16	-0.13	0.02	0.03	0.17	0.15	-0.10	-0.02	1
NL2	SP863	6@51.7	-0.15	-0.02	-0.26	0.12	0.42	0.37	0.35	0.13	-0.16	0.05	0.42	0.06	0.48	0.03	0.06	7
NL2	SP737	6@107.5	-0.17	-0.21	-0.23	-0.02	0.13	0.09	0.11	-0.10	-0.10	0.03	0.13	-0.12	-0.05	0.12	-0.29	6
NL2	TC12620	7@34.6	-0.25	-0.24	-0.25	-0.19	-0.04	-0.10	-0.11	-0.15	-0.10	-0.29	0.00	-0.16	0.11	-0.21	0.02	8
NL2	SP680	9@124.2	0.01	-0.06	-0.01	-0.13	0.02	-0.03	0.08	-0.17	-0.19	0.25	0.14	0.10	0.06	0.17	-0.20	6
NL2	SP968	10@13.1	0.02	-0.07	0.04	0.13	-0.06	0.00	-0.07	0.07	0.15	0.05	-0.24	-0.17	-0.15	0.16	-0.09	5
NL2	SP729	12@21.9	0.21	0.23	0.25	-0.06	-0.05	-0.06	0.04	-0.07	-0.03	-0.20	-0.08	-0.35	0.05	0.13	0.04	5
NL2	KS09097F12-2	12@43.8	-0.05	-0.01	-0.04	-0.16	-0.07	-0.11	0.00	-0.13	-0.10	-0.09	0.06	0.02	0.11	-0.10	0.08	2
SP1	SP310	1@30.1	-0.16	-0.11	-0.08	-0.21	0.00	-0.08	0.06	-0.18	-0.10	0.20	0.04	0.04	-0.23	0.14	-0.08	4
SP1	23685	2@21.3	-0.07	-0.12	-0.04	0.06	-0.11	-0.08	0.02	-0.10	0.23	0.12	-0.02	-0.12	-0.07	0.19	-0.38	5
SP1	SP45	2@75.6	-0.25	-0.29	-0.15	-0.07	-0.01	0.00	-0.23	-0.09	0.11	0.03	-0.07	0.11	-0.04	-0.05	-0.03	5
SP1	SP474	2@108.4	-0.28	-0.16	-0.21	-0.24	-0.03	-0.07	-0.02	-0.28	0.02	-0.22	0.05	0.02	0.09	-0.22	-0.19	9
SP1	SP94	3@147.3	-0.04	0.07	-0.04	-0.10	0.01	-0.09	0.11	-0.09	-0.07	0.03	0.18	0.29	0.06	-0.18	-0.08	4
SP1	DHV41	4@15.8	-0.30	-0.25	-0.33	0.13	0.29	0.30	-0.12	0.09	0.23	0.20	0.08	-0.09	0.00	0.30	-0.08	11
SP1	19586	4@70.1	-0.06	-0.03	-0.01	-0.32	-0.10	-0.20	0.10	-0.27	-0.26	-0.06	-0.08	0.11	0.08	-0.14	0.04	7

SP1	SP863	6@51.7	-0.20	-0.12	-0.25	0.04	0.30	0.20	0.06	0.10	0.02	0.00	0.28	0.04	0.23	-0.05	0.03	7
SP1	8536	6@91	-0.05	0.06	-0.08	0.03	0.04	0.02	0.21	-0.11	0.07	0.01	0.11	0.03	-0.06	0.01	-0.39	5
SP1	SP644	6@138	0.00	-0.03	0.02	-0.17	0.00	-0.04	0.17	-0.14	-0.21	0.09	0.01	0.09	-0.16	0.06	-0.01	5
SP1	22315	7@78.6	-0.15	-0.05	-0.24	0.20	0.25	0.25	-0.14	0.28	0.24	0.03	0.05	0.19	-0.21	-0.15	0.12	12
SP1	SP565	8@63.6	-0.14	-0.02	-0.12	-0.04	0.11	0.07	-0.08	0.03	-0.03	0.04	0.19	-0.13	0.05	0.13	0.09	6
SP1	SP680	9@124.2	0.07	-0.10	0.09	-0.24	-0.13	-0.20	0.28	-0.24	-0.27	0.25	0.18	0.07	-0.05	0.20	-0.13	10
SP1	HpmsE013	10@23.3	0.12	0.09	0.09	0.10	-0.12	-0.04	0.02	0.02	0.12	0.04	-0.13	-0.11	-0.05	0.15	-0.11	5
SP1	Gpms_101	11@67	-0.12	-0.17	-0.25	0.14	0.33	0.24	0.22	0.06	0.06	0.15	0.42	-0.25	0.43	0.35	-0.23	13
SP1	SP729	12@21.9	0.08	0.11	0.12	-0.13	-0.15	-0.15	0.01	-0.11	-0.13	-0.28	0.05	-0.42	0.35	0.07	0.09	7
SP2	SP580	1@112.1	-0.28	-0.11	-0.27	-0.04	0.14	0.10	-0.14	0.04	0.15	-0.27	0.08	-0.15	0.33	-0.18	0.09	8
SP2	SP225	2@16.3	-0.15	-0.11	-0.12	0.01	-0.04	-0.03	-0.03	-0.09	0.22	0.18	0.07	-0.03	-0.17	0.26	-0.25	6
SP2	SP127	2@58.4	-0.07	-0.06	-0.06	0.15	0.00	0.05	-0.31	0.03	0.28	-0.20	-0.26	-0.15	-0.03	-0.10	-0.16	7
SP2	SP595	2@105.8	-0.20	0.03	0.02	-0.47	-0.34	-0.43	-0.12	-0.54	-0.18	-0.08	-0.04	-0.01	0.09	-0.10	-0.27	8
SP2	8211	3@94.7	0.05	0.28	0.16	-0.15	-0.23	-0.21	-0.04	-0.14	-0.08	-0.08	0.04	0.05	0.00	-0.14	-0.04	8
SP2	Epms_386	3@159.1	-0.13	0.18	-0.22	0.03	0.22	0.18	0.19	0.14	0.03	0.20	0.37	0.27	0.35	-0.03	0.11	9
SP2	SP299	4@8.8	-0.20	-0.07	-0.09	-0.14	-0.01	-0.07	-0.25	-0.10	0.00	0.05	-0.21	0.01	-0.11	0.05	0.09	4
SP2	19586	4@70.1	-0.15	-0.03	-0.02	-0.43	-0.14	-0.27	0.03	-0.36	-0.31	0.02	-0.02	0.16	0.14	-0.12	-0.03	9
SP2	SP535	5@65.3	-0.08	-0.16	-0.02	-0.19	-0.04	-0.11	-0.14	-0.09	-0.14	-0.07	-0.07	-0.01	-0.16	-0.07	0.18	7
SP2	SP745	6@69.4	-0.03	0.34	-0.10	0.00	0.15	0.10	0.35	0.04	-0.07	-0.02	0.34	0.03	0.23	-0.05	-0.04	5
SP2	SP737	6@107.5	-0.26	-0.28	-0.20	-0.13	-0.01	-0.05	0.09	-0.20	0.09	0.08	0.13	-0.02	-0.02	0.14	-0.29	6
SP2	SP147	7@93.8	-0.29	-0.12	-0.29	-0.09	0.11	0.05	0.00	0.14	0.09	0.03	0.18	-0.01	0.10	0.03	0.30	5
SP2	19369	9@117.9	-0.09	0.01	-0.03	-0.21	-0.08	-0.13	0.13	-0.22	-0.12	0.21	0.17	0.03	0.05	0.20	-0.13	7
SP2	5683	11@85.8	-0.11	0.06	-0.10	0.08	-0.01	0.03	-0.13	0.09	0.22	-0.09	-0.13	-0.03	0.05	-0.03	0.03	5
SP2	SP729	12@21.9	0.21	0.26	0.24	-0.09	-0.12	-0.13	-0.01	-0.13	-0.26	-0.32	-0.15	-0.57	0.25	0.14	-0.01	8

Appendix 3E: QTL Effects from the MTME analyses: Chromosomes 3-12

Table 3E1 Detected QTLs and Their Effects for Trait-Environment Combinations from MTME analysis: Chromosomes 3, 4, 5 and 6

Markers	Pos	Env	DWF	NF	pt_frt	DWL	DWS	DWV	LUE	LAI	pt_lf	Axl	SL	NLE	NI	INL	SLA
ISO-2	3@19.8	NL1	0.07	-0.02	0.09	0.00	0.04	0.04	0.24	0.00	-0.04	0.03	0.06	0.10	0.13	-0.01	0.09
		NL2	0.09	0.04	0.06	-0.07	-0.08	-0.08	-0.10	-0.02	-0.06	-0.08	-0.05	-0.02	-0.05	-0.11	-0.01
		SP1	0.04	0.07	0.00	-0.01	-0.03	-0.05	0.00	0.03	0.03	-0.13	-0.06	-0.11	0.15	-0.04	0.02
		SP2	-0.01	-0.03	-0.01	-0.04	-0.08	-0.07	-0.03	-0.02	0.02	-0.18	0.02	0.02	0.12	-0.21	0.02
SP609	3@49.9	NL1	0.10	0.06	0.09	-0.10	-0.12	-0.13	0.09	-0.16	-0.14	-0.09	0.02	0.14	-0.02	-0.20	-0.22
		NL2	0.11	0.18	0.07	0.07	0.02	0.04	0.17	-0.05	0.02	0.04	0.07	0.10	0.13	0.01	-0.16
		SP1	-0.06	-0.06	0.00	-0.14	-0.11	-0.12	0.10	-0.19	-0.07	0.06	0.05	0.26	-0.11	-0.13	-0.16
		SP2	0.17	0.15	0.16	-0.01	-0.12	-0.08	0.10	-0.15	-0.11	0.13	0.05	0.03	0.11	0.11	-0.20
8211	3@94.7	NL1	0.09	0.17	0.06	0.14	-0.03	0.02	0.12	0.13	0.20	0.08	0.10	-0.02	0.06	0.15	-0.07
		NL2	0.19	0.17	0.20	0.07	-0.07	-0.03	-0.08	0.15	-0.04	-0.10	-0.07	-0.13	-0.18	-0.10	-0.04
		SP1	0.06	0.17	0.07	0.08	-0.06	0.02	0.08	0.02	0.09	-0.02	0.01	-0.20	0.05	0.11	-0.16
		SP2	-0.04	0.09	0.03	0.01	-0.09	-0.06	-0.14	0.10	0.12	-0.14	0.04	-0.05	-0.03	-0.12	0.14
D0EVU	3@135.9	NL1	-0.12	0.07	-0.13	0.09	0.10	0.11	0.23	0.17	0.13	0.13	0.15	0.15	0.19	0.08	0.14
		NL2	-0.12	-0.04	-0.10	-0.01	0.05	0.02	0.01	-0.09	0.06	0.14	0.12	0.16	0.10	0.09	0.04
		SP1	-0.06	-0.03	-0.05	0.03	0.00	-0.05	0.07	-0.03	0.07	-0.06	0.01	0.10	0.05	-0.11	-0.14
		SP2	-0.12	0.11	-0.14	0.05	0.10	0.09	0.11	0.04	0.11	0.08	0.11	0.10	0.02	0.01	-0.02
Epms_386	3@159.1	NL1	0.11	0.15	0.01	0.15	0.29	0.28	0.08	0.07	-0.07	0.24	0.24	0.48	0.06	-0.07	-0.15
		NL2	0.11	0.13	0.05	0.01	0.12	0.11	0.05	-0.03	-0.24	-0.01	0.10	0.15	0.02	-0.25	-0.10
		SP1	0.03	0.16	-0.02	-0.04	0.09	0.07	0.03	0.03	-0.12	0.21	0.17	0.33	-0.02	-0.08	0.08
		SP2	-0.01	0.12	-0.08	0.00	0.11	0.08	0.15	0.12	-0.04	0.17	0.31	0.20	0.28	-0.02	0.12
Epms_402	3@187.5	NL1	-0.25	-0.27	-0.21	0.00	0.17	0.13	0.08	0.05	0.07	0.06	0.13	0.06	-0.10	0.02	0.07
		NL2	-0.19	-0.11	-0.14	-0.09	0.06	0.02	-0.03	-0.09	-0.10	-0.04	0.12	0.02	-0.07	-0.08	-0.01
		SP1	-0.25	-0.30	-0.23	0.06	0.22	0.12	-0.04	0.04	0.09	0.01	0.14	0.09	-0.06	-0.07	-0.03
		SP2	-0.19	-0.05	-0.18	0.01	0.14	0.10	-0.02	-0.02	0.08	0.07	-0.02	0.18	-0.09	-0.07	-0.06
14521-2	4@19.4	NL1	-0.31	-0.15	-0.29	-0.02	0.07	0.05	-0.09	0.18	0.00	0.16	0.17	-0.05	-0.17	0.21	-0.11
		NL2	-0.30	-0.24	-0.26	-0.10	0.09	0.03	-0.07	0.00	-0.04	0.11	-0.04	0.01	-0.08	0.15	0.11
		SP1	-0.26	-0.19	-0.29	0.14	0.25	0.29	-0.05	0.08	0.24	0.17	0.05	-0.06	-0.06	0.26	-0.16
		SP2	-0.24	-0.12	-0.25	-0.01	0.16	0.11	-0.19	0.05	0.12	0.09	-0.14	0.04	-0.13	0.09	0.12
4706-2	4@53.2	NL1	-0.01	-0.08	0.10	-0.34	-0.25	-0.30	-0.05	-0.40	-0.21	-0.20	-0.26	-0.16	0.10	-0.10	-0.02
		NL2	0.02	0.02	0.03	-0.16	-0.08	-0.12	0.03	-0.25	-0.08	-0.09	-0.02	0.09	0.08	-0.10	-0.02
		SP1	-0.11	-0.07	-0.02	-0.28	-0.15	-0.23	-0.06	-0.25	-0.15	-0.04	-0.07	-0.02	0.12	-0.01	0.03
		SP2	-0.07	-0.07	0.08	-0.30	-0.22	-0.27	-0.09	-0.33	-0.19	-0.06	-0.14	-0.04	0.13	-0.02	-0.11
SP618	5@13.9	NL1	0.09	-0.09	0.10	-0.19	-0.14	-0.17	-0.09	-0.10	-0.13	-0.09	0.05	-0.07	0.10	0.01	0.13
		NL2	0.24	0.14	0.25	-0.02	-0.05	-0.04	0.04	0.04	-0.04	-0.03	-0.08	-0.11	0.02	0.11	0.12
		SP1	0.08	0.06	0.13	-0.20	-0.11	-0.13	0.29	-0.21	-0.26	-0.01	-0.02	-0.23	0.06	0.18	-0.07
		SP2	0.07	-0.02	0.10	-0.10	-0.11	-0.12	-0.12	-0.14	-0.11	0.05	-0.16	-0.12	0.05	0.18	-0.01
SP242	5@38.6	NL1	0.03	0.17	-0.03	0.07	0.14	0.14	0.12	-0.03	-0.08	0.04	0.07	-0.01	0.27	0.06	-0.28

		NL2	-0.12	-0.04	-0.13	0.01	0.19	0.15	0.13	-0.10	-0.10	0.05	0.20	0.09	0.13	-0.03	-0.14
		SP1	-0.09	-0.07	-0.14	0.06	0.13	0.10	-0.08	0.05	0.06	0.11	0.20	0.10	0.00	0.02	-0.03
		SP2	0.01	0.00	0.02	0.02	0.06	0.05	0.20	-0.02	-0.06	0.01	0.13	0.06	0.09	-0.06	-0.10
SP749	5@65.3	NL1	-0.06	0.00	-0.03	-0.14	-0.15	-0.16	-0.16	0.00	-0.08	-0.01	-0.13	0.07	-0.35	-0.17	0.18
		NL2	0.00	-0.13	0.08	-0.12	-0.12	-0.13	-0.09	0.01	0.05	-0.01	-0.14	0.12	-0.07	-0.06	0.23
		SP1	-0.04	-0.15	-0.01	-0.13	-0.08	-0.10	-0.12	-0.04	-0.04	-0.10	-0.24	0.03	-0.24	-0.12	0.19
		SP2	-0.02	-0.14	0.01	-0.19	-0.05	-0.12	-0.21	-0.08	-0.18	-0.16	-0.13	0.01	-0.17	-0.18	0.24
18077	6@0	NL1	-0.15	-0.07	-0.14	-0.05	0.16	0.11	0.11	-0.09	-0.07	0.19	0.21	0.10	0.11	0.18	0.00
		NL2	-0.10	-0.03	-0.09	0.01	0.15	0.13	0.18	-0.03	-0.11	0.27	0.17	0.13	0.13	0.20	-0.10
		SP1	-0.07	-0.04	-0.06	-0.09	0.11	0.06	0.21	-0.08	-0.17	0.25	0.20	0.04	0.07	0.21	-0.09
		SP2	0.03	0.08	0.00	-0.05	0.02	0.00	0.15	-0.04	-0.12	0.05	0.06	0.15	0.02	-0.10	0.00
SP745	6@69.4	NL1	0.01	0.10	-0.06	0.12	0.25	0.23	0.26	0.13	0.04	0.11	0.33	0.03	0.23	0.13	-0.02
		NL2	-0.10	0.08	-0.19	0.02	0.31	0.25	0.27	0.07	-0.15	0.05	0.33	0.05	0.33	0.07	0.09
		SP1	-0.18	0.05	-0.17	0.01	0.13	0.09	0.08	-0.03	0.06	-0.02	0.23	0.04	0.13	-0.06	-0.15
		SP2	-0.01	0.38	-0.06	0.01	0.10	0.07	0.29	0.03	-0.04	-0.06	0.27	-0.05	0.20	-0.03	-0.06
SP737	6@107.5	NL1	-0.20	-0.20	-0.23	-0.09	0.10	0.05	0.10	-0.12	-0.12	0.07	0.17	0.00	0.06	0.12	-0.22
		NL2	-0.22	-0.24	-0.27	-0.02	0.14	0.10	0.12	-0.14	-0.13	0.10	0.15	-0.06	-0.03	0.13	-0.38
		SP1	-0.06	-0.06	-0.08	-0.06	0.14	0.05	0.27	-0.13	-0.12	0.09	0.20	-0.03	-0.03	0.17	-0.26
		SP2	-0.26	-0.25	-0.23	-0.06	0.08	0.04	0.17	-0.19	0.09	0.14	0.18	0.04	-0.01	0.14	-0.36

Table 3E2 Detected QTLs and Their Effects for Trait-Environment Combinations from MTME analysis: Chromosomes 7, 8, 9 and 10

Markers	Pos	Env	DWF	NF	pt_frt	DWL	DWS	DWV	LUE	LAI	pt_lf	Axl	SL	NLE	NI	INL	SLA
13163	7@0	NL1	-0.04	0.04	-0.10	-0.04	-0.02	-0.02	-0.10	-0.09	-0.12	-0.18	-0.12	-0.23	-0.29	-0.21	-0.12
		NL2	-0.06	-0.06	-0.06	-0.07	0.00	-0.02	-0.03	-0.03	-0.07	-0.02	0.07	-0.05	0.05	-0.01	0.08
		SP1	-0.13	-0.18	-0.15	0.02	0.08	0.07	-0.07	-0.04	0.08	-0.04	-0.09	-0.03	-0.05	-0.05	-0.08
		SP2	0.03	0.10	-0.04	0.04	0.11	0.07	0.13	0.03	-0.03	0.00	0.07	0.12	-0.04	-0.09	-0.07
SP648	7@28.7	NL1	-0.11	-0.22	0.00	-0.13	-0.10	-0.12	-0.09	-0.09	0.11	-0.27	-0.10	0.03	0.32	-0.18	0.24
		NL2	-0.21	-0.24	-0.19	-0.21	-0.12	-0.16	-0.14	-0.17	-0.06	-0.25	-0.12	-0.18	-0.04	-0.14	-0.07
		SP1	-0.04	-0.01	0.05	-0.14	-0.16	-0.12	-0.10	-0.11	-0.02	-0.19	-0.17	-0.18	0.03	0.00	0.05
		SP2	-0.07	-0.09	-0.05	-0.07	-0.04	-0.04	-0.05	-0.08	0.03	-0.04	-0.07	-0.24	0.10	0.18	-0.02
SP147	7@93.8	NL1	-0.10	0.04	-0.16	-0.02	0.14	0.10	0.03	0.06	-0.19	0.16	0.15	0.09	-0.05	0.03	-0.26
		NL2	-0.08	-0.02	-0.09	-0.03	0.14	0.10	0.10	0.00	-0.09	0.06	0.22	0.02	0.07	0.04	0.08
		SP1	-0.05	-0.02	-0.15	0.14	0.29	0.28	0.04	0.28	0.03	0.14	0.17	0.13	-0.13	0.00	0.23
		SP2	-0.28	-0.11	-0.28	-0.07	0.13	0.07	0.04	0.13	0.08	0.04	0.19	-0.02	0.10	0.05	0.26
14620	7@161.2	NL1	0.13	0.01	0.12	-0.02	-0.10	-0.09	-0.08	-0.08	-0.06	-0.10	-0.09	-0.20	-0.11	-0.01	-0.14
		NL2	0.07	0.11	0.06	0.00	0.10	0.08	0.17	0.03	-0.01	0.00	0.11	0.01	0.10	0.09	0.14
		SP1	-0.02	-0.01	0.05	-0.17	-0.13	-0.15	-0.04	-0.12	-0.11	-0.11	-0.07	-0.09	-0.06	-0.02	0.07
		SP2	0.17	0.22	0.19	-0.11	-0.11	-0.13	0.01	-0.12	-0.23	-0.04	-0.01	-0.12	0.06	0.09	0.01
SP491	8@18.8	NL1	0.08	-0.05	0.07	0.00	-0.01	0.00	-0.04	-0.08	0.00	0.17	0.01	0.12	0.12	0.19	-0.12
		NL2	0.12	0.10	0.10	0.17	0.05	0.10	0.07	0.16	0.14	0.18	-0.02	0.10	0.03	0.20	0.13
		SP1	0.02	0.08	0.05	-0.07	-0.11	-0.09	-0.10	-0.01	0.02	0.10	-0.07	0.03	-0.10	0.09	0.05
		SP2	0.07	0.07	0.08	0.03	-0.04	-0.01	-0.16	0.06	-0.01	0.04	-0.08	-0.10	-0.03	0.14	0.10

SP565	8@63.6	NL1	0.01	0.13	-0.03	-0.02	0.09	0.06	0.05	0.08	0.00	0.00	0.16	-0.09	0.04	0.09	0.17
		NL2	-0.04	0.02	-0.05	-0.08	0.08	0.04	0.07	-0.04	-0.22	-0.09	0.12	-0.17	0.04	-0.03	-0.03
		SP1	-0.17	-0.14	-0.17	-0.01	0.13	0.11	-0.04	0.01	0.00	0.00	0.18	-0.15	0.06	0.11	-0.01
		SP2	-0.07	-0.02	-0.09	-0.01	0.10	0.08	0.17	0.02	-0.02	0.03	0.24	-0.01	0.07	0.03	0.00
SP460	8@94.0	NL1	0.02	-0.01	0.10	-0.17	-0.13	-0.16	-0.08	-0.12	-0.11	0.05	-0.01	-0.03	-0.09	0.01	0.07
		NL2	0.03	0.08	0.07	-0.03	0.00	-0.02	0.03	-0.05	0.00	0.14	-0.02	0.03	0.00	0.15	0.08
		SP1	0.09	0.20	0.12	-0.13	-0.07	-0.10	0.22	-0.13	-0.20	0.12	0.08	-0.01	0.00	0.11	-0.03
		SP2	-0.03	0.14	-0.01	-0.11	-0.05	-0.08	-0.03	-0.06	-0.06	-0.01	-0.04	0.07	0.03	-0.05	0.09
SP509	9@2.1	NL1	-0.02	0.04	-0.08	0.32	0.24	0.30	0.14	0.19	0.24	0.07	-0.01	0.06	0.04	0.08	0.18
		NL2	-0.19	-0.18	-0.20	0.09	0.04	0.07	-0.07	0.10	0.13	-0.05	0.00	0.03	-0.01	-0.06	0.03
		SP1	-0.04	0.09	-0.09	0.26	0.16	0.21	-0.16	0.25	0.21	-0.04	0.04	-0.03	0.05	0.00	0.04
		SP2	-0.14	-0.03	-0.19	0.22	0.22	0.25	-0.04	0.17	0.26	-0.05	0.04	0.05	-0.02	-0.10	-0.03
Hpms2_24	9@30.7	NL1	0.06	0.26	0.07	-0.10	-0.14	-0.14	0.05	-0.07	-0.02	-0.18	-0.01	-0.21	-0.06	-0.09	-0.12
		NL2	0.04	0.06	0.07	-0.04	-0.02	-0.03	0.05	-0.03	0.05	-0.07	0.04	-0.07	0.08	0.02	0.04
		SP1	0.04	-0.01	0.05	-0.12	-0.13	-0.15	0.11	-0.11	-0.04	-0.12	0.05	-0.08	0.03	-0.09	-0.07
		SP2	-0.08	-0.12	-0.01	-0.14	-0.20	-0.19	-0.14	-0.08	0.08	-0.05	-0.01	0.01	0.19	-0.08	0.03
SP677	9@65.4	NL1	-0.22	-0.06	-0.16	0.00	-0.01	-0.01	-0.17	0.01	0.10	-0.02	-0.07	-0.05	-0.13	-0.05	-0.01
		NL2	-0.01	0.01	0.07	0.00	-0.07	-0.06	-0.10	0.00	0.11	0.06	-0.08	0.02	-0.10	0.08	0.12
		SP1	-0.14	-0.02	-0.19	0.18	0.24	0.26	-0.20	0.21	0.21	-0.01	-0.06	-0.01	-0.04	0.02	0.11
		SP2	0.06	0.11	0.11	-0.06	-0.01	-0.04	-0.02	0.01	-0.16	0.00	-0.07	-0.03	-0.01	0.03	0.14
SP680	9@124.2	NL1	0.09	0.01	0.01	-0.01	-0.02	-0.01	0.03	-0.09	-0.13	0.18	0.03	0.10	-0.14	0.07	-0.44
		NL2	0.15	0.02	0.13	-0.04	-0.05	-0.05	0.04	-0.07	-0.07	0.25	0.02	0.01	-0.05	0.23	-0.17
		SP1	0.15	-0.04	0.17	-0.24	-0.24	-0.25	0.19	-0.20	-0.23	0.24	0.07	0.04	-0.10	0.19	-0.05
		SP2	0.00	-0.12	0.03	-0.13	-0.12	-0.13	0.06	-0.15	-0.06	0.24	0.10	-0.02	0.00	0.30	-0.13
2249	10@7	NL1	-0.05	-0.12	-0.06	0.13	-0.06	0.00	-0.39	0.09	0.34	0.13	-0.26	-0.04	-0.09	0.22	-0.15
		NL2	-0.03	-0.10	0.00	0.02	-0.12	-0.09	-0.20	0.02	0.16	0.02	-0.34	-0.13	-0.25	0.10	-0.02
		SP1	0.03	0.08	0.02	0.17	-0.04	0.05	-0.20	0.14	0.22	0.09	-0.21	-0.10	-0.09	0.21	0.03
		SP2	0.04	-0.04	0.07	0.14	-0.04	0.04	-0.19	0.05	0.16	0.17	-0.23	-0.04	-0.11	0.23	-0.07
SP748	10@69	NL1	-0.01	-0.05	-0.03	0.06	0.08	0.08	0.10	-0.01	0.12	0.15	0.10	0.20	0.14	0.08	0.02
		NL2	-0.08	0.03	-0.10	-0.06	-0.02	-0.04	-0.06	-0.03	-0.11	0.02	-0.01	-0.02	-0.02	-0.06	-0.16
		SP1	-0.09	0.09	-0.10	0.09	0.05	0.02	-0.04	0.04	0.14	0.22	0.09	0.09	-0.05	0.15	-0.13
		SP2	-0.18	-0.03	-0.13	0.02	-0.02	0.00	-0.06	0.05	0.20	0.12	0.01	0.16	-0.06	0.00	-0.01

Table 3E3 Detected QTLs and Their Effects for Trait-Environment Combinations from MTME analysis: Chromosomes 11 and 12

Markers	Pos	Env	DWF	NF	pt_frt	DWL	DWS	DWV	LUE	LAI	pt_lf	Axl	SL	NLE	NI	INL	SLA
8758	11@19.8	NL1	-0.09	0.12	-0.04	-0.08	0.07	0.03	0.03	-0.15	-0.09	0.05	0.08	-0.05	-0.08	0.08	-0.15
		NL2	-0.05	0.00	0.00	-0.19	-0.06	-0.11	-0.05	-0.18	-0.09	0.01	-0.01	-0.02	-0.12	0.03	-0.01
		SP1	-0.08	-0.05	-0.06	-0.15	-0.05	-0.11	-0.01	-0.09	-0.04	0.09	0.07	0.03	-0.09	0.06	0.05
		SP2	-0.08	-0.02	0.01	-0.19	-0.12	-0.16	-0.02	-0.15	-0.08	0.02	0.05	0.13	-0.02	-0.05	0.01
SP699	11@40.7	NL1	0.12	-0.14	0.06	0.17	0.07	0.10	0.17	0.08	0.23	0.10	-0.03	-0.01	-0.08	0.20	-0.02
		NL2	0.00	-0.06	0.00	0.11	-0.06	-0.01	-0.01	0.02	0.21	0.16	-0.06	0.01	0.01	0.24	-0.01
		SP1	-0.01	-0.06	-0.02	0.07	0.02	0.07	0.03	-0.01	0.08	0.10	0.01	-0.03	-0.02	0.12	-0.15
		SP2	0.21	0.10	0.24	0.10	-0.14	-0.06	0.00	-0.06	-0.02	0.11	-0.18	-0.11	-0.09	0.18	-0.23
11379	11@73.3	NL1	-0.19	-0.10	-0.22	-0.04	0.08	0.05	-0.02	-0.02	0.03	0.00	0.05	-0.04	0.07	0.00	0.03
		NL2	-0.06	0.01	-0.01	-0.03	-0.09	-0.07	-0.14	-0.08	-0.02	-0.02	-0.07	-0.13	-0.08	0.01	-0.23
		SP1	-0.10	-0.09	-0.23	0.18	0.36	0.26	0.26	0.08	0.05	0.19	0.48	-0.25	0.45	0.39	-0.27
		SP2	-0.22	-0.03	-0.23	0.10	0.07	0.10	-0.19	0.13	0.33	-0.05	-0.12	0.10	0.00	-0.10	0.05
6257	11@139	NL1	0.11	0.05	0.18	-0.21	-0.17	-0.20	-0.03	-0.17	-0.29	-0.15	-0.17	-0.06	0.11	-0.15	0.27
		NL2	0.11	0.05	0.16	-0.07	-0.07	-0.07	0.05	-0.09	0.01	-0.07	-0.02	0.10	0.15	-0.08	0.03
		SP1	0.09	0.03	0.15	-0.27	-0.22	-0.28	-0.08	-0.15	-0.26	-0.17	-0.10	0.05	0.16	-0.21	0.25
		SP2	0.02	0.00	0.08	-0.12	-0.10	-0.11	0.03	-0.17	-0.12	-0.15	0.01	0.02	0.17	-0.20	-0.09
SP724	12@20.2	NL1	0.07	0.16	0.06	0.08	0.00	0.03	-0.02	-0.11	0.08	-0.15	-0.03	-0.33	0.11	0.12	-0.16
		NL2	0.28	0.22	0.35	-0.04	-0.16	-0.13	-0.06	-0.07	0.03	-0.17	-0.19	-0.42	-0.11	0.18	-0.02
		SP1	0.20	0.18	0.19	-0.01	-0.15	-0.09	0.08	0.00	-0.06	-0.15	0.09	-0.44	0.32	0.24	0.04
		SP2	0.17	0.17	0.20	-0.03	-0.15	-0.13	-0.04	-0.01	-0.12	-0.30	-0.08	-0.58	0.25	0.20	0.08
SP518	12@42.5	NL1	0.02	-0.04	0.08	-0.28	-0.20	-0.25	-0.02	-0.09	-0.31	-0.15	-0.10	-0.09	0.05	-0.17	0.37
		NL2	-0.02	0.02	0.03	-0.21	-0.12	-0.16	-0.03	-0.17	-0.12	-0.08	0.07	0.09	0.15	-0.12	0.10
		SP1	-0.02	0.04	0.04	-0.17	-0.08	-0.16	0.01	-0.14	-0.16	-0.05	-0.05	-0.03	0.09	-0.08	0.07
		SP2	0.03	0.02	0.06	-0.06	0.00	-0.01	-0.02	-0.18	-0.15	-0.04	-0.11	0.07	-0.01	-0.13	-0.19
Gpms_117	12@75.2	NL1	-0.02	0.02	-0.05	-0.17	-0.12	-0.15	-0.35	-0.23	-0.10	0.08	0.07	0.00	-0.17	0.09	-0.39
		NL2	-0.15	-0.07	-0.21	-0.04	0.10	0.06	-0.02	0.05	-0.05	-0.12	0.02	-0.09	-0.06	-0.07	0.16
		SP1	0.02	-0.08	0.02	-0.11	-0.10	-0.13	0.04	-0.12	-0.07	0.06	-0.04	0.07	-0.22	0.01	-0.03
		SP2	0.00	0.13	-0.08	0.00	0.09	0.05	0.07	0.06	-0.02	0.09	0.07	-0.03	0.00	0.16	0.09
21782	12@97.9	NL1	0.00	-0.09	0.02	-0.02	0.20	0.15	0.33	-0.01	-0.19	-0.03	0.05	0.10	0.08	-0.13	0.26
		NL2	0.14	0.03	0.15	0.06	0.05	0.06	0.21	-0.03	-0.04	0.12	0.14	0.17	0.17	0.01	-0.17
		SP1	-0.02	0.05	0.01	-0.07	0.07	0.06	0.00	0.00	-0.17	-0.02	0.12	0.10	0.08	-0.09	0.16
		SP2	-0.02	-0.06	0.03	-0.11	-0.04	-0.06	0.02	-0.11	-0.14	-0.05	0.07	0.02	0.08	-0.12	0.03
16897-2	12@130	NL1	-0.21	-0.24	-0.20	-0.03	-0.10	-0.09	-0.29	0.03	0.11	-0.18	-0.12	-0.03	-0.06	-0.29	-0.08
		NL2	-0.16	-0.15	-0.13	0.10	-0.01	0.02	-0.08	0.05	0.21	-0.07	-0.02	0.09	0.07	-0.12	0.01
		SP1	-0.10	-0.10	-0.12	0.23	0.11	0.12	-0.19	0.17	0.27	-0.09	-0.06	0.07	-0.06	-0.14	-0.05
		SP2	-0.06	-0.02	-0.13	0.22	0.14	0.19	-0.10	0.14	0.24	-0.15	-0.12	0.18	-0.01	-0.31	-0.08

CHAPTER 4

Predicting complex traits in multiple environments by a combination of genomic prediction and crop growth modelling: an example in pepper

CHAPTER 4

Predicting complex traits in multiple environments by a combination of genomic prediction and crop growth modelling: an example in pepper

4.1. Abstract

The prediction of yield as a complex trait with variations across environments may be done using direct and indirect approaches, where the latter are based on dissection of the complex trait into component traits. The direct approach can be based on either QTL prediction (QP) or genomic prediction (GP) models for the trait itself, while the indirect approach can be based on a crop growth model (CGM) that prescribes a dissection of the complex trait into a number of component traits. We first compared the direct prediction performance of single-trait (ST) and multi-trait (MT) versions of both QP and GP models for a recombinant inbred lines population of 149 individuals in pepper. The predictive performances of the models were assessed using five yield related traits measured across four environments. The four methods differed in their predictive accuracies, ranging from 0.11 to 0.89. MT models generally had higher predictive accuracy than ST models with MT-GP being the most superior for all traits across the four environments. GP methods outperformed QP methods in both single and multi-traits situations. In the indirect prediction strategy, the CGM was applied on the breeding values of yield component traits from both QP and GP methods. The indirect strategy was implemented for within-environment and across-environment analyses. The predictive accuracies from CGM were comparable to that of the direct prediction strategy. The indirect approach seemed to work well at first sight, but this is especially due to the fact that yield appeared to be strongly driven by just one component, the partitioning to fruit. The across environment CGM indicated that we may use component traits and environmental information from one environment to predict yield in another environment.

Keywords

Complex Trait Dissection; Component Trait; Genomic Prediction; Multi-Trait model.

4.2. Introduction

In plant breeding, complex traits such as yield are difficult to improve and predict, because complex traits can be influenced by many quantitative trait loci (QTLs) of small effects, many of which will show variations across environments, QTL by environment interaction (QEI). To predict the complex trait, we may use procedures based on QTLs or breeding values of the complex trait itself, i.e., direct prediction procedures. Recently, we studied yield in pepper as an example of a complex trait (Alimi et al., 2013b). In that study, yield was measured across several environments and appeared to be influenced by a number of QTLs of small effects, with some of these QTLs displaying QEI. In this paper, we will explore various prediction strategies for yield in pepper, where yield and a number of its component traits are used. Besides the direct prediction procedure via QTLs on the basis of multi-QTL models and breeding values from genomic prediction models, we will investigate an indirect prediction strategy for yield using a crop growth model (CGM) with as inputs on the one hand yield components or their genomic predictions and on the other hand environmental variables.

The use of CGMs with implicit dissections of complex traits via physiological component traits is an exciting alternative to direct prediction of complex traits by QTLs or breeding values. CGMs are based on prior biological and environmental knowledge (Tardieu, 2003; Van Ittersum et al., 2003) and are useful for understanding complex traits in terms of underlying component traits (Bustos-Korts et al., 2016; Van Eeuwijk et al., 2010; Chapman, 2008; Hammer et al., 2006). Uptmoor et al. (2008) and Yin et al. (2005) presented appealing cases of integrated QTL and CGM approaches and showed that flowering time can be effectively predicted by substituting QTL predictions for component traits in CGMs. The component traits should be biologically meaningful, easily measurable, and they should have a relatively simple genetic basis preferentially without genotype-by-environment interactions (GEI) and/or QEI (Reymond et al., 2003). Also, a known relationship should exist between the complex trait and the component traits. Since CGMs contain explicit representations of development over time and integrate developmental and environmental information, they have the added advantage of being able to describe genotype-by-environment interactions (GEI) (Malosetti et al., 2016; Technow et al., 2015; Chenu et al., 2009; Cooper et al., 2009). Therefore, component traits from one environment can be used to predict yield in another environment, provided the structure of the CGM is correct and GEI is small or absent for the component traits.

In this paper, we will calculate breeding values for component traits from different prediction models subject to cross-validation: multi-QTL based prediction (QP) and genomic prediction (GP). QTL based prediction of the phenotype was proposed in the early 1990s (Paterson et al., 1991; Lande and Thompson, 1990) to accelerate genetic improvement. Nowadays, genome-wide dense marker maps at affordable cost have made GP an interesting alternative to QP, where the difference between GP and QP resides mainly in the use of all markers in a penalized regression context for GP versus the use of a limited set of QTL related markers in QP. The key principle of GP is to simultaneously estimate the effects of all genome-wide markers in a training population consisting of

genotyped and phenotyped individuals and then predict the genomic estimated breeding value (GEBV) of genotyped but not-phenotyped individuals in test/future generations (Meuwissen et al., 2001). The large majority of QTLs are assumed to be in linkage disequilibrium with one or more molecular markers. GEBVs are calculated as the sum of estimated marker effects for genotyped individuals in a training population. Fitting all markers simultaneously ensures that marker-effect estimates are unbiased, small effects are captured, and that there is no multiple testing issue (Jia and Jannink, 2012).

Single-trait and multi-trait versions of both QP and GP will be employed. So far, no study has compared prediction performance from both multi-trait QP and multi-trait GP methods. Multi-trait analysis helps in improving the prediction of some traits with low heritabilities or of traits that are difficult to measure by exploiting their genetic correlations with other traits of higher heritability. Many studies on QTL and association mappings have shown that the joint analysis of multiple traits helps to improve the power and precision of QTL (Alimi et al., 2013b; Jiang and Zeng, 1995) and association mappings (Galesloot et al., 2014; Stephens, 2013). Also, several studies have shown that multi-trait genomic prediction (MT-GP) methods performed better than single trait genomic prediction (ST-GP) methods (Burgueño et al., 2012; Jia and Jannink, 2012; Sørensen et al., 2012; Calus and Veerkamp, 2011).

In this paper, we first present and compare prediction accuracies from single–trait (ST) and multi-trait (MT) versions of both QP and GP models. We consider Bayesian LASSO Regression (BLR) (Legarra et al., 2011; Park and Casella, 2008) as the ST-GP model and Bayesian Latent Variable (BLV) model (Janss, 2014; Sørensen et al., 2012) as the MT-GP model. The prediction accuracy (correlation of GEBV with trait) and bias (slope of GEBV on trait) from the four models are evaluated using five yield related traits measured across four environments. The traits and environments are from the EU-SPICY project, see Alimi et al. (2013a) and Voorrips et al. (2010). In the second part of the paper, we investigate an indirect prediction strategy for yield by first predicting a set of physiological component traits of yield (Van Eeuwijk et al., 2010; Chapman, 2008; Hammer et al., 2006) and utilize a crop growth model (CGM) to model yield as a function of the predicted component traits. We apply the CGM on the breeding values from the four prediction models enumerated above. The accuracies of predicting yield through the CGM will be explored both within and across environments. The latter across environments prediction is a form of genotype-by-environment interaction (GEI) prediction, and will be compared to direct yield prediction in an environment.

4.3. Materials and Methods

4.3.1. Genotypic and Phenotypic data

The pepper population used in this study consisted of 149 individuals obtained from the sixth generation (F_6) of the segregating recombinant inbred lines (RILs) of an intraspecific cross between the large – fruited cultivar 'Yolo Wonder' (YW) and the pungent small-fruited cultivar 'Criollo de Morelos 334' (CM 334) of *Capsicum annuum*. All individuals were genotyped for 455 markers on 12 chromosomes covering 1705cM. Phenotyping experiments were carried out at two locations, i.e., Wageningen in the Netherlands (NL) and El Ejido in Spain (SP), representing temperate and Mediterranean growing conditions respectively. At both locations, experiments were done during two time periods: December – May (1) and June – December (2). This generated four experiments denoted as NL1, NL2, SP1 and SP2. All the experiments lasted for about five months except NL2 that lasted for only two months. In the NL trials, only one stem per plant was kept. Plant density was approximately 6.4 plants per m^2 (i.e. about six stems per m^2). In the SP trials, two stems per plant were kept. Plant density was approximately three plants per m^2 (i.e. six stems per m^2). In the four trials, greenhouse air temperature, humidity, CO_2 concentration and inside global radiation were registered every five minutes. A large number of traits relating to vegetative and fruit development of pepper crop was measured in the four trials (Alimi et al., 2013a).

Five traits related to yield (total weight of fruits) were selected across the four trials for the purpose of this study. These traits were increase rate of leaf area index (LAI_{rate}) which expresses mean increase in leaf area index per unit time, where time is expressed in degree-days, radiation use efficiency (RUE), which is the dry matter production (g) per megajoule (MJ) of intercepted global radiation, and partitioning into fruit (PF), which expresses the proportion of total plant biomass due to fruits. Other traits were total number of fruits (NF) and total dry weight of fruit (DWF). Both NF and DWF included the fruits harvested during the growing season and the fruits on the plant at the final destructive harvest. DWF was taken to represent measured yield. Best linear unbiased estimates (BLUEs) (Alimi et al., 2013a) which are the genotypic means for each of the five traits were used as observed phenotypes for the traits. Summary statistics for the selected traits were obtained. Correlations between the experiments were calculated for each trait, and also correlations between the traits within each experiment. These correlations were calculated from the genotypic means for the traits. We will refer to these correlations as genetic correlations.

4.3.2. Univariate and Multivariate QTL Prediction Models

The single-trait QTL prediction (ST-QP) model follows from a single trait multi-QTL analysis for each trait. This model is of the form:

$$y_i = \mu + \sum_{j \in Q} x_{ij}\beta_j + e_i, \qquad (4.1)$$

where y_i was the phenotypic response of genotype i ($i = 1...N$), μ the population mean, β_j was the additive effect of QTL j ($j \in Q$), with Q the set of selected QTLs constructed from

the set of QTLs that were identified by interval mapping (Zeng, 1994) followed by a backward elimination procedure. Genetic predictors (Lynch and Walsh, 1998) were calculated at all marker positions and intermediate positions for those marker intervals that were larger than 5cM, giving a total of 639 evaluation points. The genetic predictor for genotype *i* at genomic evaluation point *j* is denoted by x_{ij}, and e_i was the residual term, which contains both genetic (polygenic, non-detected QTLs) and non-genetic (plot error) contributions.

The multi-trait QTL prediction (MT-QP) model is a joint analysis combining the five traits within each environment in a mixed model QTL analysis (Alimi et al., 2013b; Malosetti et al., 2008). Traits were standardized to have mean zero and standard deviation one. The responses, y_{ik} ($i=1...n_G$, with n_G being the number of genotypes, and $k = 1... n_T$, with n_T being the number of traits), are modelled by first an overall mean for each trait, μ_k, then a QTL term containing products of genetic predictors, x_{ij}, for genotype *i* at genomic evaluation point *j*, and trait-specific QTL effects, β_{kj}, and finally a genetic/residual component, e_{ik} :

$$y_{ik} = \mu_k + \sum_{j \in Q} x_{ij}\beta_{kj} + e_{ik}, \qquad (4.2)$$

We assumed that the vector $\boldsymbol{e}_i = (e_{i1}, \dots, e_{in_T})$ follows a multivariate normal distribution with zero mean and a first order factor analytic variance-covariance matrix Σ i.e. $e_i \sim N(0, \Sigma)$. This model accounts for genetic correlations between traits. The QTL models were implemented in the GenStat QTL library (Payne et al., 2011), following the strategy described by Malosetti et al. (2013) and Boer et al. (2007).

4.3.3. Bayesian Genomic Prediction Models

For large numbers of markers (*M*) and relatively few genotypes/individuals (*N*), ordinary least square procedures break down. Therefore, for estimation in relation to GP models, variable selection and shrinkage estimation methods are used to tackle the problem of high dimensionality in the predictors (Habier et al., 2011; Legarra et al., 2011; De Los Campos et al., 2009; Hayes et al., 2009). These estimation methods try to reduce mean squared error (MSE) by reducing the variance of the estimator. This may however introduce bias in the estimate. The obtained penalized estimates are the solution to an optimization problem that balances model fit and model complexity. The optimization problem is generally of the form:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\{L(y, \beta) + \lambda J(\beta)\}. \qquad (4.3)$$

Where $L(y,\beta)$ is a loss function that measures lack of fit of the model to the data, $J(\beta)$ is a measure of model complexity and $\lambda \geq 0$ is a regularization parameter controlling the trade-offs between fitness and model complexity. In a Bayesian setting, shrinkage estimation is controlled by the choice of the prior density assigned to marker effects (De Los Campos et al., 2009).

4.3.4. Univariate Genomic Prediction Model: Bayesian LASSO Regression

The linear mixed model employed for the single-trait genomic prediction (ST-GP) within each environment is of the form:

$$y_i = \mu + \sum_{j=1}^{M} X_{ij}\beta_j + e_i, \qquad (4.4)$$

where y_i was an element of a vector (N x 1) of phenotypes on N individuals, μ was the overall population mean, X was a design matrix (N x M) allocating the M marker genotypes to N individuals, β_j was the allele substitution effect for marker j and assumed normally distributed $\beta_j \sim N(0, \sigma^2_{\beta_j})$, e_i was an element of the vector (N x 1) of identically and independently distributed residuals with $e \sim N(0, \sigma^2_e)$.

In LASSO (Least Absolute Shrinkage and Selection Operator), $J(\beta) = \sum_{j=1}^{M} \|\beta_j\|$, i.e. the model complexity is the sum of the absolute values of the regression coefficient (Tibshirani, 1996). This penalty combines both subset selection and shrinkage estimation since it induces a solution that may involve zeroing-out some regression coefficients and shrinkage estimates of the remaining effects. The optimization problem in (3) now becomes:

$$\hat{\beta}_{BLR} = \underset{\beta}{\operatorname{argmin}}\{(y - X\beta)'(y - X\beta) + \lambda \sum_j |\beta_j|\}. \qquad (4.5)$$

In the Bayesian paradigm, the solution to (4.5) is the posterior mode of the combination of a Gaussian likelihood and a double exponential (DE) prior density for the marker effects (de los Campos et al., 2009).

$$p(y, \beta | \sigma^2_\varepsilon, \sigma^2_\beta, \lambda) = \prod_i N(y_i | \sum_j x_{ij}\beta_j, \sigma^2_\varepsilon) \times \prod_j (\lambda/2)\exp(-\lambda|\beta_j|), \qquad (4.6)$$

The parameter λ in BLR controls the prior on regression coefficients β_j. The higher the values of λ, the higher the penalty on β_j. This ensures stronger shrinkage of coefficients that are close to zero and less shrinkage of those with large absolute values.

4.3.5. Multivariate Genomic Prediction Model: Bayesian Latent Variable

A Bayesian latent variable (BLV) model was employed as a multi-trait genomic prediction (MT-GP) model. Here, we modelled all traits within an environment following the method described in Janss (2014) and Sørensen et al. (2012). Our MT-GP model is comparable in structure to the MT-QP model (4.2) described above. However, now not only the residuals are modeled by a multiplicative, factor analytic structure, but also the SNP effects. In the MT-QP model the QTL/SNP effects were taken fixed and the residuals followed a factor analytic rank one model with an additional independent error term with trait specific variance. In our MT-GP model the QTL/SNP effects are taken random with a multivariate Normal distribution and a restriction on the variance-covariance matrix to follow a multiplicative model making the SNP effects becoming correlated between traits. Both

marker effects and their variances were jointly estimated in a single hierarchical model as the variances were also treated as unknown.

The BLV model specification was: $y_{ik} = \mu_k + \sum_j^M X_{ij}\beta_{jk} + e_{ik}$, (4.7)

Where y_{ik} denoted the phenotype of the *i*-th individual for *k*-th trait, μ_k ($k = 1,\ldots,n_T$) was the overall mean of each trait, β_{jk} was the random regression coefficient of trait k on marker j and e_{ik} is the residual term. For a full model description and an account of the distributional assumptions involved see Sørensen et al. (2012).

Inferences for the GP methods were based on Gibbs sampling of 110000 samples. The first 10000 samples were discarded as burn-in, while 500 of the remaining samples were stored, i.e., using a skip factor of 200. Visual inspections of trace plots confirmed convergence of the Markov chains. The GP models were implemented in Bayz software (Janss, 2011).

4.3.6. Yield Indirect Prediction through Crop Growth Model

The ecophysiological crop growth model (CGM) (Figure 4.1) employed here was a LINTUL-type (Light INTerception and Utilization) crop growth model that simulated the formation of pepper yield under potential growing conditions (Van Ittersum et al., 2003; Spitters and Schapendonk, 1990). The main environmental factors considered in the CGM we adopted were radiation and temperature.

For genotype *i* in environment *h*, the CGM can be mathematically written as:

$$Yield_{ih} = PF_{ih} * RUE_{ih} * \sum_1^D \left(I_{hd} * (1 - e^{-k*LAI_{ihd}})\right). \quad (4.8)$$

Yield is accumulated over the growing days *d*=1... *D*. The leaf area index (LAI) is dynamic and for genotype *i* in environment *h* on a specific day *d* ($d \leq D$) calculated as $LAI_{ihd} = LAI_{rate_{ih}} * \sum_1^d (T_{hd} - T_b)$. The term $\sum_1^d (T_{hd} - T_b)$ is the accumulated thermal time till day *d*, expressed in degree-days, and $LAI_{rate_{ih}}$ is a genotype specific increase rate of leaf area index. T_{hd} is the daily average temperature in environment *h* on day *d*, and T_b is the base temperature below which no development takes place, taken as 10°C in all environments (Marcelis et al., 2006). The increase rate of leaf area index (LAI_{rate}) was calculated as the ratio between the increase of leaf area index (LAI) and thermal time between initial and final harvest. The fraction intercepted radiation is $1 - e^{-k*LAI_{ihd}}$, where a value of 0.7 for the extinction coefficient *k* is assumed for all genotypes (Marcelis et al., 1998). Daily intercepted radiation is then calculated from this fraction and I_{hd}, the daily global radiation intensity (MJ m^{-2} d^{-1}). The daily intercepted radiation is consequently multiplied by the radiation use efficiency (RUE_{ih}) resulting in daily dry matter production. RUE represents the biomass produced per unit of intercepted radiation. It is the ratio between biomass increase and the total intercepted radiation, which was the daily intercepted global radiation summed over the total growth period. Finally, yield is calculated from the total accumulated dry matter by multiplying it by the fraction biomass

partitioned into the fruits (PF_{ih}), i.e. fruit biomass/total plant biomass. Total plant biomass was calculated as the sum of plant dry weight at final destructive harvest and the dry weight of the already harvested fruits.

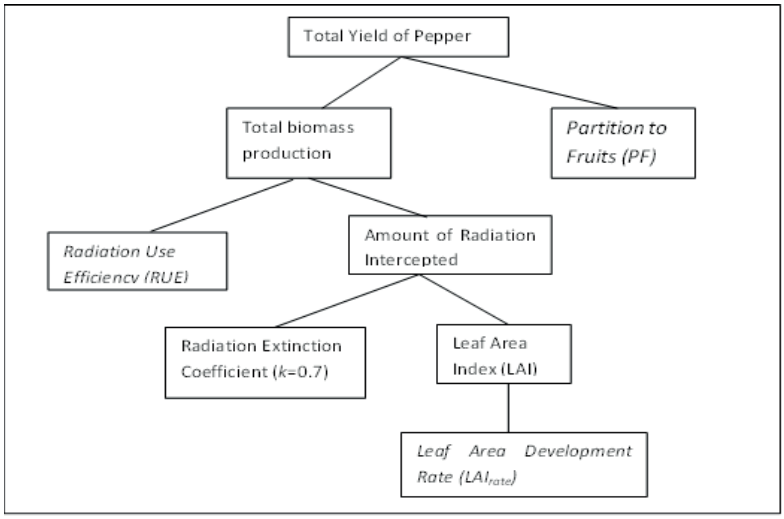


Figure 4.1 Schematic diagram of the LINTUL-type crop growth model (CGM). The figure describes how the target complex trait yield is dissected into component traits. The three yield component parameters for yield predictions are in italics (adapted from (Higashide and Heuvelink, 2009)).

4.3.7. Model Validation and Accuracy

Our models were validated via a cross-validation (CV) scheme (Efron and Gong, 1983; Kohavi, 1995). We generated training and testing sets using a five-fold CV (subsets), and repeated this CV scheme 10 times. This means that the 149 individuals were randomly divided into five non-overlapping subsets. Four subsets were taken as the training set and were used to fit the model. The fitted model was then used to obtain predictions for the fifth subset. This was repeated until all the subsets were used as the testing set. This CV scheme is similar to situation where predictions are obtained for individuals that have not been phenotyped. In our CV scheme, each testing set had 30 individuals except fold number five which had 29 individuals.

Prediction accuracies and bias of the estimates were used to evaluate performance of the different models. Prediction accuracy refers to the correlation between the genomic estimated breeding values (GEBV) from the models and the phenotypic best linear unbiased estimates (BLUE) of each individual. Bias was assessed as the coefficient of regression of GEBV on BLUE (Resende et al., 2012). Unbiased models are expected to have a slope coefficient of one, whereas values greater than 1 indicate a biased overestimation in the GEBV and values smaller than one indicate a biased underestimation in the GEBV.

4.3.8. Yield Prediction Strategies

Both direct and indirect prediction strategies were employed for yield prediction (Figure 4.2). By direct prediction, we mean the use of yield phenotypes in each environment to generate yield GEBVs for that environment through each of the four prediction models (ST-QP, MT-QP, ST-GP and MT-GP). For each environment, accuracy of yield prediction was calculated as the correlation between the yield GEBVs and yield BLUEs (phenotypes) in that environment.

In the indirect prediction strategy, yield was first predicted from the GEBVs for its three component traits (RUE, LAI_{rate} and PF) via the crop growth model. We then calculated the accuracies of predictions as the correlation between the predicted yield values and yield phenotypes. GEBVs from all four prediction models were used. Note that the GEBVs for both MT-QP and MT-GP were estimated on an analysis that included only the component traits, i.e., excluding DWF and NF.

The prediction accuracies for an indirect prediction strategy were estimated within and across environments. The within-environment indirect prediction strategy involved using GEBVs of component traits in one environment (e.g. SP1) to predict yield in the same environment. The across-environment CGM analysis is a form of genotype-by-environment (GEI) analysis where the GEBVs of component traits in one environment (e.g. SP1) were used to predict yield in another environment (e.g. NL1). The across-environment analysis envisages a situation where we wish to predict how a certain population will perform in a new environment. In such a situation, we can use phenotypic information or GEBVs from a related environment.

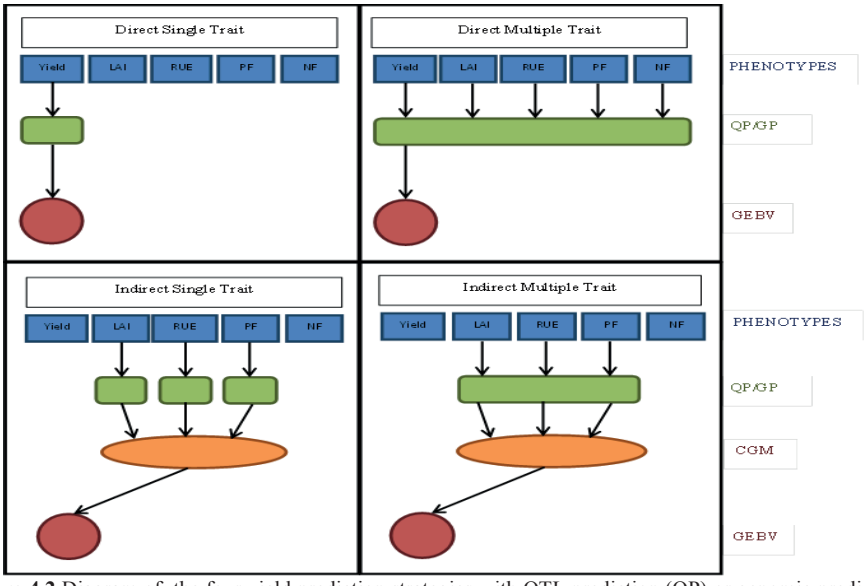


Figure 4.2 Diagram of the four yield prediction strategies with QTL prediction (QP) or genomic prediction (GP) models used to calculate genomic estimated breeding values (GEBV) for single trait or multiple traits jointly and a crop growth model (CGM) used in the indirect calculation of GEBV for Yield.

4.4. Results

4.4.1. Trait Descriptions

The phenotypic means and standard deviations for the five traits from each of the four environments are presented in Table 4.1. Yields per plant in NL trials were lower than in SP trials and yield per plant in NL2 was particularly very low. This was primarily due to the lower rate of fruit set in NL trials, especially in NL2. This difference is however reduced if yield per m^2 is considered since different plant pruning strategies were employed in the two locations. One stem was kept per plant in the Netherlands compared to two stems per plant in Spain. Hence, DWF and NF in NL trials were multiplied by two to correct for the different pruning systems and make the comparison on plant density per m^2 instead of per plant. The particularly low yield in NL2 was due to the high development rate of length growth in this trial. The trial lasted for only two months against 3-5 months for the other trials since the plants reached their maximal height rapidly. The ranges for traits means were in general highest in SP2.

Trait correlations across environments were moderate to high, ranging from 0.34 for RUE between NL2 and SP1 to 0.79 for LAI_{rate} between SP1 and SP2. The majority of these correlations were above 0.6 (Table 4.2) with overall mean of 0.61. This is an indication that measurements for a particular trait in one environment (e.g. LAI_{rate} in SP1) may be substituted with those for the same trait in another environment (e.g. LAI_{rate} in SP2). All the correlations below 0.5 involved RUE and NF in SP1.

The within trial correlations were mostly consistent in sign across the trials (Table 4.3). Very high correlations were observed among DWF, NF and PF. This was expected since PF was computed from total fruit weight and total plant biomass and also there is usually a direct relationship between number of harvested fruits and total fruit weight. RUE showed low but positive correlations to DWF and NF, while LAI_{rate} displayed very low and sometimes negative correlations to other traits.

4.4.2. Predictive Ability and Bias of the Four Prediction Models

Four different genome-wide prediction methods were compared on five traits measured across four environments. Overall, the predictive abilities of the models ranged from 0.01 from ST-QP for NF in NL2 to 0.93 from MT-GP for LAI_{rate} in SP2 (Figure 4.3). NF in both NL trials was very poorly predicted by all the methods except MT-GP. LAI_{rate} was well predicted in all environments by the four methods (0.37 – 0.93).

Table 4.1 Mean and standard deviation for each of the traits in the four environments calculated from genotypic means. NL1, NL2, SP1 and SP2 represent phenotypic trials in the Netherlands (NL) and Spain (SP) during spring (1) and autumn (2) in 2009. DWF, NF and PF stand for total dry weight of fruit, total number of fruits and the proportion of total plant biomass due to fruits respectively. RUE is the radiation use efficiency, which is the dry matter production (g) per megajoule (MJ) of intercepted global radiation while LAIrate expresses mean increase in leaf area index per unit time, where time is expressed in degree-days.

	Mean					Standard Deviation			
Trait	NL1	NL2	SP1	SP2		NL1	NL2	SP1	SP2
DWF[a]	46.24	15.54	66.89	87.13		33.54	12.78	36.68	32.77
NF[a]	24.96	12.60	28.98	37.41		14.02	9.26	11.04	11.44
PF	0.18	0.10	0.28	0.36		0.12	0.07	0.14	0.11
RUE	1.12	1.00	0.96	1.19		0.25	0.18	0.13	0.14
LAI_{rate}	2.22	3.01	1.56	1.69		0.70	0.98	0.49	0.44

[a] DWF and NF in NL trials were multiplied by two to correct for different pruning systems used in NL and SP trials in order to make the comparison on plant density per m^2 instead of per plant. One stem was kept per plant in the NL trials compared to two stems per plant in SP trials.

Table 4.2 Trait genetic correlations between environments e.g. NL1.NL2 refers to correlation of specific trait measurements (e.g. DWF) between environment NL1 and NL2. Environments and traits abbreviations are as explained in Table 4.1.

	NL1.NL2	NL1.SP1	NL1.SP2	NL2.SP1	NL2.SP2	SP1.SP2	Mean
DWF	0.72	0.60	0.61	0.53	0.62	0.58	0.61
NF	0.70	0.55	0.54	0.49	0.65	0.41	0.56
PF	0.69	0.65	0.67	0.54	0.72	0.57	0.64
RUE	0.64	0.45	0.60	0.34	0.64	0.36	0.51
LAI_{rate}	0.73	0.76	0.70	0.67	0.75	0.79	0.73
Mean	0.70	0.60	0.62	0.51	0.68	0.54	0.61

Table 4.3 Genetic correlation of traits within each environment. Environments and traits abbreviations are as explained in Table 1.

	DWF	NF	PF	RUE		Trait	DWF	NF	PF	RUE
NL1						SP1				
NF	0.85					NF	0.80			
PF	0.90	0.76				PF	0.93	0.74		
RUE	0.13	0.23	-0.11			RUE	0.40	0.37	0.34	
LAI_{rate}	0.07	0.01	-0.18	0.11		LAI_{rate}	0.09	0.12	-0.18	-0.34
NL2						SP2				
NF	0.86					NF	0.60			
PF	0.91	0.76				PF	0.89	0.51		
RUE	0.26	0.36	-0.03			RUE	0.22	0.36	0.03	
LAI_{rate}	0.19	0.09	-0.07	0.32		LAI_{rate}	-0.04	-0.01	-0.37	0.01

For most of the traits across the four environments, the single-trait Bayesian genomic prediction method (ST-GP) outperformed the single-trait QTL method (ST-QP) except in

NL2 where ST-QP performed better than ST-GP. Also, MT-GP clearly outperformed MT-QP for all the traits across all the environments. Averaging over all the traits in SP1 for example, ST-QP, MT-QP, ST-GP and MT-GP gave prediction accuracies of 0.37, 0.46, 0.59 and 0.89, respectively. These results confirm that GP methods have better predictive power than QTL methods. This might be expected as parameters from the GP methods are fitted using all available markers whereas parameters of QTL methods are fitted using selected QTL markers. Also, multi-trait models are expected to have better predictive power than single-trait models. This is mostly true for SP trials, but not for NL trials, especially NL2 where ST-QP had better prediction accuracies than MT-QP. The MT-GP clearly stood out as the best method in terms of its predictive ability for all the traits.

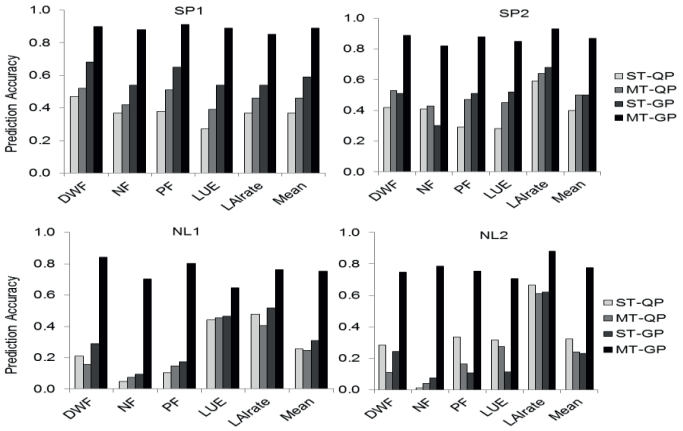


Figure 4.3 Prediction accuracies from the direct prediction strategy for each of the five traits in the four environments using the four prediction models.

Prediction accuracies differed across environments for each of the five traits irrespective of the prediction method employed. Traits were generally better predicted in SP trials than in NL trials. For example, DWF had prediction accuracies (from the four methods) ranging from 0.16 – 0.84 in NL1, 0.11 – 0.75 in NL2, 0.47 – 0.90 in SP1 and 0.42 – 0.89 in SP2. Overall, the average prediction accuracies across traits range from 0.23 – 0.78 (mean = 0.39) in NL trials and 0.37 – 0.89 (mean = 0.57) in SP trials. The difference in the accuracies of trait prediction across environments is an indication that these traits exhibit QEI (Alimi et al., 2013b).

The coefficient of regression of the GEBV on the phenotypes was calculated to measure the bias of each of the prediction models (Table 4.4). Unbiased models are expected to have a slope coefficient of one. Traits that were poorly predicted in an environment had regression coefficients very different from one (biased) while traits that were better predicted had regression coefficients very close to one (unbiased). An example is NF in NL2 with regression coefficients of 0.05, -0.18, 3.60 and 0.95 for ST-QP, MT-QP, ST-GP and MT-GP respectively. This indicates that only MT-GP gave unbiased prediction for NF

in NL2. All the models gave very good regression coefficients for LAI_{rate} in the four environments (0.86 – 1.15). Regression coefficients from MT-GP were the closest to one for all the traits across the four environments, indicating that the MT-GP model is most correctly estimating the genomic breeding values.

Table 4.4 Bias from the direct prediction strategy for each of the five traits in the four environments using the four prediction models. Environments and traits abbreviations are as explained in Table 4.1. ST-QP and MT-QP respectively stand for single-trait and multi-trait versions of QTL prediction model while ST-GP and MT-GP respectively stand for single-trait and multi-trait versions of genomic prediction model.

Env	Trait	ST-QP	MT-QP		ST-GP	MT-GP
NL1	DWF	0.64	0.49		2.03	1.12
	NF	0.13	0.27		8.23	1.08
	PF	0.29	0.47		2.45	0.96
	RUE	1.01	0.95		1.53	0.86
	LAI_{rate}	0.91	0.87		1.12	0.97
NL2	DWF	0.87	0.36		3.64	0.97
	NF	0.05	-0.18		3.60	0.95
	PF	0.77	0.49		9.66	0.95
	RUE	0.87	0.74		1.04	0.78
	LAI_{rate}	0.98	1.01		1.03	0.96
SP1	DWF	0.89	0.90		1.06	0.99
	NF	0.94	0.86		1.20	1.01
	PF	0.93	0.90		1.05	0.95
	RUE	0.78	0.80		1.26	0.96
	LAI_{rate}	0.94	0.86		1.15	0.94
SP2	DWF	0.87	0.93		1.08	1.07
	NF	0.88	0.81		1.39	1.02
	PF	0.86	0.89		1.17	1.05
	RUE	0.76	0.82		1.15	0.92
	LAI_{rate}	0.97	0.96		1.09	0.94

4.4.3. Accuracies of Yield Prediction from Crop Growth Model

For both within-environment and across-environment yield predictions using the component traits in the crop growth model (CGM), prediction accuracies vary depending on environment and prediction method employed (Table 4.5). In the within-environment CGM where component traits in a given environment were used to predict yield in the same environment, yields are better predicted in SP trials than in NL trials, irrespective of prediction method employed. For example, using ST-GP, yield in SP trials had prediction accuracies ≥ 0.5 against an accuracy of ≤ 0.2 in NL trials. The pattern observed for performances of the four predictive models under CGM is similar to the pattern seen in direct yield prediction from DWF GEBV (Figure 4.2). Here, MT-GP also gave very high

prediction accuracies across the four environments, while prediction accuracies from ST-QP, MT-QP and ST-GP were generally low, especially in NL environments (≤ 0.35). SP1 had the highest prediction accuracy of 0.92, similar to the result from direct prediction (0.90 for DWF in SP1).

The across-environment (GEI) CGM results (Table 4.5) revealed that it is possible to use component traits from one environment to predict yield in another environment. Using ST-QP result as an example, prediction accuracy for yield in NL1 environment improved from 0.19 to 0.31, 0.28 and 0.23, if component traits from NL2, SP1 and SP2 respectively were used. The prediction accuracies from MT-QP and ST-GP in NL trials increased if component traits from SP trials were employed. Prediction accuracy for yield in NL2 using MT-QP improved from 0.11 to 0.38 if component traits from SP1 were used instead of component traits in NL2 itself. However, prediction accuracies for yield in SP environments did not improve when component traits from NL environments were used. This may be due to the inherent measurement errors in the NL trials couple with the low yield in these environments.

MT-GP gave very high prediction accuracies (≥0.5) across all trial combinations. The prediction accuracies from MT-GP in the GEI CGM analyses were mostly higher than prediction accuracies from ST-QP, MT-QP and ST-GP in the within-environment CGM and direct prediction methods. All prediction accuracies in the GEI CGM were however lower than the prediction accuracies from MT-GP method in the within-environment analysis. This showed that yield in each environment was best predicted using MT-GP with component traits from the same environment, but still reasonable prediction is possible using component traits from similar environments.

Table 4.5 Accuracies of yield predictions using direct and CGM prediction strategies. Accuracy here is defined as the correlation coefficient between the estimated breeding values from the models and the genotypic means.

Trials	ST-QP	MT-QP	ST-GP	MT-GP
	Direct Prediction[a]			
NL1	0.21	0.16	0.29	0.84
NL2	0.28	0.11	0.24	0.75
SP1	0.47	0.52	0.68	0.90
SP2	0.42	0.53	0.51	0.89
	Within Environment Indirect Prediction using Crop Growth Model[b]			
NL1	0.19	0.11	0.18	0.83
NL2	0.35	0.11	0.19	0.81
SP1	0.42	0.55	0.69	0.92
SP2	0.34	0.49	0.49	0.86
	Across Environment Indirect Prediction using Crop Growth Model[c]			
NL1:NL2	0.31	0.11	0.23	0.63
NL1:SP1	0.28	0.34	0.49	0.60
NL1:SP2	0.23	0.35	0.35	0.57
NL2:NL1	0.12	0.11	0.28	0.59
NL2:SP1	0.31	0.38	0.48	0.49
NL2:SP2	0.21	0.36	0.40	0.56
SP1:NL1	0.37	0.21	0.30	0.58
SP1:NL2	0.49	0.31	0.47	0.52
SP1:SP2	0.46	0.54	0.56	0.58
SP2:NL1	0.18	0.17	0.04	0.49
SP2:NL2	0.42	0.23	0.19	0.51
SP2:SP1	0.29	0.47	0.47	0.58

[a] Yield in each environment was predicted directly from yield breeding values in that environment.

[a] In the direct prediction strategy, no way to estimate prediction accuracies for MT-QP and MT-GP analyses if they are based on only the component traits since DWF was not included in the analyses. The values from joint analyses based on all the five traits are thus reported for MT-QP and MT-GP in the direct strategy.

[b] Yield in each environment was predicted via crop growth model using component traits from the same environment.

[c] Yield in an environment was predicted via crop growth model using component traits from another environment e.g. NL1:NL2 implies that yield in NL1 was predicted using component traits from NL2.

4.5. Discussion

In this paper, we studied two important objectives with respect to the prediction of complex traits. The first objective was to compare the predictive performances of QTL prediction (QP) and genomic prediction (GP) methods. In recent years, several studies have reported on the performance of GP and QP methods as predictive models in plant breeding. For a recent review, see Heslot et al. (2015), Desta & Ortiz (2014), Daetwyler et al. (2013) and Würschum (2012). We took a step further by using the same experimental data to compare prediction accuracies from both QP and GP methods. Both single-trait and multi-trait versions of the QP and GP methods were explored resulting into four prediction models. The predictive performances of the models were characterized using five of the pepper traits measured across four environments in the EU-SPICY project (Alimi et al., 2013a; Voorrips et al., 2010). These traits were subjected to a five-fold cross validation scheme with 10 repetitions.

The four methods differed substantially in their predictive abilities. Our results showed that GP methods outperformed QP methods in both single and multi-traits situations. This is not really surprising since parameters from GP methods are fitted on all available markers while parameters of QP methods are fitted only on selected QTL markers. Unlike QP methods, the GP methods fully take advantage of the correlations between all the markers and assigned prior distribution to marker effects so as to control shrinkage estimation. Two other single-trait GP methods, Bayesian Ridge Regression (BRR) (Gianola, 2013) and Bayesian Variable Selection (BVS) (Calus et al., 2008), were explored. They were not reported as they gave very similar results to BLR in all cases, even though these methods differ in their prior assumptions about marker effects. This result is similar to the pattern observed in literature for these GP methods, hence none of them could be said to be superior to others except in specific situations for example due to genetic architecture of a trait and experimental sample sizes (Daetwyler et al., 2013; De los Campos et al., 2013; Wimmer et al., 2013; Resende et al., 2012).

Furthermore, multi-trait models generally had better predictive power than single-trait models with MT-GP being superior. This showed that in situations where phenotypic data on a large number of traits have been collected (in multiple environments), using multivariate methods that properly model underlying variance-covariance (VCOV) structures among the traits and between environments would lead to improved power to detect more QTLs than performing individual trait/environment analyses (Alimi et al., 2013b; Xu, 2013; Jia and Jannink, 2012). The joint analysis was especially suitable for complex traits such as yield, whose variations are usually due to a large number of QTLs of small effects which might remain undetected in a univariate analysis. This is because sharing information among correlated traits helped to increase prediction accuracies for traits with hitherto low accuracies; see Stephens (2013). The results we obtained in NL trials, especially NL2, are however counter-intuitive as ST-QP gave higher prediction accuracies than MT-QP for many of the traits. This was due to the loss of some putative QTLs in the MT-QP model in NL trials. A number of QTL peaks were found to be just

below the threshold. Adjusting the default settings for those analyses may improve the performance of the MT-QP model in NL trials, but predictions will remain bad anyhow.

The second objective relates to prediction of the complex trait yield as a function of GEBVs of its component traits together with environmental variables using ecophysiological modelling. These types of crop growth modelling techniques have been widely employed to combine physiological traits with environmental inputs and study plant development over time (Uptmoor et al., 2008; Yin et al., 2005; Reymond et al., 2003). This approach was termed indirect prediction in contrast to predicting yield directly from its own QTL effects or breeding values. A simple LINTUL type crop growth function (Van Ittersum et al., 2003; Spitters and Schapendonk, 1990) was employed to relate yield to three component traits namely radiation use efficiency (RUE), partitioning into the fruits (PF) and growth rate of leaf area index (LAI_{rate}). The suitability of the crop growth function to correctly predict yield was evaluated by using the BLUEs for the component traits in the CGM. This gave predictive accuracies of almost one for yield in the four environments, suggesting that there is a relationship between yield and the three component traits and that the adopted crop growth function is capable of representing this relationship. This strategy was implemented in both within-environment and across-environment (GEI) analyses.

The most superior prediction model for generating the breeding values for the component traits remains the MT-GP. It is interesting to note that the CGM performed as well as the direct prediction strategy. We showed that yield in an environment can be successfully predicted from its component traits, provided a suitable function relating yield to the component traits is available. Also, the GEI CGM indicates that in situations where similarities exist among environments, we may use component traits and environmental information from one environment to predict yield in another environment. However, population type and the different pruning systems used in the two locations are primarily responsible for the marked differences in yield in NL and SP. The population is more suited for outdoor growing system than for the system in a greenhouse. The climate conditions (light, temperature) in Spain were more suited for this population than the NL conditions. Therefore, apart from the prediction model of choice, a suitable population and comparable management and environmental settings should be used.

To investigate the importance of the structure of the CGM for predicting the target trait yield from its components or GEBVs for components, we tried a simple linear regression model to relate the target trait to its component traits, without specifically including the environmental variables temperature and radiation. The prediction accuracies from this regression (results not reported) were quite similar to those of the CGM. This result indicated that specifically including environmental variables such as global radiation and temperature in the CGM did not confer added accuracy. A stepwise regression identified PF as the most important among the three component traits for yield predictions. This follows directly from the genetic correlations between yield and these three component traits. The fact that a linear regression performed very similarly to the CGM, should not be

seen as a drawback of the CGM we employed but rather due to the way some of the component traits were generated from yield, e.g. partitioning to fruits was calculated using yield and total biomass production. It would have been desirable to estimate the component traits as much as possible independent from the resulting target trait. For example, partitioning could be based on sink strengths (represented by the potential growth rate (Marcelis, 1996)), which can be measured independent of yield but have the disadvantage that they are difficult to measure. Also, the desire to make the crop model as simple as possible has likely reduced the model to being too empirical. In pepper, yield largely depends on fruit set, which is determined by many factors (Wubs et al., 2009). Simulating yield while taking into account fruit set would increase the number of component traits and not all of them are easy to determine. The balance should be between absolute empirical and mechanistic modelling (Yin and Struik, 2010).

Recently, efforts are being made to directly incorporate CGMs into the estimation of whole genome marker effects in GP using an Approximate Bayesian Computation (ABC) method (Technow et al., 2015). Technow et al., (2015) demonstrated the use of ABC as a mechanism for incorporating substantial biological knowledge embodied in the CGMs into a GP approach and showed that their proposed approach can be considerably more accurate than a benchmark GP method in predicting performance in environments represented in the estimation set as well as in previously unobserved environments for traits determined by non-additive gene effects. A key difference in our approach to that of Technow et al., (2015) is the way in which component traits are introduced. We assume that all the components are observable / measurable for all the genotypes. This is usually not true in practice especially with more sophisticated CGMs such as for example APSIM (Holzworth et al., 2014; Keating et al., 2003) that includes particularly difficult to measure root traits. The use of ABC allows handling difficult to measure component traits as hidden variables and thus facilitate incorporating them into GP models.

The use of the combined approach of QTL/genomic prediction and CGM still holds challenges including the relative weakness of current crop models in predicting differences in complex traits between genotypes that are members of segregating populations (Yin et al., 2005) and accumulation and propagation of errors (Uptmoor et al., 2008). However, despite the shortcomings, the combined approach can still show a high accuracy as the sources of error need not be statistically independent (Tardieu, 2003). To make a strong case for CGMs, the targeted complex traits should be defined as functions of as much as possible independent component traits, where these components traits themselves can still be measured with a certain ease or where these component traits may be approximated by other traits that can be recorded quickly and cheaply by automated phenotyping devices (Horgan et al., 2014; Van der Heijden et al., 2012).

In conclusion, we found that performing multi-trait analysis instead of single-trait analyses helps in improving the prediction of yield related traits. Also, genomic prediction methodologies performed better than QTL methodologies as prediction models. Combining multi-trait analysis with genomic prediction was shown to significantly

improve the prediction of all the traits considered in this study. The within environment direct and indirect prediction of yield through crop growth model using GEBVs from multi-trait genomic prediction method were found comparable although the indirect approach was heavily influenced by the high correlation between yield and PF. Prediction accuracies in the across-environment indirect prediction scenarios were lower than those from within-environment indirect prediction, but again multi-trait genomic prediction methods for GEBVs of component traits did better than single trait prediction methods.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211347. We thank the EU-SPICY Industrial Advisory Board for support and discussions.

CHAPTER 5

A network analysis of yield and yield components across environments: an example in pepper

CHAPTER 5

A network analysis of yield and yield components across environments: an example in pepper

5.1. Abstract

For efficient multi-trait QTL mapping it is important to take into account the interconnectedness among traits. The relations between traits can change between environments and then represent a special form genotype-by-environment interaction. For a suitable data set collected on pepper, we explore various network models for understanding the directed, say causal, relationships between the target trait yield and three component traits. We looked at the mutual dependencies between the traits and the dependencies of the traits on QTLs. Conditional and unconditional networks were constructed for the set of four traits across four different environments. For unconditional networks, we map QTLs given a particular fixed dependency structure between the traits by a standard multi-trait model (MTM). For conditional networks, the dependency structure between the traits and the QTLs affecting the traits are identified simultaneously by a QTL-driven phenotype network method (QTLnet). Inference for the final reconstructed networks was done by refitting the identified models as structural equation models. Conditioning QTL mapping on network structure via QTLnet clarified trait dependence on QTLs. QTLs with direct and indirect effects were distinguished, and QTL hotspots were resolved. The most probable conditional networks, with posterior probabilities ranging between 0.28 and 0.77 showed, as expected, yield as most downstream trait in all four environments. A complex target trait as yield can be studied and predicted via its component traits. The genetic improvements of yield would benefit from improvements on the component traits.

Keywords

Correlation network; Complex and Component traits; Conditional and Unconditional networks; Structural Equation Model; Quantitative Trait Loci.

5.2. Introduction

Important target traits in plant and animal genetic studies, such as fruit and milk yield, are complex as they result from intricate interactions of multiple genetic and environmental factors. Complex traits are often interpreted as integrations over time of underlying mutually interacting component traits and environmental inputs (Bustos Korts et al., 2016; van Eeuwijk, 2015; Chapman, 2008; Hammer et al., 2006). The totality of physiological interactions among target and component traits, together with shared genetic factors are responsible for observed associations among these traits (Li et al., 2006). Hence the genetic improvements of a complex target trait as yield could benefit from improvements on the component traits, especially when the mechanism of (causal) association between the target and component traits is known. To gain insight into the relationships among traits and between traits and QTLs, combined phenotypic and genetic network models have been proposed for use in a variety of contexts, see the review article by Valente et al. (2013). Network models can be regarded as an alternative to traditional multi-trait models (MTMs). Adding QTL information to the phenotypic network allows causal inference (Li et al., 2006; Schadt et al., 2005). Network models enable the differentiation of QTL effects into direct effects and indirect effects, i.e., QTL affect a trait either directly or via another trait. This will provide an intuitive explanation for potential QTL hotspots, where a QTL influences many traits as identified by multi trait QTL analyses (Alimi et al., 2013b).

In our earlier studies on yield in pepper, the relationships between yield and its component traits, such as radiation use efficiency (RUE), leaf area index (LAI) and partitioning to fruits (PF), were studied in the form of genotypic correlations (Alimi et al., 2013b) and by ecophysiological models (Alimi et al., 2016). Pleiotropic QTLs affecting yield and its component traits were identified. It is however anticipated that these traits exert causal effects on each other according to a particular pattern. Increasing yield in pepper may result from higher RUE and PF, while higher LAI may lead to an increase in RUE. Thus RUE and PF can have direct causal effects on yield, while LAI can have an indirect causal effect on yield through its effect on RUE. Traditional MTMs (Jiang and Zeng, 1995) do not model causal relationships among the traits and no mechanism behind pleiotropic effects of QTLs is to be revealed. For example, a pleiotropic QTL can act directly on a component trait that affects yield, making the same QTL affecting yield indirectly. The QTL can also act on both traits directly with or without the traits being correlated (Hageman et al., 2011; Li et al., 2006; Neto et al., 2010; Rosa et al., 2011).

Yield, a complex trait, exhibits strong genotype by environment interaction (GEI) and QTL by environment interaction (QEI) (Alimi et al., 2013b). Equally so, some of its component traits exhibit GEI (Alimi et al., 2013a). This means that any breeding strategy for improving yield or its components needs to be conditioned on the environmental conditions. The QTLs underlying yield (directly or through any of its component traits) might vary across environments qualitatively (changes in sign of QTL effects between environments) and/or quantitatively (changes in magnitude of QTLs) (Fournier-Level et al., 2013). The main issue would then be to assess the stability of QTLs for yield and its

component traits across environments so as to understand how selection can act on molecular variation in different environments.

In this study, two types of network models, called unconditional and conditional, were fitted across a number of environments. In the unconditional network, we revisit the multi-trait QTL (MT-QTL) analysis based on four yield related traits needed for the crop growth model (CGM) earlier reported in Alimi et al. (2016). This is done for the detection of the QTLs that will be included in the unconditional network. In the unconditional network, QTL information will not be updated in response to the inferred structure of the phenotype network. The unconditional network is simply a way to display MTMs graphically without updating the genetic architecture of the traits. In the conditional network (Hageman et al., 2011; Neto et al., 2010), the genetic architecture for each trait is inferred conditional on the phenotype network. Because the final phenotype network structure is itself unknown, the procedure iterates between updating the phenotypic network structure and the genetic architecture using a Markov chain Monte Carlo (MCMC) approach. The posterior samples from network models are summarized by Bayesian model averaging (Neto et al., 2010). The posterior probabilities for the set of probable networks are estimated and the network with the highest posterior probability is selected. The effect sizes and signs for the variables (traits and QTLs) in the final reconstructed networks corresponding to the four environmental conditions were compared by refitting the final networks in the form of structural equation models (SEMs). In the discussion of this paper, we first focus on the added values of using conditional network models over unconditional models. Secondly, we look at the interpretation of QTL hotspots found in MTMs. Then, the distinction between direct and indirect QTL effects is discussed, followed by a comparison of the final conditional networks for individual environments with each other and with the CGM topology for yield and its components.

5.3. Materials and Methods

5.3.1. Genotypic and Phenotypic data

A sixth generation segregating recombinant inbred lines (RIL) population of 149 individuals generated from a cross between Yolo Wonder (YW) and Criollo de Morelos 334 (CM 334) pepper cultivars were genotyped for 455 markers assembled onto 12 chromosomes covering 1705cM. The population was phenotyped in four experiments carried out at two locations (Netherlands, NL, and Spain, SP) during two time periods. This generated four environmental conditions denoted by NL1, NL2, SP1 and SP2. Additional information on genotyping and phenotyping of this population can be found in Nicolaï et al. (2012) and Alimi et al. (2013a). To study QEI for the two SP trials, an additional average condition was defined by the average of phenotypic measurements in SP1 and SP2: SPavg.

Four traits were studied that are central to a LINTUL-type (Light INTerception and Utilization) crop growth model (CGM) (Higashide and Heuvelink, 2009). Earlier analyses on those traits were reported in Alimi et al. (2016). These traits were yield, represented by

total dry weight of fruit (DWF), the sum of dry weight of all the fruits harvested during the growing season and the fruits on the plant at the final destructive harvest; the increase rate of leaf area index (LAI) which expresses mean increase in leaf area per unit time, where time is expressed in degree-days; radiation use efficiency (RUE) which is the dry matter production (g) per megajoule (MJ) of intercepted global radiation; and partitioning into fruit (PF) which expresses the proportion of total plant biomass due to fruit. The traits were preliminarily analysed to correct for non-genetic sources of variation and obtain best linear unbiased estimates (BLUEs) as genotypic means (Alimi et al., 2013a). The BLUEs were used in the network analyses.

5.3.2. Traits Relationships from Crop Growth Model

The LINTUL CGM employed in the SPICY project relates yield (DWF) to three component traits (Alimi et al., 2016). The model assumes yield can be predicted from these component traits via a mathematical representation that also includes environment specific variables such as temperature and radiation. This means yield is assumed to be (causally) related to LAI, RUE and PF.

For genotype *i* in environment *j*, the CGM was mathematically written as:

$$DWF_{ij} = PF_{ij} * RUE_{ij} * \sum_1^D \left(I_{jd} * \left(1 - e^{-k*LAI_{ijd}}\right)\right), \quad (5.1)$$

with DWF accumulated over the growing days $d = 1 \ldots D$, k is the extinction coefficient for the intercepted light and I is the daily global light intensity (Higashide and Heuvelink, 2009).

5.3.3. Unconditional Network (MTM)

The unconditional phenotype network was reconstructed by first performing QTL mapping to identify QTLs making up the genetic architectures of the traits. The unconditional phenotype network is simply a representation of a model for multi-trait QTL-mapping, where QTLs have directed effects at traits, whereas effects between traits are absent. The residuals of the traits can be correlated. This type of network is herein referred to as MTM network. For building an MTM network, multi-trait QTL (MT-QTL) analyses were carried out for each of the original four experimental environments and the constructed environment SPavg, for detection of QTL to be included in the network.

The MT-QTL model is a joint analysis of the four traits within each environment, y_{ik} with i for genotype ($i = 1 \ldots n_G$, with n_G being number of genotypes, 149) and k for trait ($k = 1 \ldots n_T$, with n_T being number of traits, 4) in a mixed model QTL analysis cf. Alimi et al. (2013b). To facilitate convergence, traits were standardized to have mean 0 and standard deviation 1. The MT-QTL model reads:

$$y_{ik} = \mu_k + \sum_{q \in Q} x_{iq}\beta_{kq} + g_{ik} + \varepsilon_{ik}, \quad (5.2)$$

where μ_k is the overall mean for trait k, x_{iq} is the genetic predictor for genotype i at genomic evaluation point q, β_{kq} is the trait-specific QTL effect for trait k corresponding to

the additive genetic predictor at locus q, g_{ik} represents the genetic effect of genotype *i* for trait k, and ε_{ik} represents a non-genetic/residual component that cannot be distinguished from g_{ik}. We will refer to the residual by g_{ik}. The set Q represents the full set of identified QTLs from an interval mapping followed by backward elimination. We assumed that the vectors $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{in_T})$ follow a multivariate normal distribution with zero mean and unstructured variance, $\boldsymbol{G}$, i.e. $\boldsymbol{g}_i \sim N(0, \boldsymbol{G})$ with dimensions n_T x n_T. Recall, that we standardized the traits, so that the diagonal of $\boldsymbol{G}$ will contain the proportion of the variance that was not explained by the QTLs. Model (5.2) accounts for genetic correlations between traits and allows us to detect pleiotropic QTLs. The MT-QTL analyses were performed using the QTL facilities in GenStat 15 (VSNi, 2012).

5.3.4. Conditional Network (QTLnet)

The second network building method considered is the QTL-driven phenotype network approach proposed by Neto et al. (2010). This type of network was termed QTLnet. QTLnet jointly models genetic architecture and phenotype network structure using so called homogeneous conditional Gaussian regression (HCGR) models (Lauritzen, 1996). This method is termed conditional network as the genetic architecture for each phenotype is inferred conditional on the phenotype network. The correlation structure among phenotypes is explicitly modelled according to the directed graph representation of the phenotype network. The genetic model is derived from a system of linear regression equations which corresponds to HCGR (Neto et al., 2010). In the HCGR model, the phenotypes ($\boldsymbol{y}$) are distributed according to a multivariate normal distribution conditional on the QTL genotypes ($\boldsymbol{q}$) which are subsets of the marker genotypes ($\boldsymbol{m}$), while the QTL $\boldsymbol{q}$ are modelled through the mean. Using Bayesian notation, the joint probability of $\boldsymbol{y}$ and $\boldsymbol{q}$ can thus be partitioned into genetic and recombination components, respectively relating phenotypes to QTL and QTL to observed markers across the genome:

$$p(\boldsymbol{y}, \boldsymbol{q}|\boldsymbol{m}) = p(\boldsymbol{y}|\boldsymbol{q}, \boldsymbol{m})p(\boldsymbol{q}|\boldsymbol{m}) = p(\boldsymbol{y}|\boldsymbol{q})p(\boldsymbol{q}|\boldsymbol{m}). \qquad (5.3)$$

The latter part of the equation follows from conditional independence since the marker genotypes provide no additional information about the phenotypes, given the QTL genotypes are already known.

For genotype *i* and trait *k*, the phenotype model can be represented as:

$$y_{ik} = \mu_k + \sum_{q \in Q} x_{iq}\beta_{kq} + \sum_{y_v \in pa(y_k)} \theta_{kv} y_{iv} + g_{ik} + \varepsilon_{ik}, \quad \varepsilon_{ik} \sim N(0, \sigma_k^2). \qquad (5.4)$$

Model (5.4) is an extended version of Model (5.2), where an extra term is added to represent the relations between the traits. The set of traits affecting trait *k* is denoted by its parent set $pa(y_k)$,,while θ_{kv} gives the partial regression coefficients of the traits in the parent set. For the residual term, g_{ik}, we again assume an unstructured model.

A Metropolis–Hastings (M-H) algorithm (Metropolis et al., 1953) as described in Husmeier (2003) was used to estimate the posterior probability in (5.3) starting from a non-informative prior for the skeleton of the network. As the algorithm makes single

changes (add or drop an edge, or change causal direction) to the phenotypes, such phenotypes in which the parent nodes have been altered are thus remapped. The accept/reject calculation involves estimation of the marginal likelihood conditional on the parent nodes and newly mapped QTL (Neto et al., 2010). Because the phenotype network structure is itself unknown, the algorithm iterates between updating the network structure and genetic architecture using a Markov chain Monte Carlo (MCMC) approach. The posterior sample of network structures is summarized by Bayesian model averaging (Hoeting et al., 1999). The averaged network is constructed by putting together all causal relationships with maximum posterior probability or with posterior probability above a predetermined threshold. When causal signal is high, the true model has the highest posterior probability. The conditional networks were estimated using QTLnet package in R (Neto et al., 2010; R-Core-Team, 2012).

5.3.5. Structural Equation Models

In order to gain further insight into the relationships among the traits, a structural equation model (SEM) (Wright, 1921) was employed to quantify effect sizes and directions for the variables (traits and QTL) in each of the final networks for each of the environments. The final network configurations from both MTM and QTLnet were translated into a SEM. In SEM, phenotypes can be treated as both predictor (exogenous) and response (endogenous) in a system of simultaneous equations, hence functional (causal) links between phenotypes can be established. QTLs can only be exogenous as it is already established that the association of a QTL and a phenotype is causal since QTL mapping is considered a randomized experiment and genotype precedes phenotypes (Li et al., 2006). The endogenous variables in a SEM are assumed to follow a multivariate normal distribution, while exogenous variables can be either continuous or categorical as with the QTL in our case.

The quantities of importance in the SEM analysis were the goodness of fit as judged by the Akaike Information Criterion (AIC), and the path coefficient estimates. Standard errors were computed for each estimated path coefficient, and equality to zero is tested using a z-statistic. After the individual relationships within the model were assessed, path coefficients between the unconditional and conditional networks were compared. The final SEM models for MTM and QTLnet in each environment were compared on AIC. For individual traits in the SEM the R^2 statistic was calculated. SEM analyses were performed using the CALIS procedure in SAS 9.3 (SAS-Institute, 2011).

Finally, based on SEM path coefficients, the net effects for each of the traits were estimated and compared across environments and between network methods. The net effect of the QTL on a trait is the sum of the effects along all the direct and indirect paths connecting the two variables. For standardized variables, the effect of an indirect path is the product of all the path coefficients (including + and - signs) along this path. (Li et al., 2006).

5.4. Results

5.4.1. Genotypic and Phenotypic data

The broad-sense heritability (H^2), phenotypic mean and standard deviation for the four yield related traits of interest across the four environments are presented in Table 5.1. Mean trait values varied across environments with DWF showing highest variation. Mean DWF in the Dutch environments were lower than in the Spanish environments. Yield in NL2 were particularly low due to very short growing period and blossom end rot infestation (Alimi et al., 2013a). The trait H^2 varied from 0.49 – 0.95 with an average of 0.82 and were slightly higher in the SP environments than the NL environments. Among the four traits, H^2s for radiation use efficiency were lowest across all environments. For each trait, the consistent and high values for the estimates of H^2 across environments reflect low micro-environmental disturbances in the traits and good possibilities for mapping QTLs governing these traits.

Table 5.1: Traits Summary Statistics

	Heritability (H^2)				Mean				Standard Deviation			
Trait	NL1	NL2	SP1	SP2	NL1	NL2	SP1	SP2	NL1	NL2	SP1	SP2
DWF	0.82	0.76	0.89	0.87	23.12	7.77	66.89	87.13	16.77	6.40	36.68	32.77
LAI	0.67	0.88	0.90	0.94	2.22	3.01	1.56	1.69	0.70	1.00	0.49	0.44
RUE	0.49	0.73	0.73	0.75	1.12	1.00	0.96	1.19	0.24	0.20	0.14	0.14
PF	0.89	0.89	0.95	0.93	0.18	0.10	0.28	0.36	0.12	0.07	0.14	0.11

Table 5.2: Genetic correlations of traits within each environment

	DWF	LAI	RUE		Trait	DWF	LAI	RUE
NL1					SP1			
LAI	0.07				LAI	0.09		
RUE	0.13	0.11			RUE	0.40	-0.34	
PF	0.90	-0.18	-0.11		PF	0.93	-0.18	0.34
NL2					SP2			
LAI	0.19				LAI	-0.04		
RUE	0.26	0.32			RUE	0.22	0.01	
PF	0.91	-0.07	-0.03		PF	0.89	-0.37	0.03

The within environment correlations between traits were mostly consistent in sign and magnitude across the environments (Table 5.2). The correlations between DWF and the other three traits are of primary importance in this study. Very high correlations were observed between DWF and PF (0.89 – 0.93). This was not surprising since PF was computed from total fruit weight and total plant biomass. RUE showed low but positive correlations to DWF (0.13 – 0.40), while LAI displayed very low and sometimes negative correlations to DWF (-0.04 – 0.19). This suggests that LAI possibly shares little or no genetic components with DWF while PF should share almost all genetic architecture with DWF.

5.4.2. Unconditional Network (MTM)

The QTLs for the unconditional phenotype networks were obtained by MT-QTL analyses that combined the four traits in each environment (Table 5.3). Figure 5.1 shows the

detected QTLs for all traits and environments. Many of the detected QTLs for each of the traits are part of the QTLs earlier reported for these traits in Alimi et al. (2013b). The addition of an extra QTL or loss of some QTL earlier reported is due to the different number of trait used in the two studies. A total of one, zero, four, five and six QTL were detected for DWF in NL1, NL2, SP1, SP2 and SPavg respectively. The NL trials gave the smallest number of QTLs for this trait, as expected (since these had the lowest H^2 and mean phenotypic values for DWF). Many of the significant QTLs are pleiotropic with some pleiotropic effects being antagonistic. Examples of QTLs with antagonistic effects are the QTLs on chromosomes 6 and 7. These QTLs showed increasing effects from parent YW on some traits (e.g. DWF and PF) and increasing effects from parent CM334 on other traits (e.g. LAI and RUE). The QTLs also showed mainly quantitative QTL-by-environment interactions. All the QTLs picked up simultaneously in both SP1 and SP2 are also picked up for SPavg. Some QTLs detected in either SP1 or SP2 were also picked up in SPavg. None of the QTLs in SPavg showed qualitative QEI with either SP1 or SP2.

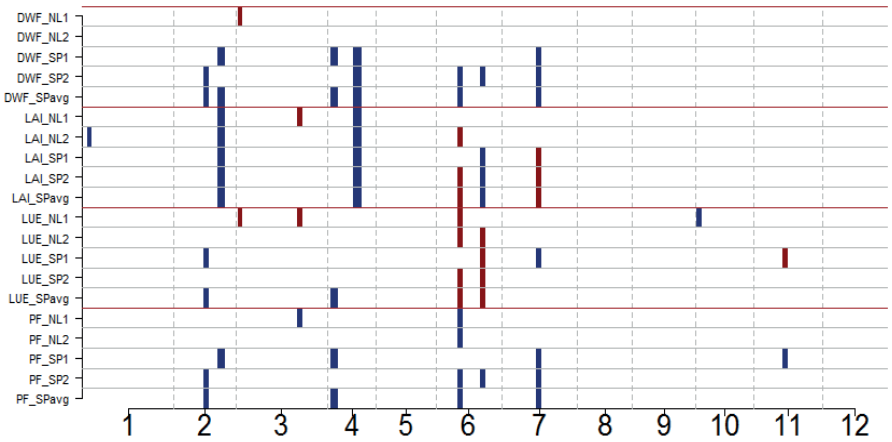


Figure 5.1 Detected QTL from MT-QTL analyses in the five environments. Blue indicates QTL with significant effect from YW allele while red indicates QTL with significant effect from CM334 allele. The 12 pepper chromosomes are on the X-axis while trait-environment combinations (e.g. PF_SP2 represents trait PF in environment SP2) are given on the Y-axis

Table 5.3 Trait-specific QTL effects estimates (Est) with their standard errors (SE) for MT-QTL analysis. QTL are denoted by their chromosome number and position e.g. Q2p105 represents a QTL on chromosome 2 at around 105cM position. Negative QTL effects mean that the YW allele gives higher trait values than the CM334 allele, and positive QTL effects mean that the CM334 allele gives higher trait values. The underlined values are significant QTL effects.

Environments	Predictors	DWF		LAI		RUE		PF	
		Est	SE	Est	SE	Est	SE	Est	SE
NL1	Q2p105	-0.089	0.082	-0.339	0.074	-0.090	0.070	-0.010	0.082
	Q3p0	0.186	0.083	-0.013	0.074	0.321	0.071	0.146	0.083
	Q3p135	-0.143	0.084	0.153	0.075	0.277	0.071	-0.206	0.084
	Q4p53	-0.139	0.085	-0.344	0.076	-0.041	0.073	0.007	0.085
	Q6p51	-0.103	0.084	0.112	0.075	0.245	0.072	-0.187	0.084
	Q10p7	-0.030	0.084	0.108	0.075	-0.215	0.072	-0.039	0.084

NL2	Q1p15	0.005	0.085	-0.371	0.064	-0.036	0.080	0.082	0.084
	Q2p105	-0.140	0.083	-0.361	0.062	-0.041	0.078	-0.046	0.082
	Q4p53	-0.102	0.085	-0.351	0.064	-0.018	0.080	-0.027	0.085
	Q6p51	-0.128	0.084	0.137	0.063	0.338	0.079	-0.222	0.083
	Q6p107	-0.095	0.084	-0.077	0.063	0.181	0.079	-0.123	0.083
SP1	Q2p75	-0.137	0.072	-0.078	0.075	-0.186	0.079	-0.055	0.072
	Q2p105	-0.289	0.074	-0.304	0.077	-0.026	0.082	-0.222	0.073
	Q4p19	-0.290	0.072	0.042	0.075	-0.087	0.079	-0.309	0.071
	Q4p53	-0.170	0.074	-0.244	0.078	0.065	0.082	-0.105	0.074
	Q6p107	-0.050	0.070	-0.157	0.073	0.326	0.078	-0.079	0.070
	Q7p78	-0.237	0.071	0.266	0.075	-0.157	0.079	-0.318	0.071
	Q11p79	-0.112	0.068	0.074	0.072	0.164	0.076	-0.233	0.068
SP2	Q2p75	-0.181	0.073	-0.001	0.067	-0.126	0.080	-0.206	0.075
	Q2p105	-0.135	0.074	-0.459	0.068	-0.071	0.081	0.049	0.077
	Q4p53	-0.269	0.074	-0.252	0.069	-0.028	0.082	-0.140	0.077
	Q6p51	-0.203	0.072	0.177	0.066	0.283	0.079	-0.277	0.074
	Q6p107	-0.194	0.071	-0.168	0.066	0.234	0.078	-0.172	0.074
	Q7p78	-0.297	0.073	0.180	0.068	0.029	0.081	-0.319	0.076
Spavg	Q2p75	-0.279	0.065	-0.072	0.073	-0.188	0.079	-0.243	0.069
	Q2p105	-0.220	0.067	-0.343	0.074	-0.086	0.080	-0.090	0.070
	Q4p19	-0.300	0.068	0.126	0.075	-0.216	0.081	-0.327	0.071
	Q4p53	-0.193	0.069	-0.279	0.076	0.096	0.083	-0.086	0.072
	Q6p51	-0.257	0.062	0.171	0.070	0.160	0.075	-0.324	0.065
	Q6p107	-0.064	0.063	-0.186	0.070	0.363	0.075	-0.079	0.065
	Q7p78	-0.266	0.063	0.253	0.070	-0.061	0.076	-0.339	0.066

5.4.3. Conditional Network

In Table 5.4, the three most probable conditional networks in each environment are given together with their posterior probabilities. The first conditional networks for Spanish environments are by far superior to their alternative networks with posterior probability > 0.6, In NL trials the posterior probabilities were relatively low. In NL2, the two most probable networks had almost equal posterior probabilities (0.36 and 0.34). The major difference in these two networks is that PF became the most downstream trait in the second network while DWF was the most downstream trait in the first network.

The most probable conditional network and the unconditional network in the SP2 environment are displayed in Figure 5.2 for comparison. The conditional networks for the remaining environments are given in Appendix 5.1. The conditional network with the highest posterior probability is reported in each case. In all environments, the phenotype networks from QTLnet have yield as most downstream trait, similar to what happens in the CGM for yield. It is noteworthy that in the QTLnet specification, when QTL detection threshold changes, the network configuration also changes slightly with PF becoming the most downstream trait instead of DWF especially in NL2. This can be explained by the very high correlation and shared genetic components between these two traits.

QTLnet improved elucidates the relationships between phenotypic traits and distinguishes between QTL with direct and indirect effects, and thereby helping to resolve the causal structure behind QTL hotspots. For instance, in SP2, MTM reported six QTL with five of them having pleiotropic and direct effects. QTLnet revealed only one QTL with pleiotropic effect and six QTL with direct and indirect effects. As an example, MTM revealed that Q7p78 had a pleiotropic effect on DWF, LAI and PF (hotspot) but QTLnet showed that Q7p78 had a direct effect only on PF through which it affected LAI and DWF indirectly. In SP2, the only QTL with a pleiotropic effect as detected by QTLnet was Q6p51 that was found to have a direct effect on both RUE and PF. No QTL directly influenced DWF in SP2. All the effects on DWF were found to be indirect through its upstream component traits. Using QTLnet, two additional QTL (Q4p19 and Q11p79) were detected for RUE while one QTL from MTM (Q6p107) disappeared. These two additional QTL are however not new as they were earlier reported for this trait in Alimi et al. (2013b).

Table 5.4: Final model assessment based on comparison of three most probable conditional networks in each environment.

Trial	Network[#]	Posterior Probability
NL1	(1|2,3,4)(2)(3|2)(4)	0.284
	(1|2,3,4)(2|3)(3)(4)	0.085
	(1|2,3,4)(2|4)(3|2)(4)	0.059
NL2	(1|2,3,4)(2|3)(3)(4)	0.364
	(1)(2|3)(3|1)(4|1,2,3)	0.341
	(1|2,3,4)(2|3)(3|4)(4)	0.091
SP1	(1|2,3,4)(2|3,4)(3|4)(4)	0.745
	(1|2,3,4)(2|3)(3|4)(4)	0.254
	(1|2,3,4)(2|4)(3|4)(4)	0.001
SP2	(1|2,3,4) (2|4) (3) (4|3)	0.768
	(1|2,3,4)(2|3,4)(3)(4|3)	0.140
	(1|2,3,4) (2|4) (3) (4)	0.073
SPavg	(1|2,3,4)(2|3,4)(3|4)(4)	0.623
	(1|2,3,4)(2|4)(3|2)(4)	0.134
	(1|2,4)(2|4)(3|1,2,4)(4)	0.133

[#]1 = DWF; 2 = LAI; 3 = RUE; and 4 = PF.

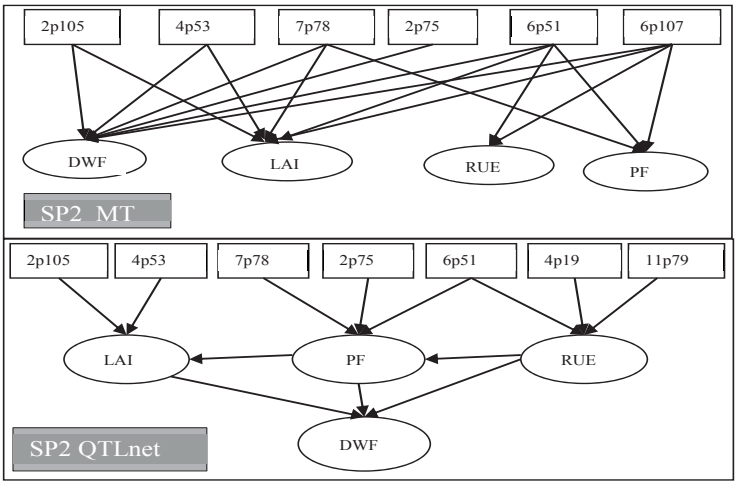


Figure 5.2 Unconditional (SP2_MTM) and Conditional (SP2 QTLnet) networks in SP2: Graphical representation of the relationships among traits and the QTLs influencing the traits in the SP2 environment. QTL are denoted by their chromosome number and position e.g. 2p105 represents a QTL on chromosome 2 at around 105cM position. Conditional networks for the remaining environments are in appendix 1

5.4.4. Structural Equation Models

The standardized path coefficients for both conditional and unconditional networks in SP2 obtained via SEM are presented in Table 5.5. The path coefficients for the remaining environments are given in Appendix 2. In all environments, QTL effect directions in both conditional and unconditional networks were defined to go from QTL to traits. All QTL paths in conditional networks as constructed by QTLnet showed significant path

coefficients. In contrast, some paths in MTM contained non-significant coefficients when retested in SEM, e.g. Q3p0 and Q4p53 for DWF in NL1 and SP1 respectively. This difference between SEM and MTM can be caused by the difference in likelihood between MTM (QTLs fixed, polygenic effects random, REML) and SEM (everything random, ML).

From the AIC values, QTLnet models fitted better in NL1 and NL2, while the MTM models fitted better in SP1 and Spavg. In SP2 there was little difference. In SP2, based on squared correlations, R^2, between fitted and observed values for the BLUEs, the SEM fitted the data fairly well. The R^2 estimates from the SEM for individual traits in each environment were mostly higher for QTLnet than for the MTM. As an example, the SEM for DWF in SP2 had R^2 of 0.31 and 0.93 for MTM and QTLnet networks, respectively. However, this was a direct consequence of upstream traits being included in the models for downstream traits in QTLnet networks.

Across environments, the QTL path coefficients mainly showed quantitative QEI i.e. difference in effect magnitude but not in direction of effect. An example is the direct influence of Q2p105 on LAI across the four environments. The coefficients ranged from strongly negative in SP2 (-0.45) to mildly negative in NL1 (-0.28). The only QTL with crossover QEI is Q11p79. This QTL negatively influenced RUE in SP2 (-0.177) but showed positive influence on RUE in SP1 (0.377). This QTL however disappeared in SPavg. Four QTLs showed slightly modified quantitative QEI in SP environments. These include Q4p19 that was found to influence PF in SP1 and SPavg, but it influenced RUE in SP2, and Q7p78 that was found to influence PF in SP2 and SPavg, but LAI in SP1. Others include Q6p51 and Q6p107.

The SEM results for the conditional network also showed that DWF was significantly influenced by PF, LAI and RUE in all environments. This is in agreement with the CGM specification. This confirms that DWF is truly downstream to these component traits and can be predicted from these traits given that a true relationship is established.

SEM allows us to resolve effects' sign and magnitude for any trait along its direct and indirect paths. The net effects reported for MTM are made up of only direct paths while the reported net effects for QTLnet take into account both direct and indirect paths (Table 5.6). The net effects for some of the traits changed markedly when a conditional network is used as against an unconditional network (e.g. net effect on PF in SP2 changed by a value of 0.448) while others only changed slightly (e.g. net effect on DWF in NL1 changed by a value of 0.007). Generally speaking, trait net effects increased when a conditional network is used. Across most environments, net effects for DWF increased when direct and indirect paths leading to it are properly resolved. Changes of the net effects in their dependence on environmental conditions provide us with a measure for environmental stability of traits. Net effect for fruit size was substantial more in the Mediterranean climate (SP trials) than in the more temperate climate (NL trials). On the contrary, in the NL trials, we observed slightly higher net effect for leaf formation (LAI) than in the SP trials, hence reduction in fruit sizes.

Table 5.5 Standardized SEM path coefficients for MTM and QTLnet networks for SP2. QTL are denoted by their chromosome number and position e.g. Q2p75 represents a QTL on chromosome 2 at around 75cM position. Negative path coefficients indicate that the YW allele is associated with higher trait values and positive path coefficients mean that the CM334 allele gives higher trait values.

SP2									
Methods	Predictors	DWF		LAI		RUE		PF	
		Est	SE	Est	SE	Est	SE	Est	SE
MTM	Q2p75	-0.286	0.060					-0.313	0.057
	Q2p105			-0.210	0.046				
	Q4p53	-0.175	0.033	-0.335	0.059				
	Q6p51	-0.182	0.068	0.160	0.070	0.297	0.073	-0.243	0.067
	Q6p107	-0.203	0.067	-0.148	0.069	0.229	0.074	-0.183	0.068
	Q7p78	-0.299	0.062	0.176	0.069			-0.311	0.066
	R^2	0.313		0.287		0.151		0.312	
	AIC = 155.09								
QTLnet	LAI	0.397	0.044						
	RUE	0.148	0.054					0.973	0.340
	PF	1.079	0.045	-0.407	0.101				
	Q2p75							-0.217	0.067
	Q2p105			-0.453	0.055				
	Q4p19					-0.197	0.063		
	Q4p53			-0.321	0.056				
	Q6p51					0.287	0.072	-0.586	0.148
	Q7p78							-0.297	0.065
	Q11p79					-0.177	0.061		
	R^2	0.925		0.512		0.182		0.415	
	AIC = 155.17								

Table 5.6: The net effects (direct + indirect) for each trait from SEM models. The effect of a direct path is the standardized path coefficient and that of an indirect path is the product of the path coefficients (including the sign) along that path. Negative net effect means that the YW allele gives higher trait values than the CM334 allele, and positive net effect means that the CM334 allele gives higher trait values.

Traits	NL1		NL2		SP1		SP2		SPavg	
	MTM	QTLnet	MTM	QTLnet	MTM	QTLnet	MTM	QTLnet	MTM	QTLnet
DWF	-0.068	-0.075	0	-0.137	-0.937	-0.730	-1.145	-1.407	-1.289	-1.238
LAI	-0.370	-0.503	-0.743	-0.894	-0.349	-0.382	-0.377	-0.292	-0.545	-0.530
RUE	0.557	0.283	0.538	0.390	0.414	0.160	0.526	-0.087	0.305	0.521
PF	-0.129	0	-0.120	0	-1.000	-0.717	-0.737	-1.185	-1.077	-1.014

5.5. Discussion

The objective of this study was to explore network models for yield and yield components for a unique pepper data set that allowed us to compare reconstructed networks across various environmental conditions. Both conditional and unconditional networks were constructed for four traits. The unconditional networks were based on a standard multi trait model (MTM) (Jiang and Zeng, 1995), while the conditional networks were based on the QTL-driven phenotype network method (QTLnet) developed by Neto et al. (2010). The final networks for each environment from both conditional and unconditional methods were used in a SEM to quantify and compare the relationships among yield and its

components. QTLnet improved detection of refined genetic architecture by distinguishing between QTLs with direct and indirect effects, thereby resolving QTL hotspots.

A number of recent papers stress the need to study networks of plant phenotypes, and the stability of phenotype relationships across environments due to genotype-by-environment interactions (Granier and Vile, 2014; Li et al., 2010; Valente et al., 2013). This need is even more pronounced with the advent of automated high-throughput phenotyping technologies capable of simultaneously recording many traits (Barócsi, 2012; Van der Heijden et al., 2012) that may exhibit moderate to strong correlations between them. Also, characterising genotypes by a single phenotype is inefficient since many QTLs influence multiple traits; hence there exists a strong need to properly account for covariations among the traits and resolving the type/direction of relationships among the traits. Our results showed that although MTMs are able to account for covariation among traits and establish QTL with pleiotropic effects, they miss out on the possibility to disentangle the paths for such effects and neither are they able to provide insight into the (causal) relationships among the traits. When the correlation between two phenotypes arises exclusively because of a pleiotropic QTL, conditioning on the QTL genotypes makes the phenotypes independent. Properly conditioning QTL analysis on network structure increased detection of refined genetic architecture as shown for DWF in SP2 for example. No QTL directly influences DWF in SP2. All the effects on DWF were found to be indirect through its upstream component traits. Conditioning QTL mapping on network structure disentangles QTL effects into direct and indirect effects. We further showed that the QTL hotspots from the MTMs resulted from ignoring network structure for the correlated traits.

The most probable conditional networks from the four environments contained DWF as the most downstream trait, as expected from the structure for the ecophysiological LINTUL-type (Light INTerception and Utilization) genotype-to-phenotype model (CGM) (Spitters and Schapendonk, 1990; Van Ittersum et al., 2003). The CGM simulates the formation of pepper yield under potential growing conditions and relates yield to its component traits. The main environmental factors considered in the CGM were radiation and temperature. These types of genotype-to-phenotype modelling techniques have been widely employed to study traits relationships in combination with environmental inputs and also to study plant development over time (Reymond et al., 2003; Uptmoor et al., 2008; Yin et al., 2005). Similar to the CGM topology, yield was established to be downstream to its three component traits, indicating that yield can be predicted from the component traits (Alimi et al., 2016). However, the relationship among the component traits as reconstructed by our statistical analysis differs from the relationships as defined by the CGM. The network model did not incorporate any environmental information. Including environmental characterizations via CGM into the network model may improve our ability to improve on yield prediction through its component traits.

When the aim of a QTL analysis that is part of breeding program is restricted to selection based on breeding values and estimation of the response to selection, using an MTM is sufficient as the interest is on the total additive genetic effects. Even if traits are indeed

causally related, no information is lost. However, where changes due to genetic and/or environmental interventions are of interest, using an MTM provides only limited information for such predictions. Examples of such interventions may be changes in temperature or other environmental/agronomic inputs or even modification of gene expression levels. The use of a network-based model will allow prediction of outcomes to interventions applied to such a network (Scutari et al., 2014; Bouwman et al., 2014; Valente et al., 2013). This is possible since the relationships depicted by conditional networks are direct, thus any change due to external interventions can be quantified. As an example, using a conditional network model, we can predict changes in our target complex trait yield across the different environments by adding environmental information to the network relating the component and target traits. This is however only feasible if the component traits have simple genetic architecture with little or no GEI (Bustos-Korts et al., 2016). Such predictions are made by representing the intervention on the causal structure among phenotypes and by knowing the genetic effect directly on each trait, as well as the dispersion parameters that describe their joint distribution. All these are possible by fitting a SEM.

Causality claims in genomic network studies stem from two facts. First, there is the analogy between randomized experimental design and genetic randomization that occurs during meiosis and secondly, the intuition that phenotypic variation is caused by genetic factors (Li et al., 2006; Neto et al., 2010). Correlation between traits is insufficient for claims for causality, even when the traits share a common QTL. Understanding of the biology governing the relationships is crucial. It is possible that two traits sharing a common QTL are actually independent (Li et al., 2010). Although our results indicate that yield in pepper can be studied and predicted via its component traits, an in-depth biological understanding of the relationship is needed before categorical claim of causality between yield and the three component traits can be made. Notwithstanding this shortcoming, our results demonstrated that properly conditioning the genetic architecture of a complex trait on the causal structure of the component traits will enhance our ability to correctly predict the complex trait. Thus, the genetic improvements of the complex trait would benefit from improvements on the component traits.

Acknowledgements

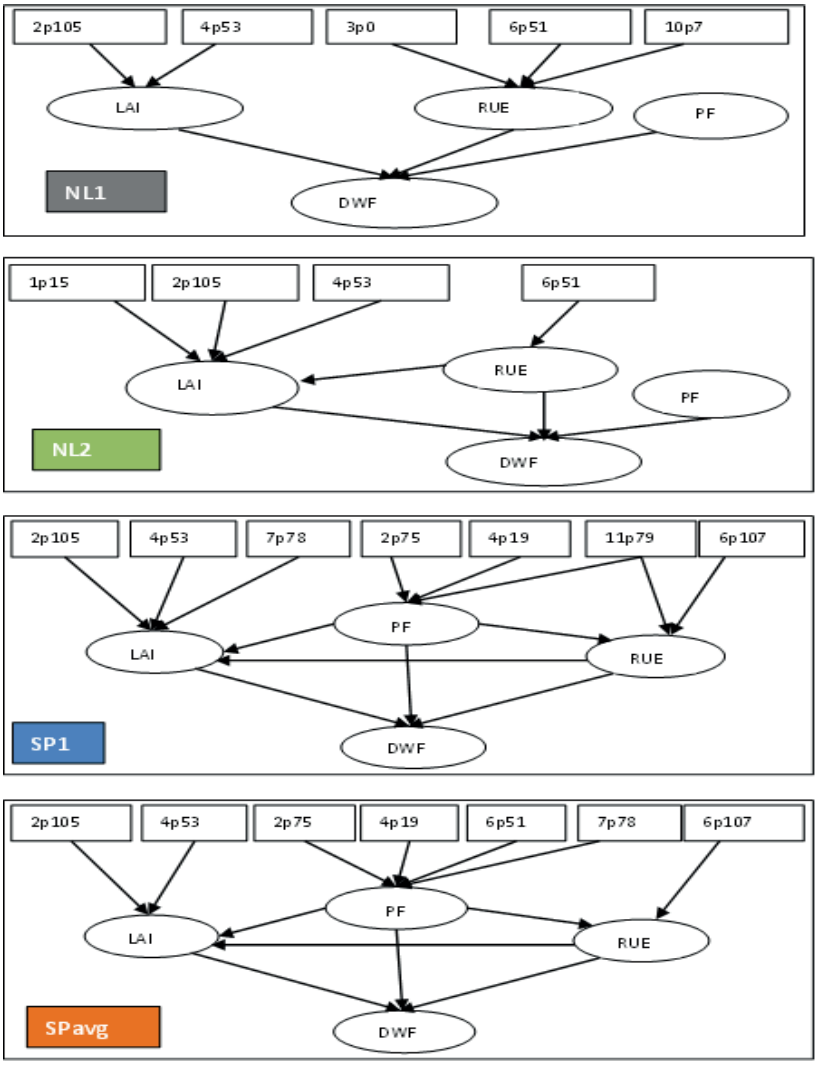
The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211347. We thank the EU-SPICY Industrial Advisory Board for support and discussions.

Key Message

Both conditional and unconditional correlation networks were constructed to study relationships among yield and yield related traits across multiple environments. A conditional network considerably improved detection of refined genetic architecture by distinguishing between QTL with direct and indirect effects thereby resolving pleiotropic QTLs.

Appendix 5A

Figure 5A1. Graphical representation of the conditional relationships among traits and the QTLs influencing the traits. All the paths presented are significant with P-val<0.05.



Appendix 5B

Table 5B1: Standardized path coefficients for MTM and QTLnet networks. QTL are denoted by Q followed by the chromosome number and p followed by cM positions. Negative effects in the QTL indicate that the YW allele is associated with higher trait values.

NL1									
		DWF		LAI		RUE		PF	
Methods	Predictors	Est	SE	Est	SE	Est	SE	Est	SE
MTM	Q2p105			-0.251	0.067				
	Q3p0	0.038	0.030			0.300	0.059		
	Q3p135			0.130	0.077	0.273	0.063	-0.059	0.032
	Q4p53	-0.106	0.030	-0.249	0.069				
	Q6p51					0.230	0.060	-0.070	0.030
	Q10p7					-0.246	0.058		
	AIC = 103.76								
QTLnet	LAI	0.219	0.102						
	RUE	0.124	0.053						
	PF	0.903	0.017						
	Q2p105			-0.280	0.067				
	Q3p0					0.336	0.061		
	Q4p53			-0.223	0.068				
	Q6p51					0.213	0.064		
	Q10p7					-0.266	0.064		
	AIC = 84.37								

NL2									
Methods	Predictors	DWF		LAI		RUE		PF	
		Est	SE	Est	SE	Est	SE	Est	SE
MTM	Q1p15			-0.335	0.056				
	Q2p105			-0.305	0.056				
	Q4p53			-0.241	0.058				
	Q6p51			0.138	0.066	0.393	0.064	-0.120	0.033
	Q6p107					0.145	0.058		
	AIC = 115.09								
QTLnet	LAI	0.213	0.038						
	RUE	0.137	0.061	0.247	0.149				
	PF	0.909	0.022						
	Q1p15			-0.354	0.055				
	Q2p105			-0.368	0.055				
	Q4p53			-0.268	0.057				
	Q6p51					0.390	0.069		
	AIC = 73.2378								

SP1									
Methods	Predictors	DWF		LAI		RUE		PF	
	Predictors	Est	SE	Est	SE	Est	SE	Est	SE
MTM	Q2p75					-0.101	0.056		
	Q2p105	-0.394	0.061	-0.296	0.066			-0.314	0.064
	Q4p19	-0.286	0.061					-0.304	0.063
	Q4p53	-0.055	0.029	-0.159	0.067				
	Q6p107			-0.161	0.051	0.302	0.066		
	Q7p78	-0.202	0.068	0.267	0.070	-0.148	0.074	-0.281	0.068
	Q11p79					0.260	0.062	-0.101	0.022
	AIC = 136.74								
QTLnet	LAI	0.323	0.046						
	RUE	0.274	0.044	-0.111	0.170				
	PF	0.907	0.045	0.038	0.169	0.675	0.170		
	Q2p75							-0.319	0.066
	Q2p105			-0.344	0.066				
	Q4p19							-0.209	0.066
	Q4p53			-0.169	0.069				
	Q6p107					0.267	0.064		
	Q7p78			0.176	0.066				
	Q11p79					0.377	0.080	-0.189	0.071
	AIC = 145.91								

SPavg									
		DWF		LAI		RUE		PF	
Methods	Predictors	Est	SE	Est	SE	Est	SE	Est	SE
MTM	Q2p75	-0.226	0.063			-0.140	0.072	-0.258	0.058
	Q2p105	-0.170	0.031	-0.422	0.053				
	Q4p19	-0.242	0.064			-0.114	0.073	-0.219	0.059
	Q4p53	-0.136	0.030	-0.251	0.055				
	Q6p51	-0.229	0.066	0.162	0.064	0.268	0.071	-0.292	0.067
	Q6p107			-0.202	0.046	0.291	0.061		
	Q7p78	-0.286	0.060	0.168	0.063			-0.308	0.065
	AIC = 127.39								
QTLnet	LAI	0.406	0.046						
	RUE	0.180	0.086	-0.738	0.260				
	PF	1.101	0.052	-0.572	0.154	-0.229	0.146		
	Q2p75							-0.158	0.063
	Q2p105			-0.426	0.052				
	Q4p19							-0.189	0.063
	Q4p53			-0.300	0.053				
	Q6p51							-0.362	0.062
	Q6p107					0.289	0.074		
	Q7p78							-0.305	0.062
	AIC = 157.19								

CHAPTER 6

GENERAL DISCUSSION

Complex Trait Predictions

CHAPTER 6

GENERAL DISCUSSION

Complex Trait Predictions

6.1. Introduction

The central theme of the research presented in this thesis revolves around the development of prediction models for complex target traits, where yield grown in different environments serves as an example case. For complex traits, a desired situation would be for the breeder to test promising candidate varieties in several conditions with the aim of selecting the best genotypes, either for a wide or otherwise narrower range of environments or conditions. This would however be expensive, laborious and time-consuming (Montes et al., 2007). The breeder therefore needs a set of tools that supports her/his ability to predict phenotypic responses of genotypes for (complex) traits under a range of environmental conditions. An important set of tools is given by prediction models capable of taking into account the heterogeneity of genetic variances and correlations that underlies genotype by environment interaction and properly accounting for both generic and specific features of experimental designs being employed. Various strategies were tried: multi-trait multi-environment (MTME) analysis as the most general form of linear mixed model (LMM), crop growth model (CGM) and causal network models. Yield being a complex trait, is based on multiple QTLs/genes with small effects that interact between themselves and with the environment. So, we expect that yield itself will be difficult to predict. As an alternative, we may try predicting yield using component traits. These component traits can enter a multi-trait analysis together with yield (chapters 2 and 3), or the components enter the CGM to predict yield together with environmental information (chapter 4), or they enter a causal network where their relationship with yield is determined (chapter 5). An approach that predicts yield via a dissection in component traits may work when the components are less sensitive to genotype-by-environment interaction (GxE) than yield itself and when they have a simpler genetic basis, i.e., only a few QTLs with large effects.

In the EU-SPICY project (which this thesis is a part of), new tools allowing automated and fast high-throughput phenotyping (HTP) were also developed, leading to reduction in the amount of manual labour expended on phenotyping experiments and contributing to cost reduction in the long run. Although the target trait yield could not be measured directly by the HTP, some of the HTP traits can be used as correlated traits for especially yield components. Since this work is part of the EU-SPICY project, some reflections on salient aspects of the project that directly influence the central theme of this thesis are presented. These include the choice of parents, type and size of the mapping population, type and size

of marker data and phenotype measurement protocols. Before discussing statistical issues related to different types of prediction models, we will touch on these other aspects that affected the performance of the prediction models.

6.2. Mapping Population

Among other factors, the choice of promising segregating populations for QTL mapping can be based on selecting phenotypically diverse parents (Hung et al., 2012). The planting material used in the EU-SPICY project was the progeny obtained from a biparental cross between two phenotypically distant pepper inbred cultivars of *Capsicum annuum*; Yolo Wonder (YW) as the female parent and CM334 as the male parent (Figure 6.1). These cultivars differed in their plant and fruit phenotypes. Yolo Wonder is a domesticated large-fruited sweet cultivar while CM334 is a wild small-fruited pungent cultivar. A total of 297 recombinant inbred lines (RIL) were obtained after 6 to 7 cycles of successive self-pollinations using the single seed descent method. The lines represent a random and large sample of all the possible descendants issued from the initial cross as there was no chromosome segment with segregation distortion. The relevance of this choice of parents in the prediction of the target trait yield can be shown by the type, magnitude and precision of the QTLs detected for yield and its component traits and the performance of the genomic prediction models. On average, the detected QTLs for yield explained about 47% genetic variance in each of the four environments (NL1, NL2, SP1 and SP2), similar to the average explained genetic variance for each of the three component traits (LAI, PF and RUE).

The sample size that we used is comparable to sample sizes used in literature for mapping QTLs for pepper and similar vegetables. For instance, Chaim et al., (2001) and Rao et al., (2003) used a population of 180 and 248 individuals respectively to map yield-related QTL in pepper. Nunome et al., (2001) used a mapping population size of 168 individuals to map fruit shape and colour development traits in eggplant. Similar sample sizes have been used in tomato (Causse et al., 2002; deVicente M. C. & Tanksley). So, we can conclude that this population allowed us to map QTLs for yield and its component traits.

However, the QTLs identified in offspring from crosses of extreme parents as used here may be of limited interest to breeders. For example, QTLs segregating in the offspring from a cross between a domesticated and an exotic parent may have most yield increasing QTL alleles coming from the domesticated parent, while most yield decreasing QTL alleles come from the exotic parent. In that situation, little progress is possible beyond the level of the domesticated parent, so no or little transgression will be observed in the offspring population. This lack of transgression was investigated at the level of estimated genotypic means (BLUEs) from the initial linear mixed model, at the level of predicted values based on multi-QTL models (both single and multi-trait) and, finally, at the level of genomic prediction models (both single and multi-trait).

We defined two statistics Q_{min} and Q_{max} (Alimi et al., 2013a) to estimate the proportion of transgressed offspring lines for each of the three approaches. Q_{min} represents the

proportion of RILs with estimated/predicted values lower than the value for the lower parent while Q_{max} stands for the proportion of RILs with estimated/predicted values higher than the value for the higher parent. When selecting for yield, a breeder will prefer crosses that will give rise to offspring with substantial Q_{max}.

Using univariate transgression obtained from the BLUEs for yield, Q_{min} were 0.03, 0.01, 0.01 and 0.01 while Q_{max} were 0.08, 0.10, 0.01 and 0.31 in the NL1, NL2, SP1 and SP2 trials respectively. For yield, values of both Q_{min} and Q_{max} were generally small across the four environments except in SP2 where Q_{max} was as high as 0.31. Taking SP2 as an example, the values of Q_{max} for yield from the BLUE, QP and GP models univariately were 0.31, 0.26 and 0.15 respectively. As expected, the values of Q_{min} were negligible with the highest being 0.01. From our QTL models, almost all the major QTLs for yield showed increasing alleles from YW (Alimi et al., 2013a; 2013b). However in the MTME model, four QTLs in SP2 trial showed increasing allele effects for yield from CM334 and are believed to be the bases for the transgression observed in this trial (0.26). Hence with regards to yield, the RILs did show some transgression especially in SP2 trial but the transgression was not replicated in other trials and thus may not be enough for them to be appealing to breeders for selection in their own breeding programme. An alternative may be to generate the progeny from a cross between parents having comparable yield but contrasting for yield components, thereby optimizing the power for detecting QTLs underlying the trait variation in yield components (Van Eeuwijk, 2015). Offspring from such parents should show interesting transgression for yield as a consequence of the presence of crossover QTLs. Example of such a cross may be between Poblano and Yolo Wonder (YW). Similar to YW, Poblano is also a large fruited pepper, hence high yielding. Unlike YW, Poblano is mildly hot and also different in its plant phenotypes such as leaf area (Brand et al., 2012). Other varieties that may be used include California Wonder (CW) and Cowhorn which are also both large fruited varieties (Saini and Sharma, 1978).

Furthermore, the issue with our choice of biparental population as regards its suitability for proper breeding and dissection of yield can be circumvented with the use of multi-parental populations and/or association panels. A multi-parent advanced generation intercross (MAGIC) population, (Cavanagh et al., 2008) as an example of a multi-parental population will be able to provide a broad genetic and phenotypic base for dissecting and identifying the genetic control of the complex multigenic trait yield. A MAGIC population can be analyzed by linkage analysis when there is no segregation distortion or structure and by association mapping methodology when segregation distortion occurs due to selection or drift (Cavanagh et al., 2008; Ehrenreich et al., 2009). In MAGIC, a number of parent lines are intercrossed for a number of generations to combine the genomes of all parents in the progeny lines (Huang et al., 2012; 2015; Verbyla et al., 2014). Since multiple parents are utilized, the population segregates for multiple QTL for multiple traits. This allows identification of gene-trait association for complex traits. With a complex target trait such as yield in mind, a MAGIC population will be a preferred resource for creating high-density maps using germplasm relevant to breeders (Cavanagh et al., 2008). MAGIC populations however take longer time to establish in comparison to

association mapping populations and require higher genotyping efforts than biparental populations (Keurentjes et al., 2011; Rakshit et al., 2012).

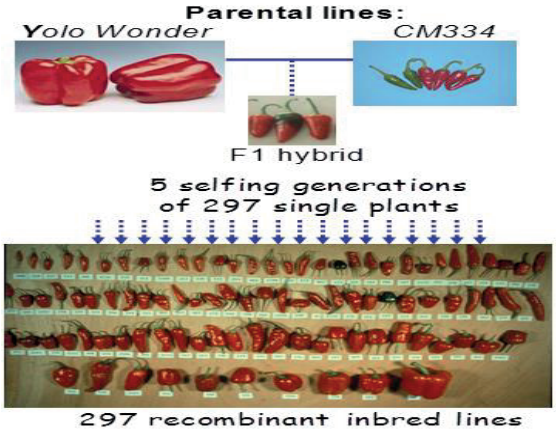


***Figure 6.1**: Plant material used in the EU-SPICY project. Differences in the fruit shape and size between the inbred lines issued from the initial cross between the two Capsicum annuum cv Yolo Wonder and CM334.*

6.3. QTL mapping resolution

The size of mapping populations, *N*, affects the precision of QTL locations as well as the precision with which genetic linkage maps can be constructed (Weller and Soller, 2004; Semagn et al., 2006). The precision of the QTL location is positively correlated with the number of individuals since there is an inverse relationship between confidence interval (CI) for QTL location and population size for any type of biparental mapping population i.e. $CI \propto C/\alpha^2 N$, where α is a standardized QTL allele substitution effect, with standardization in respect to the genetic variance, σ^2 , for a trait. The value of constant C depends on the type of mapping population ranging from 3073, 1537 and 480 for backcross, F2, and RIL_1 (first generation RIL) respectively (Weller and Soller, 2004). To map QTLs with CI of 10cM assuming effect size of 0.25σ (similar to average effect sizes in the SPICY data, see *Tables C1, D1 and E* in Alimi et al., 2013b), population sizes of about 5000, 2500 and 770 individuals are required for backcross, F2 and RIL_1 population respectively. For a 6th generation RIL (RIL_6) population, population sizes of 384, 480 and 768 individuals are required for a CI of 10cM, 8cM and 5cM respectively. In our project, although the whole set of 297 individuals were used for DNA extractions and genetic mapping of molecular markers, only 149 individuals were used for QTL estimation, hence producing a CI of about 13cM precision for a 0.25σ QTL effect size. Theoretically only large effects QTLs with poor precision can be detected with such a population. For a complex target trait such as yield whose genetic architecture is by assumption influenced by many small effect QTLs, it is desirable to have the population size increased to 500 individuals or more, so as to increase the ability to pick up some QTLs of small effects and also enhance the precision of detected QTLs.

However, with advanced QTL modelling techniques such as the multi-trait multi-environment (MTME) model which we developed further here, it is possible to harness the

correlations between environments and among various traits to increase the power of QTL detection for such complex traits, especially for QTLs with small effects. How this is achieved will be shown later. The MTME will leverage on the traits-environments combination to improve the QTL detection. In our case the MTME method significantly improved detection of QTLs having small effects (with explained genetic variance between 3% - 8%) that were not detected by simpler QTL methods. The MTME detected a total of 7, 10, 10 and 13 QTLs for yield as against 1, 2, 3 and 3 QTLs picked up by the single trait analysis in NL1, NL2, SP1 and SP2 environments respectively. These QTLs from MTME also explained higher total genetic variance than those from the single trait analysis. This confirms that although it is advisable to have adequate number of individuals for effective QTL mapping, advanced QTL methodology that takes various sources of correlations into account are useful in QTL mapping when sample size is small, aside from being able to capture small effects QTLs.

Another important factor that can influence the precision of genetic linkage maps and thus QTL resolution is the degree of marker saturation of the genome. For a RIL population with marker saturation levels of 5cM, 10cM and 20cM, the minimum numbers of individuals to be genotyped were found to be 100, 154 and 500 respectively (Silva et al., 2007). The mean between-marker distance in our case was about 4cM, indicating that the number of genotyped individuals is sufficient for QTL detection. However, a number of adjacent marker spacings were more than 5cM (108, 25% cases), with some spacings greater than 10cM (35, 8%) and even as high as 20cM (10, 2%). The three largest distances were 29.7cM, 37.4cM and 45.7cM found on chromosomes 6, 10 and 7 respectively. Therefore, the marker size of 455 used in the EU-SPICY project can be considered adequate for QTL mapping in pepper but the spread of the markers is deficient. Consequently, adding more markers to the chromosome segments with large spacing might improve our ability to map more QTLs for some of the traits.

6.4. Manual and Automated Phenotyping

Yield being a difficult to measure trait can benefit tremendously from the use of novel high-throughput phenotyping (HTP) techniques. This is made feasible since new traits that are correlated with yield and/or yield components can be measured using HTP, hence contributing to better possibilities to map QTLs for yield and its component traits and also enhancing yield prediction. In the EU-SPICY project, phenotyping was done by both manual and automated scorings. The manual scoring recorded physical characteristics related to vegetative and fruit developments. The automated scoring comprised a newly developed fluorescence device (Barócsi, 2012) and a mobile digital imaging tool (Van der Heijden et al., 2012; Song et al., 2014; and Horgan et al., 2014). Both tools permitted high throughput recording of dynamic trait expressions with some of the expressions (such as leaf area and rate of photosynthesis) related to yield component traits.

Yield is regarded as a performance-related trait determined by structural and physiological traits, hence difficult to measure directly. The structural traits are the traits expressing the development and growth of plants (e.g. cell size and leaf area index) while physiological

traits describe plant functioning (e.g. rate of photosynthesis and water content). HTP measurements are rare for structural and physiological traits since they are usually measured at organ and cellular levels unlike morphological traits that are measured at whole-plant level (e.g. stem length and number of fruits) (Li et al., 2015; Dhondt et al., 2013; Fiorani and Schurr, 2013; Furbank and Tester, 2011). It is rather difficult to obtain near accurate measurement for yield (represented as fruit dry weight) through manual measurement unlike its fresh weight counterpart. It is more feasible to automate measurements of fresh weights of fruits, leaves and stems than their dry weights counterpart. However, if accurate expressions relating fresh and dry weights are developed and water contents in the organs are accurately determined, weight-related traits such as yield can be approximated from the automated measurements of their fresh weights. Measurements of water contents in the organs can be automated using 2D imaging tools such as near-infrared cameras, multispectral line scanning cameras and active thermography. Estimates of biomass composition can then be obtained by chemometric methods although extensive calibration will be required (Fiorani and Schurr, 2013; Kümmerlen et al., 1999; Munns et al., 2010; Araus, J. L., & Cairns, 2014; Seelig et al., 2008).

Among yield component traits considered in our crop growth model, partitioning to fruits appears to be the most difficult to obtain via automated devices. However, if biomass accumulation and weights of fruits, stems and leaves are captured over time, then a more reliable and dynamic measurement for this partitioning, spanning the duration of plant development can be obtained. For such dynamic traits, estimates of the slopes of the linear relations between the traits and time may be obtained and its genetic basis in the form of QTLs investigated (See Horgan *et al*., 2015 for an example).

6.5. Complex Traits Analyses

Efficient use of the wealth of phenomics and genomics data for improved QTL mapping and genomic prediction of complex traits such as yield require appropriately designed conceptual and statistical frameworks. The ‘traditional’ QTL approach to link genetic markers to a trait is generally conducted univariately for phenotypes observed in a single environment, but this is often not sufficient for complex traits that exhibit considerable GxE. Also, in plant breeding experiments and with the advent of HTP, phenotypic measurements on a large range of traits are collected simultaneously. These traits are often genetically correlated and the genome-wide availability of genetic markers allows us to study whether these genetic correlations are caused by pleiotropic and/or closely linked QTLs (Mackay, 2001). Similarly important is the understanding of correlation of genotypic performances between multiple environments as these will impact transferability and predictability (Boer et al., 2007). These issues demand improved and novel statistical methods and strategies to adequately describe and analyse such datasets and to arrive at sound QTL results. In this thesis, a number of quantitative genetics models were considered to predict yield either directly from itself or indirectly from its components traits and also study the genetic basis of variation of yield and its component traits in different growing environments and in the presence of other (related) traits.

6.5.1. QTL methods based on linear mixed model

One of the methods we considered extensively was the linear mixed model (LMM) since it offers suitable frameworks for handling complex correlation structures describing several scenarios found in plant breeding such as 1) instances of the same trait in multiple environments; 2) multiple traits in single environments; 3) pairs of traits across environments (Boer et al. 2007; Malosetti et al. 2008; van Eeuwijk et al. 2010). These models also offer flexible tools for handling the diversity of experimental designs, imbalance due to the set of genotypes changing between trials, non-linear relationship, repeated measures during plant cycle, evaluation of genotypes within complex pedigrees, etc.

In this thesis, the LMM QTL approach was implemented in several situations commonly found in plant breeding experiments. We started with the univariate analysis of a number of pepper physiological traits which we termed single trait single environment (STSE) analysis (Alimi et al., 2013a). We first obtained unbiased estimates and genetic parameters for yield and 15 other traits univariately in each of the four environments using a model specification of Piepho et al. (2006). The phenotypic trait value for each individual was estimated taking the design of the experiments in each location into account. For the QTL estimation, we used a multiple-QTL mapping procedure (MQM) (Jansen, 1993; Arends et al., 2010) for each trait in each environment:

$$Y_i = \mu + \sum_{q=1}^{Q} x_{iq}\alpha_q + e_i, \qquad (6.1)$$

where Y_i was the phenotypic response of genotype i, μ the population mean, α_q was the additive effect of QTL q, x_{iq} was a marker-genotype indicator variable (0-1) and e_i was the residual term, which contains both genetic (polygenic, non-detected QTLs) and non-genetic (plot error) contributions. A total of four QTLs were detected for yield with only one of them (C4@35cM) found in all environments. Two main conclusions were drawn from the STSE analysis. A first conclusion concerns the presence of QTL-by-Environment interaction (QEI) as indicated by the differences in the number, level of expression, fraction of variance explained and effect sizes of QTLs for most of the traits across the four trials. The QEI were mostly quantitative and not qualitative, i.e. the QTL showed the same sign in all the environments and mostly differed in magnitude. For example the yield QTL on chromosome 2 showed QEI effects in magnitude, but not in direction (= non-crossovers). This QTL had non-significant effect 0.06 in NL1 and significant effects 0.15, 0.42 and 0.25 in NL2, SP1 and SP2 with superior alleles from parent YW. This is an indication that though many of these traits are genetically determined in any given environment, their degree of expression differs from one environment to the other. Second is the presence of QTL with pleiotropic effects. QTLs with overlapping confidence intervals were loosely declared to be the same QTL, i.e., a QTL with pleiotropic effects. The pleiotropic effects were consistent with the genetic correlations among the traits (Alimi et al., 2013a). None of these conclusions could be properly handled via the univariate analyses. Hence the need to use advanced LMM techniques capable of

combining the data from the four trials while explicitly modelling the complex variance-covariance structures among the environments, traits or trait-environment combinations.

This led us to use multi-environment (ME) model where each trait was evaluated over the four trials. Our main objective was to investigate possible GxE exhibited by yield and yield components, hoping that yield components would show less GxE than yield itself (Alimi et al., 2013b). This would then make it easier to predict yield from its component traits. The final multi-locus ME model was of the form:

$$Y_{ij} = \mu_j + \sum_{q=1}^{Q} x_{iq}\alpha_{jq} + g_{ij} + e_{ij}. \qquad (6.2)$$

Where Y_{ij} denoted the standardized phenotype of the i^{th} genotype ($i = 1,...,149$) in environment j ($j = 1,...,4$), μ_j was the environmental mean, g_{ij} represented the genetic effect of genotype i at environment j, α_{jq} was the environment-specific QTL effect for QTL q and e_{ij} represented the non-genetic component. We assumed that the vectors $\boldsymbol{g_i} = (g_{i1}, \ldots, g_{iJ})$ follow a multivariate normal distribution with zero mean and an unstructured variance-covariance matrix $\boldsymbol{G}$ i.e. $\boldsymbol{g}_i \sim N(0, \boldsymbol{G})$.

The possibility to model heterogeneity of variance and genetic correlation across environments led to more reliable tests for QTL effects and detection of QTLs with differential expressions across environments for many of the traits. For example, three QTLs were detected for yield, similar to the result of STSE. However, the effects of these QTLs were significant in more than one environment, unlike what was obtained from STSE. Detected QTLs were categorized as either constitutive or adaptive, according to the stability of their effects across different environments (Alimi et al., 2013b). The constitutive QTLs are responsible for consistent phenotypic differences between genotypes, with the favourable allele contributing to wide adaptation, at least within the range of environments in which the evaluation was made. This type of QTLs are always desirable to be introduced in elite germplasm of any breeding program as selecting the superior allele will produce a consistent improvement. The adaptive QTLs on the other hand, carry alleles with significantly varying effects across environments and thus constitute the genetic basis of (qualitative/crossover) GxE. All the QTLs for yield were constitutive with majority of the superior alleles from parent YW. This is not surprising as phenotypic differences for yield between the two parents are very pronounced in all of the environments (Alimi et al., 2013a). However across the four environments, more QTLs were picked up for the component traits than yield itself and these QTLs mostly show quantitative GxE. In NL1 for example PF, LUE and LAI had four, three and five QTLs that explained 42%, 26% and 37% variation respectively as against two QTLs for yield which explained 22% variation. Only LUE was affected by QTLs with crossover GxE on chromosome 11 (~70cM). These results show how difficult it is to dissect yield into sensible and useful component traits as the components displayed higher GxE than yield itself.

In the multi-trait (MT) analyses, a number of the resulting phenotypes could be correlated either due to pleiotropy, linkage between close-by QTLs and/or shared environment. Taking these correlations into account was important in our case, not only because of the need to understand pleiotropy, but also because it led to detection of more QTLs as a consequence of increased power due to increase in sample size relative to marginal analyses (Alimi et al., 2013b; Zhou and Stephens, 2014; Furlotte and Eskin, 2015). The specification of MT model is very similar to the ME model. In the case of MT model, instead of having environment (E) in the QTL model (5), we have trait (T). Per environment, there were 15 traits, resulting in four MT analyses. For yield, 4, 8, 9 and 10 QTLs were detected in NL1, NL2, SP1 and SP2 trials respectively including the QTLs already detected in the ME analyses. These QTLs explained higher proportions of genetic variations than STSE and ME analyses. For example, the QTLs from the MT model in SP2 together explained 44.6% genetic variance as against 37% and 28.7% explained by QTLs from STSE and ME models. MT model also provides insights into the genetic architecture of the multiple traits as it enables estimation of their genetic correlation, which is a measure of the portion of the total correlation between traits that is due to (additive) genetic effects (Furlotte and Eskin, 2015). Positive genetic correlations occurred between traits that share a common biological process (e.g. photosynthesis such as leaf area and partition to leaf with average correlation of 0.5 in the four trials) or are components of the same structure (e.g. fruit related traits such as yield and partition to fruit with average correlation of 0.9 in the four trials), and negative genetic correlations were found between components of fitness or traits in competition for resource allocation (e.g. stem and fruit such as partition to stem and yield with average correlation of -0.6 in the four trials) (Alimi et al., 2013b). Thus, understanding the underlying pleiotropic connections between quantitative traits is important for predicting correlated responses to artificial selection and understanding genetic constraints on the evolution of natural populations (Mackay et al., 2009).

Specifically for application in the EU-SPICY project, we adapted a multi-trait multi-environment (MTME) QTL model as the most general form of the LMM for identifying QTL in the presence of several sources of correlations (Alimi et al., 2013b; Malosetti et al., 2008). This model helped to identify the genome regions responsible for genetic correlations between trait-by-environment combinations and showed how genetic correlations depend on the environmental conditions. Extension to the MTME setting was achieved by combining traits across the four environments in a single LMM analysis where we specified the response trait (*Y*) to be a vector of the traits (*T*) and environments (*E*) combinations. The mean for the trait by environment combination, TE, is taken as fixed in the QTL analysis:

$$Y_{ip} = \mu_p + \sum_{q=1}^{Q} x_{iq}\alpha_{pq} + g_{ip} + e_{ip}, \qquad (6.3)$$

where μ_p (p *= 1, 2,..., 60*) is the trait by environment mean (p runs over the product of environment and traits), α_{pq} is the trait by environment-specific QTL effect for QTL q, g_{ip} represents the genetic effect of genotype i for trait by environment combination p, and e_{ip}

is the residual effect. We specified an unstructured VCOV matrix for all pairs of the trait by environment combinations, giving a total of 1830 parameters. With the MTME model, GxE and genetic correlations between traits were simultaneously modelled. The MTME method yielded many QTLs of small effects (between 3% - 8%) that were not detected in both ME and MT methods. MT and ME however had more QTLs that explained above 10% genetic variations than MTME. This might be related to the so called "Beavis effect" (Beavis, 1994, 1997; Xu, 2003). Beavis effect is used to qualify situations where simpler models fail to detect some QTLs with small effects and also result in overestimation of some effect sizes. On the one hand, the Beavis effect will cause the estimated number of QTL to be biased down ward, because the undetected QTLs are not reported. On the other hand, the average effect of the detected QTLs will be biased upward.

The average power to detect QTLs by each of the models was compared using a standard *t*-test power function for the univariate model and Hotelling's T^2 power function for the multivariate models. For a *p*-variates situation, the Hotelling's T^2 can be viewed as a combination of univariate *t*-tests. Although there may not exist any unique best test for power in multivariate settings, the Hotelling's T^2 is probably the best known test for this problem since it is the likelihood ratio test and is uniformly most powerful (UMP) among all tests that are invariant under the group of non-singular linear transformations (Wu et al., 2006; Agresti & Klingenberg, 2005; Kaplan and George, 1995; Kariya, 1981). While the obvious test for power in LMM is the Wald test statistics, nevertheless Hotelling's T^2 is almost equal to the Wald test statistics and they are asymptotically equivalent when the sample size is large ($n \rightarrow \infty$). The sample size of 149 individuals studied here is considered large enough for the use of Hotelling's T^2 in lieu of Wald test statistics. Also, the exact F-distribution of the Hotelling's T^2 converges to the Wald test χ^2-distribution when *n* is large (Wu et al., 2006). The *p*-variate Hotelling's T^2 is written as:

$$T_p^2 = N(\overline{\boldsymbol{Y}} - \boldsymbol{\mu_0})' \boldsymbol{S}^{-1} (\overline{\boldsymbol{Y}} - \boldsymbol{\mu_0}), \qquad (6.4)$$

where $\overline{\boldsymbol{Y}}$ is the vector of sample means for the variates to be tested (e.g. the effects for yield in each of the four trials as estimated from MTME) and $\boldsymbol{S}$ is their covariance/correlation matrix. We assume that all *N* observations available on *p* variables have the same multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. The aim is to test the hypotheses $H_0: \boldsymbol{\mu} = \boldsymbol{\mu_0}$ *versus* $H_a: \boldsymbol{\mu} = \boldsymbol{\mu_a}$ where at least one component of $\boldsymbol{\mu_a}$ is different from the corresponding component of $\boldsymbol{\mu_0}$. Usually, $\boldsymbol{\mu_0}$ is a vector of zeros. The power function of the critical region α for the rejection of the null hypothesis above can be represented as $\beta_\alpha(\mu_1, \ldots, \mu_p, \boldsymbol{\Sigma})$, with power equal to $1 - \beta$. Using non-centrality parameters, the power of the Hotelling's T^2 may be calculated for any value of the means and standard deviations. Since there is a simple relationship between the non-central T^2 and the non-central *F*, calculations are actually based on the non-central *F* using the formula (Muller et al., 1992):

$$\beta = \Pr(F < F'_{\alpha, df1, df2}), \qquad \text{where } df1 = p \text{ and } df2 = N - p.$$

For many of the constitutive QTLs detected by the various models, the power was better with multivariate models than the univariate model. For example, the standardized estimated effects of the yield QTL on chromosome 4 at 35cM by the MTME model were 0.31, 0.30, 0.26 and 0.24 for NL1, NL2, SP1 and SP2 trials respectively. Using these effects together with the correlations between yield trait across the four trials (see *Table A2* in Alimi et al., 2013b) and a 0.05 test level, the estimated multivariate power from Hotelling's T^2 was 0.92. Following the same pattern, the power from the ME and MT models were 0.98 and 0.95 respectively. The estimated effects from the STSE for the same QTL were 0.25, 0.20, 0.19 and 0.24 for NL1, NL2, SP1 and SP2 trials respectively giving estimated power of 0.88, 0.70, 0.65 and 0.85 for NL1, NL2, SP1 and SP2 trials respectively using univariate *t*-test statistic. This particular QTL was picked up in the four environments by all the models with very big effect sizes (≥0.24) and hence very high power. For an adaptive QTL such as the QTL on chromosome 7 at 78cM with effect sizes 0.10, 0.08, 0.05 and 0.28 in NL1, NL2, SP1 and SP2 trials respectively, the multivariate power by MTME was 0.89. Using STSE, this QTL was significant in only SP2 with effect size of 0.21 and power 0.74. An example of QTL with small effects for yield detected by MTME and only significant in SP1 is the QTL on chromosome 2 at 2cM with effect sizes 0.13, 0.04, 0.26 and 0.06 in NL1, NL2, SP1 and SP2 trials respectively yielding multivariate power of 0.83. In summary, there is a very high power of detecting constitutive QTLs across the four environments irrespective of the models while multivariate models especially the MTME increase the power of detecting adaptive QTLs and QTLs with small effects. Power calculations were done using G*Power software (Faul et al., 2009).

Yield prediction also improved significantly under the MTME since its genetic correlations with other traits were better exploited. Prediction accuracies for yield in SP2 for example improved from about 0.53 under the ME model to 0.64 under the MT and 0.81 under the MTME (Alimi et al., 2013b). The MTME was especially suitable for the complex trait yield as it led to better detection of pleiotropic QTLs with either synergistic or antagonistic effects, some complementary QTLs (qualitative GxE), differential allele expression according to environments (quantitative GxE) and an increased explained variance for the complex target trait. For instance, the QTLs from MTME in SP2 explained 56% genetic variance as against 37%, 28.7% and 44.6% explained by QTLs from STSE, ME and MT models respectively. The QTL on chromosome 3 around 140cM as detected by both MT and MTME models, is an example of an antagonistic pleiotropic QTL as it had opposite effects on different traits: YW had the increasing allele for yield related traits while CM334 had the increasing allele for stem related traits. As expected due to the population type, the majority of the yield increasing alleles are from YW. However, one particular QTL on chromosome 12 showed an increasing allele effect for yield from CM334 in NL2 and SP2 trials and can be regarded as a complementary QTL for yield and a basis for transgression.

Conversely, there are some constraints limiting the application of multivariate LMM for QTL mapping. First is the computational difficulty in fitting multivariate models for large

number of traits simultaneously. In such situations, it is desirable to implement a multivariate approach on the most relevant traits instead of performing the analysis on all available traits. A variable selection approach has been proposed to choose a subset of informative traits for multitrait QTL mapping while still maintaining optimal statistical power for QTL identification (Cheng et al., 2013). This has the obvious advantage of better biological interpretation over standard data reduction techniques such as principal components without requiring any back transformation since selections of the original traits are used (Aschard et al., 2014; Gao et al., 2014). Second is the realization that it is not in all situations that multivariate analysis is more powerful than univariate analysis as the statistical power of multitrait analysis depends on both the QTL effects and the structure of the residual covariance of the traits (Zhou and Stephens, 2014; Korol et al., 1995 and Jiang and Zeng, 1995). This was experienced with some traits in the EU-SPICY data such as LAI in NL1 and NL2 trials where some QTLs were significant only in the univariate analysis. Thus, multivariate and univariate tests should be viewed as complementary rather than competing (Zhou and Stephens, 2014).

Another constraint has to do with the use of a limited set of QTL-related markers instead of estimation of effects for all markers. This limits the usefulness of QTL models for prediction purposes. Further constraint is the lack of proper understanding of the pleiotropic paths (either having direct/indirect effects) revealed by multivariate model. A number of novel statistical methodologies have been proposed to address some of these limitations, allowing multivariate analyses to be more useful. The performances of three of such methodologies namely genomic prediction models, crop growth models and causal network models, for modelling complex traits, were also examined for the EU-SPICY data.

6.5.2. Genomic Prediction and Integrated Crop Growth Models

Genomic prediction (GP) models offer interesting alternatives to QTL based prediction (QP) models especially for complex trait predictions. The main difference between the two classes of models is in the use of all markers in a penalized regression context for GP, where all QTLs are assumed to be in linkage disequilibrium with one or more molecular markers instead of the use of a limited set of QTL-related markers in QP. Several studies have shown that multi-trait versions of GP perform better than single-trait versions (Burgueño et al., 2012; Jia and Jannink, 2012; Sørensen et al., 2012; Calus and Veerkamp, 2011). Many studies on QTL and association mappings have also shown that the joint analysis of multiple traits helps to improve the power and precision of QTL (Alimi et al., 2013b; Jiang and Zeng, 1995) and association mappings (Galesloot et al., 2014; Stephens, 2013; Xu et al., 2015). We extended this frontier by comparing multi-trait versions of GP and QP models. The predictive performances of both single-trait (ST) and multi-trait (MT) versions of GP were investigated and compared with performances of ST and MT versions of QP using yield and its component traits.

The four methods (STQP, STGP, MTQP, MTGP) differed substantially in their predictive potentials for yield. GP models showed higher predictive accuracies than QP models with

MT outperforming ST situations. Yield in SP2 for example had prediction accuracies of 0.42, 0.53, 0.51 and 0.89 using ST-QP, MT-QP, ST-GP and MT-GP models respectively. These results confirm that GP methods are better predictive models than QTL methods. This is not really surprising since parameters from GP methods are fitted on all available markers while parameters of QP methods are fitted only on selected QTL markers. Similarly, the linkage disequilibrium (LD) between markers and QTL is exploited better by GP than by QP, leading to higher prediction accuracies (Habier et al., 2010; Hayes et al., 2009). Also, MT models exploited the genetic correlations among the traits leading to improved predictive accuracies. Unlike QP methods, the GP methods fully take advantage of the correlations between all the markers and control shrinkage estimation by assigning a prior distribution to marker effects. Prediction accuracies differed across environments for each of the five traits irrespective of the prediction method employed. Traits were generally better predicted in SP trials than in NL trials. This is probably due to the type of population used as this population is more suited for outdoor growing system than for the system in a greenhouse. The climate conditions (light, temperature) in Spain were more suitable for this population than the NL conditions. Yield for example, had prediction accuracies ranging from 0.16 – 0.84 over methods in NL1, 0.11 – 0.75 in NL2, 0.47 – 0.90 in SP1 and 0.42 – 0.89 in SP2 while radiation use efficiency (RUE) had accuracies ranging from 0.44 – 0.65, 0.11 – 0.71, 0.27 – 0.89 and 0.28 – 0.85 in NL1, NL2, SP1 and SP2 respectively. These differences in prediction accuracies further confirm that GxE is an important component of the genetic variability for these traits (Alimi et al., 2013b).

The extents to which both QP and GP models predict yield from the component traits for new genotypes and/or environments were explored by linking the prediction methods with crop growth model (CGM). A CGM can suggest writing a complex target trait as a function of a set of simpler component traits and a set of environmental input covariables (Bustos et al., 2015; Chenu et al., 2008; Hammer et al. 2010). CGMs with known/predicted genotypic parameters are a potentially useful tool to understand which traits can be advantageous in a given environment, and also to identify management practices that contribute to improved crop productivity (Yin et al. 2004; Hammer et al. 2006). Here we integrated QTL/genomic prediction and CGM approaches and showed that the target trait yield can be predicted via its component traits namely radiation use efficiency (RUE), partitioning into the fruits (PF) and growth rate of leaf area index (LAI_{rate}) together with environmental covariables such as temperature, thermal time and daily global radiation intensity (I). For genotype i in environment j, the CGM was mathematically written as:

$$Yield_{ij} = PF_{ij} * RUE_{ij} * \sum_1^D \left(I_{jd} * \left(1 - e^{-k*LAI_{ijd}}\right)\right), \qquad (6.5)$$

with yield accumulated over the growing days $d = 1 \ldots D$. The leaf area index (LAI) is dynamic and for genotype i in environment j on a specific day d ($d \leq D$) calculated as $LAI_{ijd} = LAI_{rate_{ij}} * \sum_1^d \left(T_{jd} - T_b\right)$. The term $\sum_1^d \left(T_{jd} - T_b\right)$ is the accumulated thermal time till day d, expressed in degree-days, and LAI_{rate} is a genotype specific increase rate of leaf area index. T_{jd} is the daily average temperature in environment j on day d, and T_b is

the base temperature below which no development takes place, taken as 10°C in all environments (van Ittersum et al., 2003; Marcelis et al., 2006). The increase rate of leaf area index (LAI_{rate}) was calculated as the ratio between the increase of leaf area index (LAI) and thermal time between initial and final harvest. The fraction intercepted radiation is $1 - e^{-k*LAI_{ijd}}$, where a value of 0.7 for the extinction coefficient k is assumed for all genotypes (Marcelis et al., 1998). Daily intercepted radiation is then calculated from this fraction and I_{jd}, the daily global radiation intensity (MJ m^{-2} d^{-1}). The daily intercepted radiation is consequently multiplied by the radiation use efficiency (RUE_{ij}) resulting in daily dry matter production. RUE represents the biomass produced per unit of intercepted radiation. It is the ratio between biomass increase and the total intercepted radiation, which was the daily intercepted global radiation summed over the total growth period. Finally, yield is calculated from the total accumulated dry matter by multiplying it by the fraction biomass partitioned into the fruits (PF_{ij}), i.e. fruit biomass/total plant biomass. Total plant biomass was calculated as the sum of plant dry weight at final destructive harvest and the dry weight of the already harvested fruits.

The prediction accuracy of the target trait depends on the accuracy of the prediction of each of the components, and on the ability of the ecophysiological functions to correctly describe the processes leading to the target trait. Since CGM produces GxE as an emerging property of the interaction between the physiological parameters and the environmental information, the CGM we adopted has the added advantage of being able to describe GxE as it contained explicit representations of development over time and integrated developmental and environmental information (Tardieu, 2003; Chapman et al., 2008; Chenu et al., 2009; Cooper et al., 2009). Therefore, using component traits and environmental covariables from one environment to predict yield in another environment was a possibility. The across-environment analysis envisages a situation where we wish to predict how a certain population will perform in a new environment.

We noted that the integrated CGM approach performed creditably well in predicting the complex target trait yield. The accuracies obtained using integrated CGM are close to the direct prediction strategy of GP and QP models. The accuracy of prediction using integrated CGM approach also achieved the highest values with MT-GP method. The MT-GP in the integrated CGM had prediction accuracies of 0.83, 0.81, 0.92 and 0.86 in NL1, NL2, SP1 and SP2 respectively, similar to 0.84, 0.75, 0.90 and 0.89 from the direct prediction approach. For the across-environment integrated CGM analyses using breeding values from ST-GP, prediction accuracy for yield in NL1 environment improved from 0.18 to 0.23, 0.49 and 0.35, if component traits from NL2, SP1 and SP2 were used respectively. MT-GP on the other hand gave accuracies of 0.63, 0.60 and 0.57 when yield in NL1 was predicted using component traits from NL2, SP1 and SP2 respectively. Overall, the prediction accuracies for yield in NL trials increased if component traits from SP trials were employed while prediction accuracies for yield in SP environments did not improve when component traits from NL environments were used. This showed that difficult to measure complex traits such as yield can be successfully predicted from its component traits. The success of such predictions depends on (i) a well-defined CGM

relating the complex and component traits, (ii) a well-defined set of environmental covariables with corresponding CGM parameters, and (iii) a well understood genetic basis of the CGM parameters (Bustos et al., 2015; Slafer and Rawson, 1994; Snape et al., 2001).

However apart from the prediction model of choice, a suitable population and comparable management and environmental settings should be used across the environments for proper GxE analyses with the CGM. To make a strong case for the integrated CGM approach, the target complex traits should be defined as functions of as much as possible independent component traits having well understood genetic basis, where these components traits themselves can still be measured with a certain ease or where these component traits may be approximated by other traits that can be recorded quickly and cheaply by automated phenotyping devices. The environmental covariables should be well defined. Also, the CGM should attain balance between being too empirical and too mechanistic. Model parameters may have little biological meaning when the CGM is too empirical while it is difficult to model all plant processes with a consistent mechanistic detail because our level of understanding of the biological processes become limited with a lowering of the level of analysis (Yin and Struik, 2010). It is therefore unlikely that one could arrive at any reasonable and purely mechanistic model satisfactory to understand, explain, learn and predict (biological) outcomes. Furthermore, despite huge efforts aimed at extending the use of CGM to enable genetic prediction, the focus has been mainly restricted to linking CGM and QP, while the use of CGM in combination with GP models has been largely ignored, with exception of Technow et al. (2015) and Onogi et al. (2016). Explicit incorporation of biological knowledge through CGM into GP models for complex traits have the potential to open up novel avenues towards accounting for epistatic effects since the relationship between the underlying component traits may be non-additive (Holland, 2001) .

6.5.3. Causal network model

Network type models provide alternative representation of the biological knowledge of a complex target trait such as yield, by showing intricate interactions of multiple genetic (and possibly environmental factors) influencing the target trait. The use of network models show how plant traits are interconnected in networks of dependencies as a result of gene-to-gene interactions and also the stability of such networks across environments due to GxE (Granier and Vile, 2014; Li et al., 2010; Valente et al., 2013). A number of network algorithms have been developed to build and understand gene-to-gene interaction architecture underlying relationships among traits (Tasaki et al., 2015; Schadt et al. 2005; Neto et al. 2008, 2013; Hageman et al., 2011). We defined both conditional and unconditional correlation networks to study putative (causal) relationships among yield and its three component traits across the four environments. The unconditional networks were based on standard multi trait QTL model (MTM) (Jiang and Zeng, 1995; Alimi et al., 2013b) while the conditional networks were based on the QTL-driven phenotype network method (QTLnet) developed by Neto et al. (2010). QTLnet jointly models genetic architecture and phenotype network structure using homogeneous conditional Gaussian regression (HCGR) models (Lauritzen, 1996). This method is termed conditional network

as the genetic architecture for each phenotype is inferred conditional on the phenotype network. The correlation structure among phenotypes is explicitly modelled according to the directed graph representation of the phenotype network. The genetic model is derived from a system of linear regression equations which corresponds to the HCGR (Neto et al., 2010). In the HCGR model, the phenotypes (***y***) are distributed according to a multivariate normal distribution conditional on the QTL genotypes (***q***) which are subsets of the marker genotypes (***m***), while the QTL ***q*** are modelled through the mean. The joint probability of ***y*** and ***q*** can thus be partitioned into genetic and recombination components, respectively relating phenotypes to QTL and QTL to observed markers across the genome. For genotype *i* and trait *p*, the phenotype model was represented as:

$$Y_{ip} = \mu_p + \sum_{q=1}^{Q} x_{iq}\alpha_{pq} + \sum_{y_v \in pa(y_p)} \theta_{pv} y_{vi} + e_{ip}, \qquad e_{ip} \sim N\left(0, \sigma_p^2\right), \qquad (6.6)$$

where μ_p was the overall mean for trait *p*, α_{pq} was the genetic effects for trait *p*, and x_{iq} represented the genetic effect predictors derived from the conditional QTL genotype probabilities. The notation pa(y_p) represented the set of parent phenotype nodes that directly affect y_p, that is, θ_{pv} are the partial regression coefficients relating phenotypes having covariance structure that depends exclusively in the relationships among phenotypes and e_{ip} represented the normally-distributed residual component.

With MTM we could establish QTLs with pleiotropic effects for correlated traits (see also Alimi et al., 2013b). The QTLnet can disentangle the paths for pleiotropic QTL by conditioning on QTL genotypes. For instance, in SP2, MTM reported six QTL with five of them having pleiotropic and direct effects while QTLnet revealed only one QTL with pleiotropic effect and six QTL with direct and indirect effects. As an example, MTM revealed that Q7@78 has a pleiotropic effect on yield, LAI and PF (an example of hotspot) but QTLnet showed that Q7@78 has direct effect only on PF through which it affects LAI and yield indirectly. No QTL directly influences yield in SP2. All the effects on yield were found to be indirect through its upstream component traits. Hence, the QTL hotspots from the MTM resulted from ignoring network structure for the correlated traits. QTLnet improved detection of refined genetic architecture by distinguishing between QTL with direct and indirect effects and resolving QTL hotspots.

The final conditional networks across the four environments are similar in skeleton to the CGM representation. Similar to the CGM topology, yield was established to be downstream to its three component traits, indicating that yield can be predicted from the component traits. The extent of the accuracy of such prediction was roughly estimated using the R^2 values for yield obtained from fitting SEM to the final configuration from the two models. The R^2 values for yield in each environment were higher for QTLnet than for the MTM. As an example, the SEM for yield in SP2 had R^2 of 0.31 and 0.93 for MTM and QTLnet models respectively, indicating that properly resolving direct and indirect paths leading to yield would improve its prediction. Additionally, the network model captured putative relationships among the component traits, which were largely ignored in the CGM representation. However, the network model did not incorporate any environmental

information. Hence, including environmental characterizations via the CGM into the network model may increase our ability to improve on yield prediction through its component traits especially for new environments. Furthermore, since the use of network models is of particular importance in quantifying changes due to intervention, effects of environmental characterisation on the target complex trait can thus be quantified as an example of such intervention. This is possible since the relationships depicted by conditional networks are direct, thus any change due to such interventions can be easily quantified (Bouwman et al., 2014; Valente et al., 2013). We can predict changes in the target complex trait across the different environments by adding environmental information to the network relating the component and target traits. This is however only feasible if the component traits have simple genetic architecture without GEI. Such predictions are made by representing the intervention on the causal structure among phenotypes and by knowing the genetic effect directly on each trait, as well as the dispersion parameters that describe their joint distribution. All these are possible by fitting a SEM (Bouwman et al., 2014).

The suites of (statistical) methodologies presented in this thesis are not exhaustive for predicting the phenotypic response of genotypes for complex traits under a range of environmental conditions. In recent years, several other statistical (mainly non-parametric) models have been introduced for genome-enabled prediction. These models include, among others, kernel regressions (Bennewitz et al. 2009), random forest (González-Recio and Forni, 2011), support vector regression (Moser et al. 2009), reproducing kernel Hilbert spaces (RKHS) mixed model (Gianola et al. 2006, 2008) and artificial neural networks (ANN) (Okut et al. 2011; Gianola et al. 2011). They have been widely used for pattern recognition, classification and prediction problems in other fields of application such as image processing and reconstruction (Takeda et al. 2007; Hastie et al. 2009). Applications of these techniques in plant breeding are in their infancy as majority of plant breeders still rely on the use of more common and readily available parametric methods. The attractiveness of these non-parametric methods lies in the fact that they are able to handle the multiplicity of potential interactions (collinearity) arising as a result of e.g. hundreds of thousands of markers, and that most of the assumptions of parametric methods (e.g., linearity, multivariate normality, proportion of segregating loci, spatial within-chromosome effects) required for an orthogonal decomposition of variance are violated in artificial and natural populations (Gianola et al., 2006).

6.6. Concluding Remarks

The results of the analyses presented in this thesis have contributed to a better understanding of the genetics of yield-related physiological traits in pepper and represent an important step in the improvement of the target trait yield. Yield profited from joint analysis with other traits through exploitation of its genetic correlation with the other traits. Such joint analysis led to better detection of QTLs with small effects which usually remain undetected in a univariate analysis, hence higher power and better prediction accuracy for the complex trait. It was shown that the accuracy of predicting yield improved tremendously (from < 0.4 to > 0.8) when effects of all markers are estimated

simultaneously via multivariate whole genome prediction methods instead of using a limited set of QTL-related markers. Incorporating ecophysiological representation of yield where yield was written as a function of a set of simpler component traits and a set of environmental input variables, into the purely statistical whole genome prediction model showed that difficult to measure complex traits can be successfully predicted from their component traits. This integrated approach was able to describe and model GxE since it contained explicit representations of development over time and integrated developmental and environmental information, hence increases prospect of inter-environment prediction. The approach also has added advantage of opening up possibilities for modelling epistatic effects since the relationship between the underlying component traits may be non-additive. Using conditional network-type correlation models to represent the biological relationship of multiple traits, genetic and environmental factors influencing target trait yield, we were able to confirm relationships between yield and its component traits as depicted by the ecophysiological model. The conditional network model refined the genetic architecture of yield and its component traits by distinguishing between QTLs with direct and indirect effects. Incorporating this refined genetic architecture into complex trait dissection as proposed by CGM can be utilized to construct multi-trait whole genome prediction models for complex traits. This would lead to improvement in the prediction of the target complex trait and thus genetic gain in genome assisted selection for the complex trait.

References

Agresti, A., & Klingenberg, B. (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(4), 691-706.

Alimi, N. A., Bink, M. C. A. M., Dieleman, J. A., Magán, J. J., Wubs, A. M., Palloix, A., and Eeuwijk, F. A. (2013b). 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. A., Bink, M. C. A. M., Dieleman, J. A., Nicolaï, M., Wubs, M., Heuvelink, E., Magan, J., Voorrips, R. E., Jansen, J., Rodrigues, P. C., Heijden, G. W. A. M., Vercauteren, A., Vuylsteke, M., Song, Y., Glasbey, C., Barocsi, A., Lefebvre, V., Palloix, A., and Eeuwijk, F. A. (2013a). Genetic and QTL analyses of yield and a set of physiological traits in pepper. *Euphytica* **190**, 181-201.

Alimi, N. A., Bink, M. C. A. M., Janss, L. L. G., Wubs, A. M., Dieleman, J. A., Magán, J. J., Heuvelink, E., Palloix, A., and Eeuwijk, F. A. (2016). Predicting complex traits in multiple environments by a combination of genomic prediction and crop growth modelling: an example in pepper. **To be submitted**.

Alimi, N.A., Bink, M.C.A.M., Dieleman, J.A., Sage-Palloix, A.M., Voorrips, R.E., Lefebvre, V., Palloix, A., Eeuwijk, F.A.v., 2010. Exploratory QTL analyses of some pepper physiological traits in two environments. Advances in Genetics and Breeding of Capsicum and Eggplant: Proceedings of the XIVth EUCARPIA Meeting on genetics and breeding of Capsicum and Eggplant. Editorial Universidad Politécnica de Valencia, Valencia, Spain, Valencia, Spain, pp. 295-300.

Andersen, J. R., and Lübberstedt, T. (2003). Functional markers in plants. *Trends in Plant Science* **8**, 554-560.

Anhalt, U. C. M., Heslop-Harrison, J. S., Piepho, H. P., Byrne, S., and Barth, S. (2009). Quantitative trait loci mapping for biomass yield traits in a Lolium inbred line derived F2 population. 170, 99-107.

Ansell, P., Furbank, R., Gunasekera, K., Guo, J., Benn, D., Williams, G., & Sirault, X. (2013). Flexible scientific data management for plant phenomics research. *Semantics for Biodiversity (S4BioDiv 2013)*, 63.

Araus, J. L., & Cairns, J. E. (2014). Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science*, **19**(1), 52-61.

Arends, D., Prins, P., Jansen, R. C., Broman, K. W. (2010). R/qtl: high-throughput multiple QTL mapping. *Bioinformatics* **26**, 2990-2992.

Barchi, L., Bonnet, J., Boudet, C., Signoret, P., Nagy, I., Lanteri, S., Palloix, A., and Lefebvre, V. (2007). A high-resolution, intraspecific linkage map of pepper (Capsicum annuum L.) and selection of reduced recombinant inbred line subsets for fast mapping. *Genome* **50**, 51-60.

Barchi, L., Lefebvre, V., Sage-Palloix, A.-M., Lanteri, S., and Palloix, A. (2009). QTL analysis of plant development and fruit traits in pepper and performance of selective phenotyping. TAG Theoretical and Applied Genetics 118, 1157-1171.

Barócsi, A. (2012). Intelligent, net or wireless enabled fluorosensors for high throughput monitoring of assorted crops. *Measurement Science and Technology* **24**, 025701.

Bauer, A. M., Hoti, F., von Korff, M., Pillen, K., Léon, J., and Sillanpää, M. (2009). Advanced backcross-QTL analysis in spring barley (H. vulgare ssp. spontaneum) comparing a REML versus a Bayesian model in multi-environmental field trials. *Theoretical and Applied genetics* **119**, 105-123.

Beaumont, M. A., and Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics* **5**, 251-261.

Beavis, W. D. (1994). The power and deceit of QTL experiments: Lessons from comparative QTL studies. . *In Proceedings of the Forty-ninth Annual Corn and Sorghum Research Conference (Washington, DC, American Seed Trade Association)*, 250–266.

Beavis, W. D. (1997). QTL analyses: Power, precision, and accuracy *In Molecular Dissection of Complex Traits* **ed. A. H. Paterson (Boca Raton, FL, CRC Press, 1997)**, 145–162.

Beckmann, J., and Soller, M. (1986). Restriction fragment length polymorphisms and genetic improvement of agricultural species. *Euphytica* **35**, 111-124.

Ben Chaim, A., Borovsky, Y., Falise, M., Mazourek, M., Kang, B.C., Paran, I., Jahn, M., (2006a). QTL analysis for capsaicinoid content in Capsicum. *Theoretical and Applied Genetics* **113**, 1481-1490.

Ben Chaim, A., Borovsky, Y., Rao, G., Gur, A., Zamir, D., Paran, I., (2006b). Comparative QTL mapping of fruit size and shape in tomato and pepper. *Israel Journal of Plant Sciences* **54**, 191-203.

Ben Chaim, A., Grube, R.C., Lapidot, M., Jahn, M., Paran, I., (2001a). Identification of quantitative trait loci associated with resistance to cucumber mosaic virus in Capsicum annuum. *Theoretical and Applied Genetics* **102**, 1213-1220.

Ben Chaim, A., Paran, I., Grube, R.C., Jahn, M., van Wijk, R., Peleman, J., (2001b). QTL mapping of fruit-related traits in pepper (Capsicum annuum). *Theoretical and Applied Genetics* **102**, 1016-1028.

Bennewitz J, Solberg T, Meuwissen THE (2009). Genomic breeding value estimation using nonparametric additive regression models. *Genet Select Evol*;41:20.

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

Bishop CM (2006). Pattern Recognition and Machine Learning. Springer, Singapore.

Blum, E., Mazourek, M., O'Connell, M., Curry, J., Thorup, T., Liu, K.D., Jahn, M., Paran, I., (2003). Molecular mapping of capsaicinoid biosynthesis genes and quantitative trait loci analysis for capsaicinoid content in Capsicum. *Theoretical and Applied Genetics* 108, 79-86.

Boer, M. P., Wright, D., Feng, L. Z., Podlich, D. W., Luo, L., Cooper, M., and 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.

Boote, K., Kropff, M., and Bindraban, P. (2001). Physiology and modelling of traits in crop plants: implications for genetic improvement. *Agricultural Systems* **70**, 395-420.

Borevitz, J. O., and Nordborg, M. (2003). The impact of genomics on the study of natural variation in Arabidopsis. *Plant physiology* **132**, 718-725.

Bouwman, A. C., Valente, B. D., Janss, L. L., Bovenhuis, H., and Rosa, G. J. (2014). Exploring causal networks of bovine milk fatty acids in a multivariate mixed model context. *Genetics Selection Evolution* **46**, 2.

Brand, A., Borovsky, Y., Meir, S., Rogachev, I., Aharoni, A., & Paran, I. (2012). pc8. 1, a major QTL for pigment content in pepper fruit, is associated with variation in plastid compartment size. *Planta*, 235(3), 579-588.

Broman, K. W., and Sen, S. (2009). A guide to QTL mapping with R/qtl. Springer, New York; London.

Brown, T. B., Cheng, R., Sirault, X. R., Rungrat, T., Murray, K. D., Trtilek, M., ... & Borevitz, J. O. (2014). TraitCapture: genomic and environment modelling of plant phenomic data. *Current opinion in plant biology*, *18*, 73-79.

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. *Crop Science* **52**, 707-719.

Bustos-Korts, D., Malosetti, M., Chapman, S., & van Eeuwijk, F. (2016). Modelling of Genotype by Environment Interaction and Prediction of Complex Traits across Multiple Environments as a Synthesis of Crop Growth Modelling, Genetics and Statistics. *In Crop Systems Biology* (pp. 55-82). Springer International Publishing.

Calus, M. P., and Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution* **43**, 1-14.

Caranta, C., Lefebvre, V., Palloix, A., (1997a). Polygenic resistance of pepper to potyviruses consists of a combination of isolate-specific and broad-spectrum quantitative trait loci. *Molecular Plant-Microbe Interactions* **10**, 872-878.

Caranta, C., Palloix, A., Lefebvre, V., Daubeze, A.M., (1997b). QTL for a component of partial resistance to cucumber mosaic virus in pepper: Restriction of virus installation in host-cells. *Theoretical and Applied Genetics* **94**, 431-438.

Causse, M., Saliba-Colombani, V., Lecomte, L., Duffe, P., Rousselle, P., & Buret, M. (2002). QTL analysis of fruit quality in fresh market tomato: a few chromosome regions control the variation of sensory and instrumental traits. *Journal of experimental botany*, **53**(377), 2089-2098.

Cavanagh, C., Morell, M., Mackay, I., and 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.

Chaim, A.B., Borovsky, Y., De Jong, W., Paran, I., 2003. Linkage of the A locus for the presence of anthocyanin and fs10.1, a major fruit-shape QTL in pepper. *Theoretical and Applied Genetics* **106**, 889-894.

Chapman, S. (2008). Use of crop models to understand genotype by environment interactions for drought in real-world and simulated plant breeding trials. *Euphytica* **161**, 195-208.

Cheng R, Borevitz J, Doerge RW (2013). Selecting informative traits for multivariate quantitative trait locus mapping helps to gain optimal power. *Genetics* **195**,683-691.

Chenu, K., Chapman, S. C., Tardieu, F., McLean, G., Welcker, C., and Hammer, G. L. (2009). Simulating the yield impacts of organ-level quantitative trait loci associated with drought response in maize: a "gene-to-phenotype" modeling approach. *Genetics* **183**, 1507-1523.

Collins, N. C., Tardieu, F., and Tuberosa, R. (2008). Quantitative trait loci and crop performance under abiotic stress: where do we stand? *Plant Physiology* **147**, 469-486.

Cooper, M., van Eeuwijk, F. A., Hammer, G. L., Podlich, D. W., and Messina, C. (2009). Modeling QTL for complex traits: detection and context for plant breeding. *Current Opinion in Plant Biology* **12**, 231.

Crossa, J., de Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., and Yan, J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* **186**, 713-724.

Cullis, B.R., Smith, A.B., Coombes, N.E., (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural Biological and Environmental Statistics* **11**, 381-393.

Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de Los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* **193**, 347-65.

Darvasi, A., and Soller, M. (1994). Optimum spacing of genetic markers for determining linkage between marker loci and quantitative trait loci. *Theoretical and Applied Genetics* **89**, 351-357.

Darvasi, A., and Soller, M. (1995). Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* **141**, 1199-1207.

De los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2013). Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* **193**, 327-345.

De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* **182**, 375-385.

De Swart, E.A.M., Groenwold, R., Stam, P. Voorrips, R.E., 2007. QTLs for growth and growth related traits in Capsicum annuum L. In: E.A.M. de Swart: Potential for breeding sweet pepper adapted to cooler growing conditions. PhD Thesis, *Wageningen University*, p. 75-92.

deVicente M. C. & Tanksley, S. D. (1993). QTL analysis of transgressive segregation in an interspecific tomato cross. *Genetics*, **134**(2), 585-596.

DeWitt, D., and Bosland, P. W. (1996). "Peppers of the world. An identification guide," Ten Speed Press.

Dhondt, S., Wuyts, N., and Inzé, D. (2013). Cell to whole-plant phenotyping: the best is yet to come. *Trends in plant science* **18**, 428-439.

Efron, B., and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician* **37**, 36-48.

Ehrenreich, I. M., Hanzawa, Y., Chou, L., Roe, J. L., Kover, P. X., & Purugganan, M. D. (2009). Candidate gene association mapping of Arabidopsis flowering time. *Genetics*, **183**(1), 325-335.

Ersoz, E. S., Yu, J., and Buckler, E. S. (2007). Applications of linkage disequilibrium and association mapping in crop plants. *In* "Genomics-assisted crop improvement", pp. 97-119. Springer.

Fabre, J., Dauzat, M., Nègre, V., Wuyts, N., Tireau, A., Gennari, E., Neveu, P., Tisné, S., Massonnet, C., and Hummel, I. (2011). PHENOPSIS DB: an information system for Arabidopsis thaliana phenotypic data in an environmental context. *BMC plant biology* **11**, 77.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses.*Behavior research methods*, **41**(4), 1149-1160.

Ferreira, A., Silva, M. F. d., and Cruz, C. D. (2006). Estimating the effects of population size and type on the accuracy of genetic maps. *Genetics and Molecular Biology* **29**, 187-192.

Fiorani, F., and Schurr, U. (2013). Future scenarios for plant phenotyping. *Annual review of plant biology* **64**, 267-291.

Fitzpatrick, M. J., Ben-Shahar, Y., Smid, H. M., Vet, L. E., Robinson, G. E., & Sokolowski, M. B. (2005). Candidate genes for behavioural ecology. *Trends in Ecology & Evolution*, **20(2)**, 96-104.

Fournier-Level, A., Wilczek, A. M., Cooper, M. D., Roe, J. L., Anderson, J., Eaton, D., Moyers, B. T., Petipas, R. H., Schaeffer, R. N., and Pieper, B. (2013). Paths to selection on life history loci in different natural environments across the native range of Arabidopsis thaliana. *Molecular ecology* **22**, 3552-3566.

Freimer, N., and Sabatti, C. (2004). The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. *Nature genetics* **36**, 1045-1051.

Furbank, R. T., and Tester, M. (2011). Phenomics–technologies to relieve the phenotyping bottleneck. *Trends in plant science* **16**, 635-644.

Galesloot, T. E., van Steen, K., Kiemeney, L. A., Janss, L. L., and Vermeulen, S. H. (2014). A Comparison of Multivariate Genome-Wide Association Methods. *PloS one* **9**, e95923.

Gianola, D., and van Kaam, J. B. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* **178**, 2289-2303.

Gianola, D., Okut, H., Weigel, K. A., & Rosa, G. J. (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC genetics*, ***12*(1)**, 87.

Gianola, D., R. L. Fernando, and A. Stella, (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* **173**: 1761–1776.

Goddard, M. E., and Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* **10**, 381-391.

González-Recio, O., and S. Forni, 2011 Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genet. Sel. Evol*. **43**: 1–12.

Granier, C., and Vile, D. (2014). Phenotyping and beyond: modelling the relationships between traits. *Current opinion in plant biology* **18**, 96-102.

Grigoryev, D. N., Ma, S. F., Irizarry, R. A., Ye, S. Q., Quackenbush, J., & Garcia, J. G. (2004). Orthologous gene-expression profiling in multi-species models: search for candidate genes. *Genome Biol*, **5(5)**, R34.

Gupta, P. K., & Rustgi, S. (2004). Molecular markers from the transcribed/expressed region of the genome in higher plants. *Functional & integrative genomics*, *4*(3), 139-162.

Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet Sel Evol*, **42**:5.

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

Hackett, C. A., Meyer, R. C., and Thomas, W. T. B. (2001). Multi-trait QTL mapping in barley using multivariate regression. *Genetics Research* **77**, 95-106.

Hackett, C.A., (2002). Statistical methods for QTL mapping in cereals. *Plant Molecular Biology* **48**, 585-599.

Hageman, R. S., Leduc, M. S., Korstanje, R., Paigen, B., and Churchill, G. A. (2011). A Bayesian framework for inference of the genotype–phenotype map for segregating populations. *Genetics* **187**, 1163-1170.

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

Hammer, G., Cooper, M., Tardieu, F., Welch, S., Walsh, B., van Eeuwijk, F., Chapman, S., and Podlich, D. (2006). Models for navigating biological complexity in breeding improved crop plants. *Trends in Plant Science* **11**, 587-593.

Hartmann, A., Czauderna, T., Hoffmann, R., Stein, N., and Schreiber, F. (2011). HTPheno: an image analysis pipeline for high-throughput plant phenotyping. *BMC bioinformatics* **12**, 148.

Hastie T, Tibshirani R, Friedman J, Franklin J (2009). The elements of statistical learning: data mining, inference and prediction. *Math Intelligencer*; **27(2)**:83–5

Hayes B, Bowman P, Chamberlain A, Verbyla K, Goddard M (2009). Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet Sel Evol*, **41**:51.

Hayes, B., Bowman, P., Chamberlain, A., and Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* **92**, 433.

Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Science* **52**, 146-160.

Higashide, T., and Heuvelink, E. (2009). Physiological and Morphological Changes Over the Past 50 Years in Yield Components in Tomato. *Journal of the American Society for Horticultural Science* **134**, 460-465.

Hill, W., and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226-231.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with discussion and rejoinder by authors). *Statistical science* **14**, 382-417.

Holland, J. B. (2001). Epistasis and plant breeding. In: *Janick J, editor, Plant Breeding Reviews*, Volume **21**, pp. 27 - 92. Hoboken, NJ: John Wiley & Sons, Inc.

Holland, J. B. (2007). Genetic architecture of complex traits in plants. *Current Opinion in Plant Biology* **10**, 156-161.

Horgan, G. W., Song, Y., Glasbey, C. A., van der Heijden, G. W., Polder, G., Dieleman, J. A., ... & van Eeuwijk, F. A. (2015). Automated estimation of leaf area development in sweet pepper plants from image analysis. *Functional Plant Biology*, **42**(5), 486-492.

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the next challenge. *Nature Reviews Genetics* **11**, 855-866.

Huang, B. E., George, A. W., Forrest, K. L., Kilian, A., Hayden, M. J., Morell, M. K., and Cavanagh, C. R. (2012). A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant biotechnology journal* **10**, 826-839.

Huang, B. E., Verbyla, K. L., Verbyla, A. P., Raghavan, C., Singh, V. K., Gaur, P., ... & Cavanagh, C. R. (2015). MAGIC populations in crops: current status and future prospects. *Theoretical and Applied Genetics*, **128**(6), 999-1017.

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**(11), 4488-4493.

Hung, H. Y., Browne, C., Guill, K., Coles, N., Eller, M., Garcia, A., ... & Holland, J. B. (2012). The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity*, ***108***(5), 490-499.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* **19**, 2271-2282.

Jannink, J.-L., and Walsh, B. (2002). Association mapping in plant populations. *Quantitative genetics, genomics and plant breeding*, 59-68.

Jannink, J.L., Bink, M.C., & Jansen, R. C. (2001). Using complex plant pedigrees to map valuable genes. *Trends in plant science* **6**, 337-342.

Jannink, J.L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics* **9**, 166-177.

Jansen, R. C. (1993). INTERVAL MAPPING OF MULTIPLE QUANTITATIVE TRAIT LOCI. *Genetics* **135**, 205-211.

Jansen, R. C., & Nap, J. P. (2001). Genetical genomics: the added value from segregation. *TRENDS in Genetics*, *17*(7), 388-391.

Jansen, R. C., and Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447-1455.

Janss, L. (2011). bayz manual. Leiden, the Netherlands: Bayesian Solutions.

Jia, Y., and Jannink, J.-L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192, 1513-1522.

Jiang, C. J., and Z. B. Zeng, (1995). Multiple-Trait Analysis of Genetic-Mapping for Quantitative Trait Loci. *Genetics* **140**: 1111-1127.

Johnson, R.A., Wichern, D.W., (2002). Applied multivariate statistical analysis. Prentice Hall.

Kao, C.-H. (2000). On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics* **156**, 855-865.

Kao, C.-H., Zeng, Z.-B., and Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203-1216.

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, **2**(2), 101-118.

Kargbo, A., and Wang, C. Y. (2010). Complex traits mapping using introgression lines in pepper (Capsicum annuum). *African Journal of Agricultural Research* **5**, 725-731.

Kariya, T. (1981). A robustness property of Hotelling's T^2-test. *Ann. Statist*. **9**, 211-214.

Keurentjes, J. J., Willems, G., van Eeuwijk, F., Nordborg, M., & Koornneef, M. (2011). A comparison of population types used for QTL mapping in Arabidopsis thaliana. *Plant Genetic Resources*, **9**(02), 185-188.

Kim, H.J., Han, J.H., Kim, S., Lee, H.R., Shin, J.S., Kim, J.H., Cho, J., Kim, Y.H., Lee, H.J., Kim, B.D., Choi, D., (2011). Trichome density of main stem is tightly linked to PepMoV resistance in chili pepper (Capsicum annuum L.). *Theoretical and Applied Genetics* **122**, 1051-1058.

Kim, K.T., Choi, H.S., Chae, Y., Oh, D.G., Kim, B.D., (2004). Mapping QTL associated with Phytophthora root rot resistance in chilli (Capsicum annuum). *In: McCreight, J.D., Ryder, E.J. (Eds.), Advances in Vegetable Breeding*, pp. 251-255.

Klasen, J. R., Piepho, H. P., and Stich, B. (2012). QTL detection power of multi-parental RIL populations in Arabidopsis thaliana. *Heredity* **108**, 626-632.

Klukas, C., Chen, D., & Pape, J. M. (2014). Integrated analysis platform: an open-source information system for high-throughput plant phenotyping. *Plant physiology*, ***165***(2), 506-518.

Knott, S. A., and Haley, C. S. (2000). Multitrait least squares for quantitative trait loci detection. *Genetics* **156**, 899-911.

Knott, S., and Haley, C. (1992). Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genetical Research* **60**, 139-151.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *In "IJCAI"*, Vol. **14**, pp. 1137-1145.

Korol AB, Ronin YI, Kirzhner VM (1995). Interval mapping of quantitative trait loci employing correlated trait complexes. *Genetics* **140**, 1137-1147.

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant methods* **9**, 29.

Korte, A., Vilhjalmsson, B. J., Segura, V., Platt, A., Long, Q., and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics* **44**, 1066-1071.

Kümmerlen, B., Dauwe, S., Schmundt, D., and Schurr, U. (1999). Thermography to measure water relations of plant leaves. *Handbook of computer vision and applications* **3**, 763-781.

Lampinen J, Vehtari A (2001): Bayesian approach for neural networks review and case studies. *Neural Networks*, **14**:257-274.

Lande, R., and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**, 743-756.

Lander, E. S., and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185-99.

Lauritzen, S. L. (1996). Graphical models, *volume* **17** *of Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York.

Lee, H.R., Cho, M.C., Kim, H.J., Park, S.W., Kim, B.D., 2008. Marker Development for Erect versus Pendant-Orientated Fruit in Capsicum annuum L. *Molecules and Cells* **26**, 548-553.

Lefebvre V (2005) Molecular markers for genetics and breeding: development and use in pepper (Capsicum spp.). *In: Lörz H and Wenzel G (eds) Molecular marker systems in plant breeding and crop improvement. Biotechnology in Agriculture and Forestry*, vol **55**. Springer, Berlin, pp 189–214

Lefebvre, V., Daubeze, A.M., van der Voort, J.R., Peleman, J., Bardin, M., Palloix, A., (2003). QTL for resistance to powdery mildew in pepper under natural and artificial infections. *Theoretical and Applied Genetics* **107**, 661-666.

Lefebvre, V., Kuntz, M., Camara, B., Palloix, A., (1998). The capsanthin-capsorubin synthase gene: a candidate gene for the y locus controlling the red fruit colour in pepper. *Plant Molecular Biology* **36**, 785-789.

Lefebvre, V., Palloix, A., (1996). Both epistatic and additive effects of QTL are involved in polygenic induced resistance to disease: A case study, the interaction pepper - Phytophthora capsici Leonian. *Theoretical and Applied Genetics* **93**, 503-511.

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., and Fritz, S. (2011). Improved Lasso for genomic selection. *Genetics Research* **93**, 77.

Li, C. (1991). Method of path coefficients: a trademark of Sewall Wright. *Human biology*, 1-17.

Li, J., & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 95, 221-227.

Li, L., Long, Y., Zhang, L., Dalton-Morgan, J., Batley, J., Yu, L., ... & Li, M. (2015). Genome wide analysis of flowering time trait in multiple environments via high-throughput genotyping technique in Brassica napus L. *PloS one*, **10**(3), e0119425.

Li, R., Tsaih, S.-W., Shockley, K., Stylianou, I. M., Wergedal, J., Paigen, B., and Churchill, G. A. (2006). Structural model analysis of multiple quantitative traits. *PLoS genetics* **2**, e114.

Li, Y., Tesson, B. M., Churchill, G. A., and Jansen, R. C. (2010). Critical reasoning on causal inference in genome-wide linkage and association studies. *Trends in genetics* **26**, 493-498.

Li, Y.-F., Kennedy, G., Ngoran, F., Wu, P., and Hunter, J. (2013). An ontology-centric architecture for extensible scientific data management systems. *Future Generation Computer Systems* **29**, 641-653.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., Schabenberger, O., 2006. SAS for mixed models. *SAS Institute*, Cary, NC.

Lynch, M., and Walsh, B. (1998). "Genetics and analysis of quantitative traits," Sinauer.

Mackay, T. F. C. (2001). The genetic architecture of quantitative traits. *Annual Review of Genetics* **35**, 303-339.

Mackay, T. F., Stone, E. A., & Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, ***10***(8), 565-577.

MacMillan, K., Emrich, K., Piepho, H. P., Mullins, C. E., and Price, A. H. (2006). Assessing the importance of genotype x environment interaction for root traits in rice using a mapping population II: conventional QTL analysis. *TAG* **113**, 953-964.

Malosetti, M. (2006). "Mixed model methodology for the identification of genetical factors underlying trait variations in plants," PhD thesis, Wageningen Universiteit.

Malosetti, M., Ribaut, J. M., Vargas, M., Crossa, J., and van Eeuwijk, F. A. (2008). A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (Zea mays L.). *Euphytica* **161**, 241-257.

Malosetti, M., Ribaut, J.-M., and 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**.

Malosetti, M., Visser, R. G. F., Celis-Gamboa, C., and Eeuwijk, F. A. (2006). QTL methodology for response curves on the basis of non-linear mixed models, with an illustration to senescence in potato. *Theoretical and Applied Genetics* **113**, 288-300.

Malosetti, M., Voltas, J., Romagosa, I., Ullrich, S. E., and van Eeuwijk, F. A. (2004). Mixed models including environmental covariables for studying QTL by environment interaction. *Euphytica* **137**, 139-145.

Malosetti, M., Bustos-Korts, D., Boer, M. P., & van Eeuwijk, F. A. (2016). Predicting Responses in Multiple Environments: Issues in Relation to Genotype × Environment Interactions. *Crop Science*.

Marcelis, L. (1996). Sink strength as a determinant of dry matter partitioning in the whole plant. *Journal of Experimental Botany* **47**, 1281.

Marcelis, L. F. M., Elings, A., Dieleman, J. A., De Visser, P. H. B., Brajeul, E., Bakker, M. J., and Heuvelink, E. (2006). Modelling dry matter production and partitioning in sweet pepper. *In "Acta Horticulturae",* Vol. **718**, pp. 121-128.

Marcelis, L. F. M., Heuvelink, E., and Goudriaan, J. (1998). Modelling biomass production and yield of horticultural crops: A review. *Scientia Horticulturae* **74**, 83-111.

Martinez, O., and Curnow, R. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480-488.

MAYES, S., PARSLEY, K., SYLVESTER-BRADLEY, R., MAY, S., and FOULKES, J. (2005). Integrating genetic information into plant breeding programmes: how will we produce varieties from molecular variation, using bioinformatics? *Annals of applied biology* **146**, 223-237.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics* **21**, 1087-1092.

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

Mimura, Y., Kageyama, T., Minamiyama, Y., Hirai, M., (2009). QTL Analysis for Resistance to Ralstonia solanacearum in Capsicum Accession 'LS2341'. *Journal of the Japanese Society for Horticultural Science* **78**, 307-313.

Mimura, Y., Minamiyama, Y., Sano, H., and Hirai, M. (2010). Mapping for Axillary Shooting, Flowering Date, Primary Axis Length, and Number of Leaves in Pepper (Capsicum annuum). *Journal of the Japanese Society for Horticultural Science* **79**, 56-63.

Minamiyama, Y., Tsuro, M., Kubo, T., Hirai, M., (2007). QTL analysis for resistance to Phytophthora capsici in pepper using a high density SSR-based map. *Breeding Science* **57**, 129-134.

Montes, J. M., Melchinger, A. E., and Reif, J. C. (2007). Novel throughput phenotyping platforms in plant genetic studies. *Trends in plant science* **12**, 433-436.

Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, H. W. Raadsma et al., (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet. Sel. Evol*. **41**: 56.

Muller, K. E., LaVange, L. M., Ramey, S. L., & Ramey, C. T. (1992). Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications. *Journal of the American Statistical Association*, **87**(420), 1209–1226. http://doi.org/10.1080/01621459.1992.10476281

Munns, R., James, R. A., Sirault, X. R., Furbank, R. T., and Jones, H. G. (2010). New phenotyping methods for screening wheat and barley for beneficial responses to water deficit. *Journal of Experimental Botany*, erq199.

Nadeau, J. H., Burrage, L. C., Restivo, J., Pao, Y.-H., Churchill, G., and Hoit, B. D. (2003). Pleiotropy, homeostasis, and functional networks based on assays of cardiovascular traits in genetically randomized populations. *Genome research* **13**, 2082-2091.

Neto, E. C., Ferrara, C. T., Attie, A. D., and Yandell, B. S. (2008). Inferring causal phenotype networks from segregating populations. *Genetics* **179**, 1089-1100.

Neto, E. C., Keller, M. P., Attie, A. D., and Yandell, B. S. (2010). Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The annals of applied statistics* **4**, 320.

Neto, E.C., A. T. Broman, M. P. Keller, A. D. Attie, B. Zhang et al., (2013). Modeling causality for pairs of phenotypes in system genetics. *Genetics* **193**: 1003–1013.

Nicolaï, M., Cantet, M., Lefebvre, V., Sage-Palloix, A.-M., and Palloix, A. (2013). Genotyping a large collection of pepper (Capsicum spp.) with SSR loci brings new evidence for the wild origin of cultivated C. annuum and the structuring of genetic diversity by human selection of cultivar types. *Genetic resources and crop evolution*, **60**(8), 2375-2390.

Nicolaï, M., Pisani, C., Bouchet, J., Vuylsteke, M., and Palloix, A. (2012). Discovery of a large set of SNP and SSR genetic markers by high-throughput sequencing of pepper (Capsicum annuum). *Genetics and Molecular Research* **11**, 2295-2300.

Nunome, T., Ishiguro, K., Yoshida, T., & Hirai, M. (2001), development traits in eggplant (Solanum melongena L.) based on RAPD and AFLP markers. *Breeding science*, **51**(1), 19-26

O'Hara, R. B., and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis* **4**, 85-117.

Okut, H., D. Gianola, G. J. M. Rosa, and K. A. Weigel, 2011 Prediction of body mass index in mice using dense molecular markers and a regularized neural network. *Genet. Res*. **93**: 189–201.

Onogi, A., Watanabe, M., Mochizuki, T., Hayashi, T., Nakagawa, H., Hasegawa, T., & Iwata, H. (2016). Toward integration of genomic selection with crop modelling: the development of an integrated approach to predicting rice heading dates. *Theoretical and Applied Genetics*, 1-13.

Panozzo, J. F., Eckermann, P. J., Mather, D. E., Moody, D. B., Black, C. K., Collins, H. M., Barr, A. R., Lim, P., and Cullis, B. R. (2007). QTL analysis of malting quality traits in two barley populations. *Australian Journal of Agricultural Research* **58**, 858-866.

Park, T., and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103**, 681-686.

Pasyukova, E. G., Vieira, C., and Mackay, T. F. (2000). Deficiency mapping of quantitative trait loci affecting longevity in Drosophila melanogaster. *Genetics* **156**, 1129-1146.

Paterson, A. H., Tanksley, S. D., and Sorrells, M. E. (1991). DNA markers in plant improvement. *Advances in Agronomy* **46**, 39-90.

Payne, R., Murray, D., Harding, S., Baird, D., and Soutar, D. (2011). An introduction to GenStat for Windows. *VSN International*: Hemel Hempstead, UK.

Piepho, H.-P. (2000). A Mixed-Model Approach to Mapping Quantitative Trait Loci in Barley on the Basis of Multiple Environment Data. *Genetics* **156**, 2043-2050.

Piepho, H.P., E.R. Williams and M. Fleck, (2006). A note on the analysis of designed experiments with complex treatment structure. *HortScience* **41**: 446-452.

Piepho, H.P., Moehring, J., (2007). Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* **177**, 1881-1888.

Pillai, K. C. S. (1985). Hotelling's T^2. *In Encyclopedia of Statistical Sciences* **6** (Edition by S. Kotz, N. L. Johnson and C. B. Read), 669-673. Wiley, New York.

Platt, A., Vilhjálmsson, B. J., and Nordborg, M. (2010). Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**, 1045-1052.

Powell, W., Machray, G. C., and Provan, J. (1996). Polymorphism revealed by simple sequence repeats. *Trends in plant science* **1**, 215-222.

Rakshit, S., Rakshit, A., & Patil, J. V. (2012). Multiparent intercross populations in analysis of quantitative traits. *Journal of genetics*, **91**(1), 111-117.

Rao, G. U., Ben Chaim, A., Borovsky, Y., and Paran, I. (2003). Mapping of yield-related QTLs in pepper in an interspecific cross of Capsicum annuum and C. frutescens. *Theoretical and Applied Genetics* **106**, 1457-1466.

R-Development-Core-Team, 2011. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria.

Resende, M. F. R., Muñoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., Jokela, E. J., Martin, T. A., Peter, G. F., and Kirst, M. (2012). Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). *Genetics* **190**, 1503-1510.

Reymond, M., Muller, B., Leonardi, A., Charcosset, A., and Tardieu, F. (2003). Combining quantitative trait Loci analysis and an ecophysiological model to analyze the genetic variability

of the responses of maize leaf growth to temperature and water deficit. *Plant Physiology* **131**, 664-75.

Ribaut, J. M., and Hoisington, D. (1998). Marker-assisted selection: new tools and strategies. *Trends in Plant Science* **3**, 236-239.

Robertson, A. (1967). The nature of quantitative genetic variation. *Heritage from Mendel*, 265-280.

Rosa, G. J., Valente, B. D., de Los Campos, G., Wu, X.-L., Gianola, D., and Silva, M. A. (2011). Inferring causal phenotype networks using structural equation models. *Genet Sel Evol* **43**(6).

Rosyara, U. R., Gonzalez-Hernandez, J. L., Glover, K. D., Gedye, K. R., and Stein, J. M. (2009). Family-based mapping of quantitative trait loci in plant breeding populations with resistance to Fusarium head blight in wheat as an illustration. *TAG* **118**, 1617-1631.

Saini, S. S., & Sharma, P. P. (1978). Inheritance of resistance to fruit rot (Phytophthora capsici Leon.) and induction of resistance in bell pepper (Capsicum annuum L.). *Euphytica*, 27(3), 721-723.

Sari-Gorla, M., Calinski, T., Kaczmarek, Z., and Krajewski, P. (1997). Detection of QTL× environment interaction in maize by a least squares interval mapping method. *Heredity* **78**.

SAS-Institute (2011). "Sas/graph 9. 3: Reference," *SAS Institute*.

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

Saxton, A., 2004. Genetic Analysis of Complex Traits Using SAS. SAS Publ., Cary.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., and Zhang, C. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics* **37**, 710-717.

Seelig, H. D., Hoehn, A., Stodieck, L., Klaus, D., Adams Iii, W., and Emery, W. (2008). The assessment of leaf water content using leaf reflectance ratios in the visible, near-, and short-wave-infrared. *International Journal of Remote Sensing* **29**, 3701-3713.

Semagn, K., Bjørnstad, Å., and Ndjiondjop, M. (2006). An overview of molecular marker methods for plants. *African Journal of Biotechnology* **5**.

Shoemaker, J. S., Painter, I. S., and Weir, B. S. (1999). Bayesian statistics in genetics: a guide for the uninitiated. *Trends in Genetics* **15**, 354-358.

Sillanpää, M. J., and Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**, 1373-1388.

Silva, L. d. C. e., Cruz, C. D., Moreira, M. A., and Barros, E. G. d. (2007). Simulation of population size and genome saturation level for genetic mapping of recombinant inbred lines (RILs). *Genetics and Molecular Biology* **30**, 1101-1108.

Sisson, S., and Hurn, M. (2004). Bayesian point estimation of quantitative trait loci. *Biometrics* **60**, 60-68.

Slafer, G. (2003). Genetic basis of yield as viewed from a crop physiologist's perspective. *Annals of Applied Biology* **142**, 117-128.

Slafer, G. A., & Kernich, G. C. (1996). Have changes in yield (1900-1992) been accompanied by a decreased yield stability in Australian cereal production?. *Crop and Pasture Science*, *47*(3), 323-334

Snape, J. W., Butterworth, K., Whitechurch, E., & Worland, A. J. (2001). Waiting for fine times: genetics of flowering time in wheat. *Euphytica*, *119*(1-2), 185-190.

Soller, M., Brody, T., and 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.

Solti, Á., Lenk, S., Mihailova, G., Mayer, P., Barócsi, A., and Georgieva, K. (2014). Effects of habitat light conditions on the excitation quenching pathways in desiccating *Haberlea rhodopensis* leaves: An Intelligent FluoroSensor study. *Journal of Photochemistry and Photobiology B: Biology* **130**, 217-225.

Song, Y., Glasbey, C., Horgan, G., Polder, G., Dieleman, J., and van der Heijden, G. (2014). Automatic fruit recognition and counting from multiple images. *Biosystems Engineering* **118**, 203-215.

Sørensen, L. P., Janss, L., Madsen, P., Mark, T., and Lund, M. S. (2012). Estimation of (co) variances for genomic regions of flexible sizes: application to complex infectious udder diseases in dairy cattle. *Genetics Selection Evolution* **44**, 18.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). "Causation, prediction, and search," MIT press.

Spitters, C., and Schapendonk, A. (1990). Evaluation of breeding strategies for drought tolerance in potato by means of crop growth simulation. *Plant and Soil* **123**, 193-203.

Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PloS one* **8**, e65245.

Sugita, T., Yamaguchi, K., Kinoshita, T., Yuji, K., Sugimura, Y., Nagata, R., Kawasaki, S., Todoroki, A., (2006). QTL analysis for resistance to phytophthora blight (Phytophthora capsici Leon.) using an intraspecific doubled-haploid population of Capsicum annuum. *Breeding Science* **56**, 137-145.

Sukhwinder, S., Hernandez, M. V., Crossa, J., Singh, P. K., Bains, N. S., Singh, K., and Sharma, I. (2012). Multi-Trait and Multi-Environment QTL Analyses for Resistance to Wheat Diseases. *PLoS ONE* **7**, e38008.

Sun, G., & Schliekelman, P. (2011). A genetical genomics approach to genome scans increases power for QTL mapping. *Genetics*, **187**(3), 939-953.

Syvänen, A.-C. (2005). Toward genome-wide SNP genotyping. *Nature genetics* **37**, S5-S10.

Takeda, H., Farsiu, S., & Milanfar, P. (2007). Kernel regression for image processing and reconstruction. *Image Processing, IEEE Transactions on*,***16***(2), 349-366.

Tanksley, S., Young, N., Paterson, A., and Bonierbale, M. (1989). RFLP mapping in plant breeding: new tools for an old science. *Nature Biotechnology* **7**, 257-264.

Tardieu, F. (2003). Virtual plants: modelling as a tool for the genomics of tolerance to water deficit. *Trends in Plant Science* **8**, 9-14.

Tasaki, S., Sauerwine, B., Hoff, B., Toyoshiba, H., Gaiteri, C., & Neto, E. C. (2015). Bayesian Network Reconstruction Using Systems Genetics Data: Comparison of MCMC Methods. *Genetics*, **199**(4), 973-989.

Technow, F., Messina, C. D., Totir, L. R., & Cooper, M. (2015). Integrating crop growth models with whole genome prediction through approximate Bayesian computation. *PloS one*, **10**(6), e0130855.

Thabuis, A., Palloix, A., Pflieger, S., Daubeze, A.M., Caranta, C., Lefebvre, V., 2003. Comparative mapping of Phytophthora resistance loci in pepper germplasm: evidence for conserved resistance loci across Solanaceae and for a large genetic diversity. *Theoretical and Applied Genetics* **106**, 1473-1485.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Uptmoor, R., Schrag, T., Stützel, H., and Esch, E. (2008). Crop model based QTL analysis across environments and QTL based estimation of time to floral induction and flowering in Brassica oleracea. *Molecular Breeding* **21**, 205-216.

Valente, B. D., Rosa, G. J. M., Gianola, D., Wu, X.-L., and Weigel, K. (2013). Is Structural Equation Modeling Advantageous for the Genetic Improvement of Multiple Traits? *Genetics* **194**, 561-572.

Van der Heijden, G., Song, Y., Horgan, G., Polder, G., Dieleman, A., Bink, M., Palloix, A., van Eeuwijk, F., and Glasbey, C. (2012). SPICY: towards automated phenotyping of large pepper plants in the greenhouse. *Functional Plant Biology* **39**, 870-877.

Van Eeuwijk, F. (2015). How to dissect complex traits and how to choose suitable mapping resources for system genetics?: Comment on" Mapping complex traits as a dynamic system" by L. Sun and R. Wu. *Physics of life reviews*.

Van Eeuwijk, F. A., Bink, M., Chenu, K., and Chapman, S. C. (2010). Detection and use of QTL for complex traits in multiple environments. *Current Opinion in Plant Biology* **13**, 193-205.

Van Ittersum, M., Leffelaar, P., Van Keulen, H., Kropff, M., Bastiaans, L., and Goudriaan, J. (2003). On approaches and applications of the Wageningen crop models. *European Journal of Agronomy* **18**, 201-234.

Vargas, M., van Eeuwijk, F., Crossa, J., and Ribaut, J.-M. (2006). Mapping QTLs and QTL × environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods. *Theoretical and Applied Genetics* **112**, 1009-1023.

Varshney, R. K., Chabane, K., Hendre, P. S., Aggarwal, R. K., and Graner, A. (2007). Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. *Plant Science* **173**, 638-649.

Verbeke, G., and Molenberghs, G. (2000). "Linear Mixed Models for Longitudinal Data" Springer.

Verbyla, A. P., Eckermann, P. J., Thompson, R., and Cullis, B. R. (2003). The analysis of quantitative trait loci in multi-environment trials using a multiplicative mixed model. *Australian Journal of Agricultural Research* **54**, 1395-1408.

Verbyla, A. P., George, A. W., Cavanagh, C. R., & Verbyla, K. L. (2014). Whole-genome QTL analysis for MAGIC. *Theoretical and Applied Genetics*,**127**(8), 1753-1770.

Vilhjálmsson, B. J., and Nordborg, M. (2012). The nature of confounding in genome-wide association studies. *Nature Reviews Genetics* **14**, 1-2.

Visscher, P. M., & Goddard, M. E. (2004). Prediction of the confidence interval of quantitative trait loci location. *Behavior genetics*, 34(4), 477-482.

Voorrips, R. E., Palloix, A., Dieleman, J. A., Bink, M. C. A. M., Heuvelink, E., Heijden, G. W. A. M. v. d., Vuylsteke, M., Glasbey, C., Barócsi, A., Magán, J., and Eeuwijk, F. A. v. (2010). Crop growth models for the -omics era: the EU-SPICY project. *In "Advances in Genetics and Breeding of Capsicum and Eggplant : Proceedings of the XIVth EUCARPIA Meeting on genetics and breeding of Capsicum and Eggplant"*, pp. 315-321. Editorial Universidad Politécnica de Valencia, Valencia, Spain, Valencia, Spain.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van De Lee, T., Hornes, M., Friters, A., Pot, J., Paleman, J., and Kuiper, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic acids research* **23**, 4407-4414.

VSNi (2012). GenStat for Windows 15th Edition. *VSN International*, Hemel Hempstead, UK.

Wang, L.H., Zhang, B.X., Lefebvre, V., Huang, S.W., Daubeze, A.M., Palloix, A., (2004). QTL analysis of fertility restoration in cytoplasmic male sterile pepper. *Theoretical and Applied Genetics* **109**, 1058-1063.

Weller, J., and Soller, M. (2004). An analytical formula to estimate confidence interval of QTL location with a saturated genetic map as a function of experimental design. *Theoretical and Applied Genetics* **109**, 1224-1229.

West-Eberhard, M. J. (2003). "Developmental plasticity and evolution," *Oxford University Press*, New York [etc.].

Williams, E.R., John, J.A., (1999). Construction of resolvable designs with nested treatment structure. *Biometrical Journal* **41**, 341-349.

Wimmer, V., Lehermeier, C., Albrecht, T., Auinger, H. J., Wang, Y., & Schön, C. C. (2013). Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics*, ***195***(2), 573-587.

Wright, S. (1921). Correlation and causation. *Journal of agricultural research* **20**, 557-585.

Wu, S., Li, H., Casella G. (2006). Tests with optimal average power in multivariate analysis. *Statistica Sinica* **16**, 255-266.

Wubs, A. M., Heuvelink, E., and Marcelis, L. F. M. (2009). Abortion of reproductive organs in sweet pepper (Capsicum annuum L.): a review. *Journal of Horticultural Science and Biotechnology* **84**, 467-475.

Würschum, T. (2012). Mapping QTL for agronomic traits in breeding populations. *Theoretical and Applied Genetics* **125**, 201-210.

www.spicyweb.eu Smart tools for Prediction and Improvements of Crop Yield – KBBE 211347. (F. A. van Eeuwijk, ed.).

Xu, S. (2013). Mapping QTL for Multiple Traits. *In "Principles of Statistical Genomics"*, pp. 209-222. Springer New York.

Xu, Y., Hu, W., Yang, Z., & Xu, C. (2015). A multivariate partial least squares approach to joint association analysis for multiple correlated traits. *The Crop Journal*.

Yi, N., Shriner, D., Banerjee, S., Mehta, T., Pomp, D., and Yandell, B. S. (2007). An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. *Genetics* **176**, 1865-1877.

Yin, X., and Struik, P. C. (2010). Modelling the crop: from system dynamics to systems biology. *Journal of Experimental Botany* **61**, 2171-2183.

Yin, X., Struik, P. C., Eeuwijk, v. F. A., Stam, P., and Tang, J. (2005). QTL analysis and QTL-based prediction of flowering phenology in recombinant inbred lines of barley. *Journal of Experimental Botany* **56**, 967-976.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457-68.

Zeng, Z.-B., Kao, C.-H., and Basten, C. J. (1999). Estimating the genetic architecture of quantitative traits. *Genetical research* **74**, 279-289.

Zhou, X., & Stephens, M. (2014). Efficient algorithms for multivariate linear mixed models in genome-wide association studies. *Nature methods*, **11**(4), 407.

Zygier, S., Chaim, A. B., Efrati, A., Kaluzky, G., Borovsky, Y., and Paran, I. (2005). QTLs mapping for fruit size and shape in chromosomes 2 and 4 in pepper and a comparison of the pepper QTL map with that of tomato. *Theoretical and Applied Genetics* **111**, 437-445.

Summary

Breeders aim at selecting genotypes that show a superior performance for target (often complex) traits in a target population of environments. Target traits are commonly a function of the genotype, determined by a large number of loci, and of the environmental conditions. For most traits, differences between genotypic responses are not constant across environmental conditions, leading to genotype-by-environment interaction (GEI). GEI can also be modelled as an explicit function of quantitative trait loci (QTL: the genomic regions with genetic differences that influence one or more traits) across environments. Here, GEI can be interpreted as differential QTL effect sizes across environments (QTL-by-environment interaction, QEI). QEI can be modelled with a mixed model methodology that explicitly accounts for the change in QTL effect size across environments. QTLs can thus be classified as 'constitutive' if they are consistently detected across most environments. QTLs are said to be 'adaptive' when they are detected only in specific environmental conditions, or when there is a change in the QTL effect with a change in the level of an environmental factor.

A complementary strategy to characterize the genetic basis of the complex target trait (e.g. yield) is to dissect it into a number of physiological component traits using crop growth models (CGMs). The idea is that component traits have a simpler genetic basis and less GEI than the target trait and manipulation of the target trait can proceed via its component traits. For that reason, understanding the interconnectedness among plant phenotypes has become a key objective in QTL mapping. Network type models provide alternative representations of the biological knowledge of a complex target trait such as yield, by showing intricate interactions of multiple genetic (and possibly environmental factors) influencing the target trait. The use of network models allows investigating how plant traits are interconnected in networks of dependencies as a result of gene-to-trait, trait-to-trait and gene-to-gene interactions. Furthermore, we can study the stability of such networks across environments due to GEI.

In this thesis, we present the results of a number of statistical techniques that were used to understand the genetics of yield in pepper as an example of complex trait measured in a number of environments. Main objectives were; i) to propose a number of mixed models to detect QTLs for multiple traits and multiple environments, ii) to extend the multi-trait QTL models to a multi-trait genomic prediction model, iii) to study how well the complex trait yield can be indirectly predicted from its component traits, and iv) to understand the 'causal' relationships between the target trait yield and its component traits.

For this research as part of the EU-SPICY project (http://www.spicyweb.eu/), we have used a bi-parental pepper (*Capsicum annuum*) population comprising 149 individuals from the sixth generation of recombinant inbred lines (RIL) of an intraspecific cross between the large – fruited inbred cultivar 'Yolo Wonder' (YW) and the small-fruited cultivar 'Criollo de Morelos 334' (CM 334). The 149 RILs were characterized genotypically with 455 markers assembled into 12 pepper chromosomes, covering 1705cM. A total of 16

physiological traits were evaluated across four different trials and various types of genetic parameters were estimated. Trait heritabilities were generally large (ranging between 0.43 – 0.96 with an average of 0.86) while many of the traits displayed heterosis and transgression.

In chapters 2 and 3, different multiple-QTL mapping methods were employed to estimate location, heritability and direction of the QTLs. We qualitatively investigated QTL pleiotropy (a QTL region affecting more than one trait) and we discussed our results in the light of previously reported QTLs for these and similar traits in pepper. All the QTLs for yield were constitutive with the majority of the superior alleles coming from parent YW. We assumed that yield would benefit from joint analysis with other traits and so deployed two other mixed model based multi-response QTL approaches: a multi-trait approach (MT) and a multi-trait multi-environment approach (MTME). The approaches were compared in terms of number of QTLs detected for each trait, the explained variance, and the accuracy of prediction for the final QTL model. For yield as well as the other traits, MTME was superior to ME and MT in the number of QTLs, the explained variance and accuracy of predictions. Many of the detected QTLs were pleiotropic and showed quantitative QEI. The results confirmed the feasibility and strengths of novel mixed model QTL methodology to study the architecture of complex traits.

Since the main interest of this research included improvement of complex trait prediction, in chapter 4, we explored both single-trait and multi-trait versions of genomic prediction (GP) models as alternatives to the QTL-based prediction (QP) models. We extended the frontier in this research area by comparing the predictive performances of multi-trait versions of QP and GP models. The methods differed in their predictive accuracies with GP methods outperforming QP methods in both single and multi-traits situations. We further integrated QTL/genomic prediction with CGM approaches and showed that the target trait yield can be predicted via its component traits namely radiation use efficiency (RUE), partitioning into fruits (PF) and growth rate of leaf area index (LAI_{rate}) together with environmental covariables such as temperature, thermal time and daily global radiation intensity (I). The CGM approach was implemented for within-environment and across-environment analyses. The predictive accuracies from the CGM were comparable to the direct prediction strategy. The CGM approach seemed to work well at first sight, but this is especially due to the fact that yield appeared to be strongly driven by just one component, the partitioning to fruits. The across environment CGM indicated that we may use component traits and environmental information from one environment to predict yield in another environment.

In chapter 5, we constructed both conditional and unconditional networks across the four environments. The unconditional networks were based on standard multi-trait model (MTM) while the conditional networks were based on the QTL-driven phenotype network method (QTLnet). The final networks for each environment from both conditional and unconditional methods were used in a structural equation model to assess the causal relationships. Conditioning QTL mapping on network structure via QTLnet improved

detection of refined genetic architecture by distinguishing between QTL with direct and indirect effects, thereby removing non-significant effects found in MTM and resolving QTL hotspots (pleiotropy). The most probable conditional networks from the four environments are similar in skeleton to the relationships defined by the CGM. Similar to the CGM topology, yield was established to be downstream to its three component traits, indicating that yield can be studied and predicted from its component traits. Thus, the genetic improvements of yield would benefit from improvements on the component traits.

Finally, complex trait prediction can be enhanced by a full integration of the methods described in the different chapters. Recent research efforts have been channelled to incorporating both multivariate whole genome prediction models and crop growth models. Further research is required, but we hope that the present thesis presents useful steps towards better prediction models for complex traits exhibiting genotype by environment interaction.

Acknowledgement

I owe a lot of gratitude to so many people for the successful completion of this thesis. It is impossible for me to mention everybody by name, kindly accept my sincere apology if your name is not mentioned. Like the Yoruba proverb “agbajo owo la fi n soya, ajeje owo kan ko gberu dori”, which loosely means “we can boast but with a collective effort; alas, one hand can't lift a heavy load”, I couldn’t have done this alone. Top on the list are members of my supervision team headed by my promotor Professor Fred van Eeuwijk and daily supervisors Alain Palloix (of blessed memory) and Marco Bink for their painstaking supervision and numerous supports during the course of my PhD journey. It is highly unfortunate that Alain couldn’t witness the completion of this thesis. His tremendous professional tutelage and personal supports during the early part of my PhD journey cannot be overstated. Coming from a statistics background, he invited me to GAFL centre of INRA at Montfavet and introduced me to the rudiments of genetics and breeding of pepper plant. During my stay in Montfavet, Alain and his wife hosted me in their home and ensured that I got an extended stay at the centre’s guest apartment. My heartfelt condolence to his wife and family. May his soul rest in peace (Amen). I sincerely appreciate the invaluable guidance, criticisms and commendations I received from Marco. Marco assumed the driver’s seat during the course of this thesis and was always challenging me to be proactive. He always ensured I set definite timeline for my tasks and will not hesitate to stay on my neck to ensure that deadlines are met. I considered myself privileged to have been in the safe hands of Fred. Despite Fred’s extremely busy schedule, my work always received prime time attention from him as his comments to my manuscripts always arrive during the middle of the night. I have benefitted immensely from his insightful comments and editing on my manuscripts. Carrying out some of his suggestions could be challenging but worthwhile.

I would like to say a very big thank you to all my colleagues at Biometris and GAFL for the tremendous professional and social supports I received. To my colleagues in GAFL, I cannot forget our trip to the Alps and numerous dinners in Avignon. The very warm atmosphere at Biometris always made me feel at home. I say a big thank you to Marcos, Martin and Paul Keizer for their supports and the fruitful discussions we had especially on GenStat. I would also like to thank Evert-Jan and Saskia for giving me the opportunity to teach in the advanced statistics course. On my arrival in Wageningen, Gerrit Gort drove me from the train station to my apartment. He was surprised at the sheer size of my home town when he saw some images on the internet. He asked in bewilderment “are there roads among the streets?” I am also very grateful to my fellow PhD colleagues for the bond of friendship. I still hold the birthday party you organized for me in my heart. Thanks so much. To Daniela, you are a gem! To Dinie and Hanneke, thanks a lot for the seamless ways you always handled my secretarial and other needs.

I am highly indebted to all my colleagues in the SPICY project for the data used in this thesis and for their contributions to my manuscripts. I should especially mention Anja Dielema and Juan Jose Magan for the quality of the phenotypic data. The efforts of Alain,

Marys Nicolai, Roeland Voorrips and V. Lefebvre at improving the quality of the marker data are appreciated. In a special way, I would like to thank the management of GAFL centre of INRA for the financial support I received during the first three years of my PhD study.

I deeply appreciate my siblings, relatives and friends for their unconditional love and supports. The past two years has been very challenging but I have been able to cope due to your unflinching supports and encouragements. I sincerely appreciate the numerous show of love, supports and motivations from my brothers: Lukman, Yahaya, Kamorudeen, Buniyaminu and Shakirullah; thank you very much for being part and parcel of my life. To my dear elder brother, Dr. B.A. Alimi, words cannot adequately convey my appreciation. Your reward is with Him. Olalekan Ibraheem, Yemi Adeyemo and Wale Arogundade, thanks a lot for the shoulders of support you provided for my family. I owe you! To my Nigerian and Ghanaian folks in Wageningen, I am very grateful. You provided laughter, the best medicine for the stress, challenges and frustrations of PhD life in Wageningen. You also provided brotherly love and assorted African food on order. Naomie, Aisha, Valerie, Paulina, Ronke, Bidemi, the Hoek's family, Malik, Femi, Tosin, Abdullah and others many thanks for making me always feel wanted and loved.

Finally to the first Dr. in my nuclear family, Omobolanle Atoke and my lovely son, Kiyaan Adesola, thank you, thank you, thank you for the understanding and moral supports through thick and thin. Bola, your sacrifices during our trying moments are already yielding positive results. It's a matter of time! K-bobo, your high level of energy and inquisitiveness are very infectious and pleasing.

Curriculum Vitae

Nurudeen Adeniyi Alimi was born on the 4th of April 1981 in Ibadan, Nigeria. In 1996, he obtained his secondary school leaving certificate and he finished higher national diploma in mechanical engineering from the Polytechnic Ibadan, in 2002. He later obtained bachelor of Agriculture with specialization in Agronomy from University of Ibadan in 2005. He worked briefly as chief operating officer at Suhulat Digital System before proceeding to Universiteit Hasselt in Belgium where he bagged two masters degrees in applied statistics and Biostatistics in 2009. Because of his desire to bridge the gap in his agronomy and statistics degrees, he opted for PhD in plant breeding and statistical genetics which eventually culminated into this thesis. He has attended several short courses on statistical modelling, plant breeding and analysis of omics data. Nurudeen was involved in teaching advanced statistics course for MSc students at Wageningen University. He can be reached at magajinurudeen@gmail.com.

Publication List

Alimi, N. A., et al. (2010). Exploratory QTL analyses of some pepper physiological traits in two environments. Advances in Genetics and Breeding of Capsicum and Eggplant: Proceedings of the XIVth EUCARPIA Meeting on genetics and breeding of Capsicum and Eggplant, Valencia, Spain, Editorial Universidad Politécnica de Valencia, Valencia, Spain.

Alimi, N. A., et al. (2013). "Genetic and QTL analyses of yield and a set of physiological traits in pepper." *Euphytica* 190(2): 181-201.

Alimi, N. A., et al. (2013b). "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. A., et al. (2013c). "Multivariate QTL Analyses and Predictions of Yield Related Traits in Pepper." Advances in Genetics and Breeding of Capsicum and Eggplant : Proceedings of the XVth EUCARPIA Meeting on genetics and breeding of Capsicum and Eggplant, Turin, Italy.

Alimi, N. A., et al. (2016). Accuracies of Predictions of Complex Traits Within and Across Environments: an Application to Yield in Pepper. Submitted to G3.

Alimi, N. A., et al. (?). Causal Networks for Yield Phenotypes Across a range of Environments: A case study of Pepper. To be submitted to TAG.

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)

Review of literature (5.2 ECTS)
- QTL Methodologies for correlated physiological traits across a range of environmental conditions (2010)

Writing of project proposal (3.5 ECTS)
- QTL Methodologies for correlated physiological traits across a range of environmental conditions (2011)

Post-graduate courses (8.3 ECTS)
- Linear models and estimation of genetic parameters; Synbreed Summer School, Garching, Germany (2010)
- The art of modelling; WUR (2010)
- Identity by descent approaches to genomic analysis of genetic traits; WUR (2012)
- Statistical methods for genome-enabled selection; Synbreed Summer School, TUM Institute for Advanced Study, Garching, Germany (2012)

Laboratory training and working visits (4.3 ECTS)
- Insight into the biological/genetical make-up of pepper plant; INRA, Avignon (2010-2012)

Deficiency, refresh, brush-up courses (2.8 ECTS)
- French language course; INRA TALQ Institute (2010-2011)
- Statistical genetics: concepts and methodology; WUR (2010)

Competence strengthening / skills courses (3.1 ECTS)
- PhD Competence assessment; WGS (2010)
- Techniques for writing and presenting a scientific paper; WGS (2011)
- Career perspective; WGS (2013)
- Last stretch of the PhD programme; WGS (2013)

PE&RC Annual meetings, seminars and the PE&RC weekend (1.8 ECTS)
- PE&RC Day (2010)
- PE&RC Weekend (2010)
- PE&RC Weekend last year (2013)

Discussion groups / local seminars / other scientific meetings (6 ECTS)
- Mathematics, Statistics and Modelling in Production Ecology and Research Conservation (Maths & Stats) (2010-2014)
- Biometris Statistical Genetics meeting (2011-2013)
- INRA/Avignon discussion group (2009-2011)

International symposia, workshops and conferences (7.9 ECTS)
- Molecular Aspects of Plant Development; Vienna University of Economics, Austria (2010)
- XIV EUCARPIA meeting on Genetics and Breeding of Capsicum & Eggplant; Valencia, Spain (2010)
- XV EUCARPIA meeting of the Biometris in Plant Breeding Section; Hohenheim, Germany (2012)
- XV EUCARPIA meeting on Genetics and Breeding of Capsicum & Eggplant; Torino, Italy (2013)

Lecturing / supervision of practicals / tutorials (3 ECTS)
- Advanced Statistics (2013-2014)