Transcriptional regulation underlying the quantitative genetic response of maize to grey leaf spot disease

by

Nanette Christie

Submitted in the partial fulfillment of the requirements for the degree

Philosophiae Doctor

in the Faculty of Natural & Agricultural Sciences

Bioinformatics and Computational Biology Unit

Department of Biochemistry

University of Pretoria

Pretoria

2014

Declaration

I, Nanette Christie, declare that the thesis, which I hereby submit for the degree *Philosophiae Doctor* at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature: .....

Date: .....

Acknowledgments

I wish to express my sincere gratitude to the following people:

- Prof. Dave Berger for his valued supervision and leadership on this project, for his patience, motivation, enthusiasm and unique guidance throughout this study (and throughout all my post-graduate years) and in reviewing this thesis. I could not have asked for a better PhD supervisor!
- Prof. Zander Myburg for his co-supervision, highly regarded advice, input, enthusiasm and contribution to the review of this thesis.
- The members from the Maize eQTL team for valuable inputs and support. In particular Prof. Pangirayi Tongoona, Dr. Bridget Crampton, Dr. Shane Murray, Dr. Maryke Carstens, Jeanne Korsman and Jacqueline Meyer for contributing work or data to make this study possible.
- Prof. Fourie Joubert for his support, advice and willingness to assist in this project and Johann Swart for all his help and support as systems administrator.
- Molecular Plant-Pathogen Interactions (MPPI) colleagues and all my labmates in the Bioinformatics and Computational Biology Unit for their support and friendship. In particular, thanks to Pieter Burger and Oliver Bezuidt, for always having time for a cup coffee on the balcony. And a special thanks to Karen van der Merwe for her assistance.
- PANNAR and Technology Innovation Agency (TIA) for their involvement and funding of the project and the National Research Foundation (NRF) of South Africa for funding the project.
- The University of Pretoria and FABI for providing facilities and a stimulating academic environment.

- Barry Christie, my lovely husband, for his endless encouragement, patience, paryers and full support throughout the project and for always believing in me and loving me.

- My parents, Gerhard and Yda Coetzer, for always supporting me in everything I undertake, their positive encouragement throught my life me and for always believing in me.

- My two sisters, Eugenie and Gerda, and grandmother, Ouma Nancy, for their prayers, encouragement and for always being there to listen.

- My in-laws, Henry, Jetty, Willie and Anri Christie, for their patience, support, motivation and love.

- My spiritual family from Every Nation in Pretoria and Nelspruit for wonderful friendship and support. A special thanks to Mirjam Klix, my dear friend from Germany, for encouragement in times when I needed it most!

- My Heavenly Father for giving me talents and the ability to work on this project, for carrying me through every phase and for teaching me to put my trust in Him throughout this project.

Abstract

Cercospora zeina causes grey leaf spot (GLS), a yield-limiting disease on maize. The main objective of this study was to exploit maize gene expression data to dissect the quantitative disease response to *C. zeina* infection. The project addresses the hypothesis that there is an underlying DNA polymorphism that gives rise to a change in gene expression, which in turn affects GLS disease severity. Genomic and functional annotation of the reporters on an Agilent 44K maize microarray was carried out. This microarray was used for global gene expression profiling of earleaf samples collected from 100 recombinant inbred sub-tropical maize lines exposed in the field to *C. zeina*. Gene expression profiles together with GLS severity scores were used in a weighted gene co-expression network analysis to identify co-expression modules associated with disease severity. Quantitative trait locus (QTL) mapping for GLS severity was combined with expression QTL (eQTL) analyses to investigate the molecular basis of the quantitative response to GLS. An eQTL data analysis pipeline was developed in Galaxy. The overlap of phenotypic QTLs with *cis*- and *trans*-eQTLs revealed putative causal candidate genes and potential mechanisms responsible for the QTLs, respectively. Regulatory network models were constructed for *trans*-eQTL hotspots coinciding with phenotypic QTLs. A genetic basis for coordinated expression responses to GLS disease was identified. For the susceptible response, the results lead to the hypothesis that a calmodulin-related protein with a *cis*-eQTL acts as a global regulator of various pathogenesis-related proteins that are activated too late after infection started. For the resistant response, it is hypothesised that a serine threonine-protein kinase with a *cis*-eQTL acts as a post-translational global regulator regulating phosphatases and kinases involved in activation of defense gene expression. The outcomes of this study were: i) the development of a systems genetics strategy and ii) several hypotheses of maize transcriptional responses to *C. zeina* which need to be validated with further studies. These results extend the current knowledge of GLS resistance and could

aid in the improvement of maize varieties.

Preface

A major challenge in current biological research is to understand the molecular basis of quantitative traits. This study is part of a collaborative project investigating the genomics of the quantitative genetic response to grey leaf spot (GLS), caused by the fungus *Cercospora zeina* Crous & U. Braun, in a maize (*Zea mays* ssp. *mays* L.) population derived from two subtropical inbred lines that have been bred for maize growing conditions in southern Africa. The objective was to gain an understanding of the molecular basis of the quantitative genetic response, by identifying genes and pathways associated with GLS disease in maize. The project addresses the hypothesis that there is an underlying DNA polymorphism that gives rise to a change in gene expression, which in turn affects GLS disease severity.

Chapter 1, entitled "Exploiting gene expression data to dissect quantitative traits in plants", provides a comprehensive review of recent literature. It starts with an overview of quantitative trait locus (QTL) mapping and cloning. It then focuses on expression QTL (eQTL) mapping (which aims to dissect the molecular basis of gene expression variation) and associating eQTLs with phenotypic QTLs. Recent global eQTL studies in plants are compared to reveal common features and limitations of current technologies. An overview is given on how systems genetics promises to elucidate the complex molecular networks underlying phenotypic traits. Finally, the host and pathogen are introduced, and an overview of general plant defense mechanisms is given as a basis for the biological question under study.

Chapter 2 describes the development of the Maize Microarray Annotation Database. The aim was the genomic and functional annotation of Agilent 44K microarray reporters, which were used in Chapter 3 for global gene expression profiling on earleaf samples collected from 100 maize recombinant inbred lines (RILs). The annotations are available in the Maize Microarray Annotation Database (`http://MaizeArrayAnnot.bi.up.ac.`

za/), as well as through a GBrowse annotation file that can be uploaded to the MaizeGDB genome browser as a custom track. The content of this chapter has been published in the Plant Methods journal (Coetzer *et al.*, 2011).

Chapter 3 describes how genome-wide gene expression profiles as well as GLS severity scores across the individuals in the maize RIL population were used in a weighted gene co-expression network analysis (WGCNA) to identify gene co-expression modules relating to *C. zeina* disease severity. Hypotheses of driver/hub genes as regulators and of biological processes associated with the GLS disease response are given.

Chapter 4 describes how quantitative trait locus (QTL) mapping for GLS severity was combined with expression QTL (eQTL) analyses to investigate the molecular basis of the quantitative disease response to *C. zeina* infection. It outlines the development of a Galaxy workflow for global eQTL analysis as well as an overlap analyses between phenotypic QTLs and eQTLs. Finally, putative regulatory network models are presented for *trans*-eQTL hotspots coinciding with phenotypic QTLs. Hypotheses of regulators and mechanisms that could explain the phenotypic QTLs are given.

Chapter 5 provides a meta-analysis, where results from the previous two chapters are combined to answer a final question: "What is the nature of the genetic and transcriptional variation affecting responses to grey leaf spot disease in maize?" An overview of the systems genetics strategy that was developed to incorporate the analysis of gene co-expression with phenotypic QTL and eQTL mapping are given in this chapter.

Finally, the thesis ends with concluding remarks in **Chapter 6**. A critical discussion of the strengths and limitations of this study is given, as well as the importance of the results and findings in the context of the scientific field.

Publication from this thesis:

Coetzer, N., Myburg, A. A. and Berger, D. K. (2011) Maize microarray annotation database. Plant Methods **7**, 31.

Poster presentations:

2013 Joint Conference of the South African Genetics Society (SAGS) and South African Society for Bioinformatics and Computational Biology (SASBCB) - Stellenbosch, Western Cape, South Africa. **Title**: Expression QTL data analysis pipeline

2012 52^{nd} Maize Genetics Conference - Portland, Oregon, USA. **Title** (co-author): Re-annotation of the Agilent maize microarray based on the B73 genome sequence

2012 South African Association of Botanists (SAAB) - University of Pretoria, Pretoria, South Africa. **Title**: Maize Microarray Annotation Database

2011 International Society for Computational Biology (ISCB) Africa African Society for Bioinformatics and Computational Biology (ASBCB) Conference on Bioinformatics - Cape Town International Convention Centre, Cape Town, South Africa. **Title**: Maize Microarray Annotation Database

Contents

List of Figures

List of Tables

List of Abbreviations

ABA	Abscisic Acid
ABC	Adenosine Triphosphate (ATP)-Binding Cassette
AFLP	Amplified Fragment Length Polymorphism
ATP	Adenosine Triphosphate
Avr	Avirulence
BLAST	Basic Local Alignment Search Tool
bp	Base pair
cDNA	Complementary Deoxyribonucleic Acid
CIM	Composite Interval Mapping
cM	Centimorgan
DH	Doubled Haploid
DAP	Days After Planting
EST	Expressed Sequence Tag
ET	Ethylene
ETI	Effector-Triggered Immunity
ETS	Effector-Triggered Susceptibility
eQTL	Expression Quantitative Trait Locus
e-trait	Expression Trait

FDR	False Discovery Rate
FGS	Filtered Gene Set
FTP	File Transfer Protocol
GA	Gibberellic Acid
gDNA	Genomic Deoxyribonucleic Acid
GLS	Grey Leaf Spot
GO	Gene Ontology
GS	Gene Significance
GST	Glutathione S-Transferase
GWA	Genome Wide Association
HR	Hypersensitive Response
HS	Hotspot
IBM	Intermated B73 Mo17
IM	Interval Mapping
JA	Jasmonic Acid
KS	Kolmogorov–Smirnov
LD	Linkage Disequilibrium
LOD	Likelihood of Odds
LR	Likelihood Ratio
MAPK	Mitogen-Activated Protein Kinase
Mb	Megabase
ME	Module Eigengene
MM	Module Membership

mRNA	Messenger Ribonucleic Acid
NA	Not Available
NAM	Nested Association Mapping
NAT	Natural Antisense Transcripts
NBS-LRR	Nucleotide-Binding Site plus Leucine Rich Repeat
PAMP	Pathogen-Associated Molecular Pattern
PCD	Programmed Cell Death
PCR	Polymerase Chain Reaction
PR	Pathogenesis-Related
PRR	Pattern Recognition Receptor
PTI	Pathogen-Associated Molecular Pattern-Triggered Immunity
QTL	Quantitative Trait Locus
R protein	Resistance Protein
RFLP	Restriction Fragment Length Polymorphism
RIL	Recombinant Inbred Line
RLK	Receptor-Like Kinase
RNA-seq	Ribonucleic Acid-based sequencing
ROS	Reactive Oxygen Species
RT-qPCR	Reverse Transcription Quantitative Polymerase Chain Reaction
SA	Salicylic Acid
SFP	Single Feature Polymorphism
SMA	Single Marker Analysis
SNP	Single-Nucleotide Polymorphism

SNR	Signal to Noise Ratio
SSR	Simple Sequence Repeat
STK	Serine Threonine-Protein Kinase
TFBS	Transcription Factor Binding Sites
TDM	Transcript Derived Marker
TOM	Topological Overlap Matrix
UTR	Untranslated Region
WGCNA	Weighted Gene Co-expression Network Analysis
WGS	Working Gene Set

Chapter 1

Literature review: Exploiting gene expression data to dissect quantitative traits in plants

1.1 Introduction

The official 2013 world population is estimated at 7.1 billion by the United States Census Bureau (USCB). Approximately 800 million people go hungry per day, underlining the importance of food security. It is estimated that in the next 50 years, the world needs to produce more food than that was produced during the past 10,000 years. The current global food production trend is thus challenged and the need arises for more advanced methods in crop protection and food production.

Various research fields are addressed to increase crop protection and food production. These include conventional plant breeding, plant nutrition, horticulture, entomology and plant pathology. Genetic research has become the backbone to understand the underlying mechanisms that influence traits in the above-mentioned fields. Once the genetic basis of a specific phenotypic trait is determined, this could be utilised in various ways, such as crop improvement through marker-assisted breeding, genetic modification or targeted pathway expression in plants.

A major challenge in current biological research is to understand the molecular basis of quantitative traits. This literature review starts with an overview of quantitative trait locus (QTL) mapping and cloning, where the long-term goal is to identify genes

and specifically polymorphisms responsible for phenotypic variation. It then focuses on expression QTL (eQTL) mapping, which aims to dissect the molecular basis of gene expression variation. Associating eQTLs with phenotypic QTL, is hoped to provide a better understanding of the molecular basis of phenotypic traits. A few recent global eQTL studies in plants are compared to reveal common features and limitations of current technologies. Finally incorporating gene co-expression networks in a systems genetics context promises to elucidate the complex molecular networks underlying phenotypic traits.

1.2 QTL mapping

Genetic variation is important to the process of natural selection. Since gene alleles determine distinct traits that can be passed on from parents to offspring, favorable traits are passed on to the population as a whole through natural selection. Genetic variation can be brought about through polyploidy (changes in number of chromosomes), gene or point mutations, recombination, changes in chromosome structure or transposition (mobile genetic elements).

A molecular marker is a heritable and measurable DNA mutation, which may or may not have an effect on phenotype. Examples of molecular variation that are typically used as markers are variation at single nucleotides (single-nucleotide polymorphisms, or SNPs), short di-, tri-, or tetra-nucleotide tandem repeats (microsatellites), longer tandem repeats (minisatellites), small insertions/deletions, and insertion sites of transposable elements (Mackay, 2001). In progeny from a segregating population, markers on different chromosomes are inherited independently. However due to recombination, the closer markers are located on the same chromosome, the less likely they are to be separated in the progeny and are therefore said to be genetically linked.

Quantitative traits refer to a phenotype that shows a quantitative distribution in trait values, for example crop yield or disease resistance. To gain an understanding of the genetic basis of quantitative traits, QTL mapping was introduced in the 1980s (Lander and Botstein, 1989). QTLs are loci on the genome with physical boundaries defined by linked molecular markers that quantitatively affect the phenotype of interest. QTLs are often found on different chromosomes and the number of QTLs that explain the variation

of a phenotypic trait indicates the complexity of the genetic architecture of the trait. A trait can for example be controlled by many genes of small effect, or by a few genes of large effect. Whether these loci interact is also of interest.

QTL mapping approaches can be classified into linkage-based methods and association-based methods (Figure 1.1). Linkage-based QTL mapping uses related individuals and aim to identify segregating markers that predict the phenotype. Predictive markers and causal loci are linked and tend to segregate together, if not disrupted by recombination. Association mapping uses natural populations or unrelated individuals with historical recombination. Thus only tightly linked markers will predict the phenotype, and the causal locus is mapped with more accuracy. Both approaches require a mapping population to provide the genotypic and phenotypic variation. Marker genotypes and organismal phenotypes need to be scored in both cases. For a specific marker, individuals in a population are partitioned into groups based on genotype. A significant difference between trait phenotypic means of the genotype groups indicate the markers that are most likely linked to the trait. Genes located within the boundaries of the identified QTL, are candidate causal genes that affect phenotypic variation.

1.2.1 Linkage-based QTL mapping

Linkage-based QTL mapping, also called bi-parental QTL mapping, requires a controlled segregating population of related individuals, phenotypic data describing the trait of interest and a molecular marker based linkage map. It is often referred to as primary or coarse QTL mapping, since it only allows for an approximate mapping of QTLs to genomic regions, each typically containing up to hundreds of candidate genes. The actual DNA sequences, coding or non-coding, responsible for QTLs can only be detected by subsequent cloning of QTLs (see the section on QTL cloning on page 9).

When the aim of a QTL study is to find the underlying genetic causes of trait variation, large phenotypic differences between the parental lines of a population is valuable. The main types of segregating populations for self-pollinating species include (i) F_2 populations, which are produced by selfing the F_1 individuals in segregating populations generated by crossing the selected parents, (ii) backcross populations, which are generated by crossing the F_1 with either of the parents, (iii) doubled haploids (DHs), which are formed when haploid plants from F_1 (after wide crossing, chromosome elimination pro-

duces haploid embryos which are rescued and cultured) undergo chromosome doubling, and (iv) recombinant inbred lines (RILs), which are developed through single seed descent of an F_2 population, repeated for several generations until complete homozygosity is achieved. RILs and DHs are immortal populations that can be replicated over locations and years. Each line in a RIL population contains a random mixture of genotypes from the original parents (Collard *et al.*, 2005).

A genetic linkage map consisting of a set of molecular marker loci that are evenly spaced and span the genome with average intermarker distances of 5 to 10 centimorgan (cM) is optimal. A map function, either Haldane or Kosambi, is used to translate from recombination frequency to distance or vice visa. Computer software for genetic linkage mapping such as MapMaker (Lander *et al.*, 1987) or JoinMap (van Ooijen, 2006) can be used to generate a cM-scale map per linkage group.

The most popular methods for QTL detection are single marker analysis (SMA), interval mapping (IM), composite interval mapping (CIM), multiple interval mapping (MIM) and Bayesian methods. SMA does not require a linkage map and statistical tests for identifying markers that significantly correlate with trait of interest include t-tests, analysis of variance (ANOVA) or linear regression. The null hypothesis tested at each molecular marker is that the mean trait phenotype value does not differ between genotypic classes. IM uses a linkage map to examine intervals between adjacent pairs of linked markers simultaneously. This method compensates for recombination between the markers and the QTL (Lander and Botstein, 1989). CIM combines interval mapping with linear regression and includes additional genetic markers in the statistical model, in addition to an adjacent pair of linked markers for interval mapping. CIM is more precise and effective at mapping QTLs, particularly when linked QTLs are involved (Zeng, 1993). MIM aims to analyse multiple QTL together with epistasis, through a model selection procedure to search for the best genetic model for the specific quantitative trait (Kao *et al.*, 1999). Bayesian QTL mapping methods take advantage of the uncertainties in QTL number, location, and effects, by studying their joint distributions (Zou and Zeng, 2008). Current QTL mapping methods are adapting and improving to take challenges such as QTL by environment/trait interactions into account. van Eeuwijk *et al.* (2010) gives an overview of advanced mixed model and Bayesian QTL approaches that are appropriate for many types of breeding populations. These approaches also contain features for the detection

and predictive modeling of the genetic basis of genotype-environment interactions.

A variety of software programs for QTL mapping are available. Some of the most popular programs are QTL Cartographer, MapMaker/QTL, R/QTL, MapQTL, Qgene and SAS. Interval mapping methods calculate a likelihood of odds (LOD) score or likelihood ratio (LR) at each interval, which indicates the probability of detecting a QTL at that position. Finally, a profile of scores are plotted along the genome for each trait, and whenever the peak exceeds a specified significance level, the presence of a QTL is suggested (Figure 1.1 (e)). Permutation tests are commonly used to determine the empirical threshold for significance.

QTLs are commonly used for marker-assisted selection (selecting plants on the basis of their marker genotypes), understanding trait architecture, providing insights into genetic relationships among traits and for identifying chromosome regions for isolating and cloning genes (Collard *et al.*, 2005). QTL mapping in crop plants has now become routine due to the progress made in this area during the last two decades. In rice and maize respectively, more than 8,500 and 1,700 QTL have been detected (http://www.gramene.org/qtl).

1.2.2 Association mapping

Association mapping is the study of statistical associations, based on linkage disequilibrium (LD), between genetic markers and phenotypic traits in natural populations (see the section on LD on the next page). Association mapping can be split into two broad categories namely candidate gene association mapping and genome-wide association mapping (GWA) studies (Risch and Merikangas, 1996). Candidate-gene association mapping requires candidate genes to be selected based on prior knowledge, for example from linkage-based QTL mapping. An independent set of random markers is required to infer genetic relationships. It is a low-cost, hypothesis driven, and trait-specific approach. GWA studies on the other hand, is a comprehensive approach to systematically search the genome for causal genetic variation. A large number of markers are tested for association with various complex traits (Zhu *et al.*, 2008*a*). The number of markers required in an association mapping study depends on the scale and pattern of LD (Nordborg and Tavaré, 2002).

The basic statistical test for association analysis in an ideal situation would be lin-

ear regression, ANOVA, t-test or chi-square test (Zhu *et al.*, 2008*a*). However, since population structure is a confounding factor in association mapping, different statistical approaches have been designed to incorporate this, in order to avoid spurious genotype–phenotype associations. TASSEL (Trait Analysis by aSSociation Evolution and Linkage) is a popular software package used for association mapping in plants, and is continually updated as new methods are developed (Bradbury *et al.*, 2007). EMMA (Efficient Mixed-Model Association) corrects for population structure and genetic relatedness, and is often used in model organism association mapping studies (Kang *et al.*, 2008*b*). In order to detect genetic effects of moderate size with association mapping, large numbers of divergent lines are required.

In plant studies to date, most of the successful GWA experiments have uncovered loci previously known to affect a trait, or when traits were not extensively evaluated before, associated regions were identified. GWA studies could be useful when inbred lines are available because they can be grown in replicate under controlled conditions, and multiple phenotypes can be studied while controlling environmental noise. Also, once these lines are genotyped, they can be repeatedly phenotyped. However, since the mapping population is heavily structured, elevated false-positive rates are expected. Atwell *et al.* (2010) performed a GWA study of 107 phenotypes in a common set of *Arabidopsis thaliana* inbred lines. Many common alleles with major effect were identified. However, the results were confounded by complex genetics and population structure, which made it difficult to distinguish true from false associations. Previously identified candidate loci were significantly overrepresented among the identified associations. These were at least confirmed to be good candidates for follow-up experiments.

Linkage disequilibrium

LD is the non-random association of alleles at different loci. It is the occurrence of some combinations of alleles in a population more often or less often than would be expected from a random formation of haplotypes, based on their frequencies. The difference between observed and expected haplotype frequencies in a population at two loci, is considered the deviation D.

To identify SNPs or haplotypes significantly associated with phenotypic trait variation, the correlation between a pair of loci, called r^2 is the most relevant LD measurement.

A r^2 value of 0 indicate that loci are in complete linkage equilibrium and a r^2 value of 1 indicate that loci are in complete linkage disequilibrium. The relationship of r^2 relative to either genetic or physical distance between measured loci, gives an indication of, on average how fast LD decays across the genome, and how many markers are needed for an association mapping experiment. For example, if LD decays within a short distance, mapping resolution is expected to be high, but a large number of markers are required. Generally, LD extends to a much longer distance in self-pollinated crops, such as wheat and *Arabidopsis*, than in cross-pollinated species, such as maize. A r^2 cut-off value of 0.1 or 0.2 is typically used to describe the LD decay (Zhu *et al.*, 2008*a*). Factors that can strongly influence patterns of LD, include recombination, genetic drift, inbreeding, mutation and gene flow.

It is important to distinguish between LD and linkage. LD refers to the correlation between polymorphisms that is caused by their shared history of mutation and recombination, and linkage to the correlated inheritance of loci through the physical connection on a chromosome (Flint-Garcia *et al.*, 2003). Tight linkage may result in high levels of LD.

Compared to animals, LD has not been studied extensively in plants. A few studies investigated LD in maize and *Arabidopsis* across various population and marker types. In maize, the patterns of LD (measured as r^2) vary substantially. Tenaillon *et al.* (2001) used a diverse set of maize germplasm to examine sequence diversity at 21 loci on chromosome 1. They showed that interlocus LD decreased to less than 0.25 within 200 bp on average. Remington *et al.* (2001) investigated six candidate genes in a diverse set of 102 inbred lines. They reported that intragenic LD decayed to less than 0.1 within 1,500 bp. Conversely, Rafalski (2002) showed that in elite maize populations, LD extends to greater than 100 kb for the *adh1* and *y1* loci. Lu *et al.* (2011) genotyped 287 tropical and 160 temperate inbred lines with 1,943 SNP markers. They found that the LD decay distances across the genome, chromosomal regions and germplasm groups varied significantly. Specifically, the LD decay distance was two to ten times larger in the temperate germplasm (10–100 kb) compared to the tropical germplasm (5–10 kb). Flint-Garcia *et al.* (2003) speculates that the different rates of LD decay in maize mainly reflect differing levels of population bottleneck, for example the progression from diverse landraces to diverse inbreds to elite inbreds. In contrast, LD in *Arabidopsis*, a highly selfing species,

is more constant and extends further in general. In a study of 163 genome-wide SNPs in 76 *Arabidopsis* accessions showed that LD decayed within approximately 250 kb (Nordborg *et al.*, 2002). Hagenblad and Nordborg (2002) reported the same result after they sequenced 14 short fragments from a 400 kb region of the flowering time locus FRIGIDA. Later, Kim *et al.* (2007) studied genome wide LD in 19 *Arabidopsis* accessions using 341,602 non-singleton SNPs. They established that LD decays within 10 kb on average, which is significantly faster than previously estimated. This result is currently used as the standard.

1.2.3 Advantages and disadvantages of the two approaches

The main drawbacks of bi-parental QTL mapping are that a QTL mapping population (i) usually has a small number of recombination events per chromosome, which results in limited mapping resolution, (ii) is initiated by only two parents, which results in a limited allele number, and (iii) is time-consuming to generate (or not possible to generate), which results in increased research time or restricted utility. Association mapping overcomes these drawbacks, mainly because it is based on natural populations with recombination events that occurred throughout the entire evolutionary history of the mapping population. Thus it also searches for functional variation in a much broader germplasm context (Ingvarsson and Street, 2011). The major goal of association mapping is to discover the causative SNP itself, in contrast to bi-parental mapping where additional steps are required to narrow down the QTL region. Two drawbacks of association mapping are that larger populations are required and population structure becomes important.

Association mapping is useful for discovering common variants with an effect throughout the species, thus discovering associations of broad application. Such a locus may account for only a small amount of the variation, but enough of the alleles are present to affect the mean of a specific genotypic class. Bi-parental mapping on the other hand is useful for discovering rare alleles that control a phenotype, since the population has many copies of the rare allele, which typically has a major effect. The goal of an experiment will thus determine the approach. Association mapping of an appropriate population is the best approach to dissect the more general factors controlling a phenotype. Whereas the best approach to discover, analyse, and test genes of major effect is to use bi-parental populations of divergent parents and CIM (Flint-Garcia *et al.*, 2003).

With the rapid decrease in sequencing and genotyping costs, GWA studies in plants are expected to increase. It also proves to be an excellent complement to bi-parental QTL mapping. In several plant species, diverse germplasm panels are being established for whole-genome association analysis (Zhu *et al.*, 2008*a*). For example, the maize nested association mapping (NAM) population, is a collection of 5,000 RILs made by crossing 25 diverse bi-parental inbred lines with the reference line B73 (www.panzea.org) (Yu *et al.*, 2008). NAM captures the advantages of both linkage mapping and association mapping. Joint linkage mapping takes advantage of the shared B73 line in all 25 subfamilies to identify QTLs for specific traits, at an improved resolution (Li *et al.*, 2011), and GWA uses deep genotyping results from the 25 founder lines in a projection onto the progeny for improved resolution. Kump *et al.* (2011) evaluated the NAM population for resistance to southern leaf blight (SLB) disease in the maize. Thirty-two QTLs, with mostly small, additive effects on SLB resistance were identified using joint-linkage analysis. GWA tests of maize HapMap SNPs revealed SLB resistance associated SNPs within and outside of QTL intervals, many of which were within genes that were previously shown to be involved in plant disease resistance. Tian *et al.* (2011) used a GWA study of the maize NAM panel to investigate the genetic basis of important leaf architecture traits. They deduced that such traits are dominated by small effects with little epistasis, environmental interaction or pleiotropy. Specifically, they identified that variations at the *liguleless* genes contribute to more upright leaves.

1.3 QTL cloning

QTL cloning is a major objective of QTL mapping. It is the identification of the DNA sequences responsible for QTLs and thus for variation in the trait of interest. These causal DNA sequences represent polymorphisms between the parental lines of the segregating population, and can be in coding regions (within genes) or in non-coding regions such as promoters (regions responsible for regulation of gene expression) or regulatory RNAs (non-coding RNAs responsible for post-transcriptional regulation of gene expression). Depending on the number of recombination events and thus the number of individuals in the segregating population, QTLs span large genomic regions including hundreds of genes and significant effort is required for the identification of the causal polymorphism.

1.3.1 Positional cloning

For plant QTLs, positional cloning is currently the most successful QTL cloning approach. It involves QTL Mendelisation, which is best accomplished by the construction of a new experimental population of near-isogenic lines (NILs). NILs differ only at the alleles of the target QTL segment (Alonso-Blanco and Koornneef, 2000). After recruitment of polymorphic markers for that region, QTL fine mapping allows a more precise estimate of the cM region spanned by the QTL. Subsequent physical mapping of this interval on the DNA sequence will reveal the candidate genes to be selected for evaluation (Salvi and Tuberosa, 2005).

Functional testing of candidate genes can be carried out by over-expressing or down-regulating the target gene through genetic engineering or RNA interference (RNAi) (Waterhouse and Helliwell, 2003), genetic complementation of a known mutant (Doebley *et al.*, 1997), reverse genetics tools such as transferred DNA (T-DNA) or transposon-tagged populations (Maes *et al.*, 1999), TILLING (Targeting Induced Local Lesions IN Genomes) (Mccallum *et al.*, 2000) or gene replacement (Iida and Terada, 2004). However, validation of the causative polymorphism can be a major challenge, especially when the polymorphism responsible for the QTL effect reside in non-coding regions, for example in regulatory regions (promoters, enhancers or silencers), microRNA loci, transposon insertions or at regions controlling chromatin methylation or organisation (Salvi and Tuberosa, 2005).

Wang *et al.* (2005) used positional cloning in maize to identify the molecular basis of a *teosinte glume architecture* QTL, called *tga1*, of large effect. They investigated the fact that teosinte kernels are tightly encased in structures called cupulate fruitcases, whereas maize kernels are borne uncovered on the surface of the ear, so that humans can easily use it as a food source. Wang *et al.* (2005) developed a set of molecular markers near *tga1* and screened 3,106 F_2 plants segregating for *tga1*. The locus was eventually mapped to a 1 kb segment with homology to SBP (squamosa-promoter binding protein) transcriptional regulators. DNA sequence analysis of the SPB gene in maize and teosinte showed seven fixed DNA differences. One of these differences encodes a non-conservative amino acid substitution that may affect protein function, and the other six differences potentially affect gene regulation. Mutant recovery was performed as a functional proof that this gene can be used to distinguish the above-mentioned maize and teosinte phenotypes,

however exactly how *tga1* regulates ear development remains to be determined.

1.3.2 Association mapping with candidate genes

Association mapping with candidate genes (see section on page 5) is a QTL cloning method that does not require detailed linkage information. It specifically targets genes with known functions in the trait of interest. In apple for example, a key gene belonging to the *polygalacturonase* gene family called Md-PG1, was identified as a candidate gene because it co-localised with the statistical interval of a major hotspot QTL associated to several fruit texture sub-phenotypes. To investigate a region of approximately 16 kb containing the *Md-PG1* gene, Longhi *et al.* (2013) used a candidate gene based association mapping approach. A collection of 77 apple cultivars was analysed using 40 markers, and an average LD extent of 2 kb was defined for this region. This rapid LD decay confirmed the suitability of the candidate gene based approach. *Md-PG1* was validated as the main locus responsible for a QTL impacting fruit texture in apple, and new functional alleles associated to the fruit texture properties were discovered.

1.3.3 QTL tagging

Soller and Beckmann (1987) described the theoretical potential of insertional mutagenesis or QTL tagging as a means of cloning QTL. It entails mutagenesis as a result of integration of novel DNA sequences into the germ line. This process requires the phenotypic screening of an insertionally mutagenised population for the target trait, in order to identify lines with an altered phenotype compared to what is expected. However, since a very large number of potential insertion sites exist in the genome, but only a limited number of target sites that can affect any particular trait, a complete screening experiment would involve up to 20,000 plants. Also, effects of allelic variants (at any single QTL) on phenotype value are expected to be small, thus a few stages of replicate testing are required per insert for accurate conclusions. The functionally modified or inactivated gene could be rescued using standard molecular procedures. In plants, QTL tagging could be based on transfer DNA (T-DNA), RNA interference (RNAi), DNA-transposons or retrotransposons (Salvi and Tuberosa, 2005).

Singh *et al.* (2012) investigated 2 major barley QTLs located on chromosome 4H, affecting malting quality traits, using mutagenesis via the Activator (*Ac*)/Dissociation (*Ds*)

transposon tagging system. *Ds* transposon lines were created with stable transformation methods and transient *Ac* expression was used to transpose *Ds* elements. Since the *Ds* transposons tend to re-insert into genic regions that are close to the site of excision, lines with Ds loci in the region of the two above-mentioned QTLs were used as a launch pad for its reactivation. Reactivation was accomplished through hybridisation with the *Ac*TPase-expressing line and via the transient expression of *Ac*TPase in immature embryos using *Agrobacterium tumefaciens*. BLAST was used to analyse the *Ds* flanking sequences from a subset of new lines, and most of these were predicted to affect malting quality traits. A few important genes that were tagged by *Ds* include the *β-GAL1*, the *β-amylase-like* gene and the ATP-binding cassette (ABC) transporter. The resulting new source population provides new transposon mutants for functional genomic studies.

1.3.4 Functional genomics

Functional genomics can benefit QTL cloning by reducing the number of candidate genes in a QTL interval. Apart from identifying genes that are functionally related to the trait of interest and physically located in the QTL interval, transcriptional profiling can be an important indicator of candidates, since gene transcription is a primary intermediate between the information encoded in the genome and the final phenotype (Cubillos *et al.*, 2012). Specifically, gene expression profiling of lines with contrasting QTL genotypes, for example the parental lines, can be used to generate lists of differentially expressed genes. An even better approach would be to exploit the genetic variation in the entire segregating population by identifying expression QTLs (eQTL), loci controlling the level of gene expression, for each gene in the genome. Candidate genes may be revealed by an agreement analysis between loci controlling expression variation and loci controlling phenotypic variation. This will be discussed in the next section.

1.4 Genetical genomics

1.4.1 Expression QTL mapping

The concept of "genetical genomics" or eQTL mapping, was introduced by Jansen and Nap (2001). Genetic and gene expression approaches have been brought together to study the

genetic basis of gene expression. eQTL mapping involves treating the level of expression of single genes individually as quantitative traits. Similar to QTL mapping, statistical tests are performed between the markers and each gene expression trait respectively, which makes it possible to dissect the genetic loci explaining gene expression variation.

The molecular basis of an eQTL is a DNA polymorphism that gives rise to differential gene expression. Since gene regulation operates to produce differential amounts of messenger RNAs (mRNAs) and in turn proteins, polymorphisms affecting the various stages of gene regulation, including transcription but also RNA splicing, RNA transport, RNA stability and translation into protein, are good candidate causal polymorphisms (Latchman, 2005). In summary, genes could be differentially expressed in genotypically diverse individuals due to (i) *cis*-elemental variation in promoter sequences affecting transcription initiation, (ii) polymorphisms in the intronic regions effecting splicing, or (iii) changes in untranslated regions (UTRs) affecting mRNA stability and potentially differential RNA degradation (Holloway and Li, 2010). Furthermore, (iv) polymorphisms in the coding regions of genes (such as transcription factors) creating dysfunctional or hyperactive proteins, (v) copy number variation or (vi) genomic rearrangements (such as translocations, insertions and deletions) can also cause eQTLs.

All large eQTL studies distinguish between two groups of eQTLs namely *cis* (local) and *trans* (distant), based on the distance between the eQTL and the gene encoding the transcript being measured. A *cis*-eQTL represents a polymorphism located within the region of the target gene. A typical example is a promoter polymorphism, causing differential expression of the target gene between groups of individuals with different promoter alleles (Figure 1.2). This can happen for example if the polymorphism in one group prevents effective binding of a regulatory gene, and thus decreases transcription of that gene. In principle, a *cis*-eQTL could affect transcription initiation, the rate of transcription, or transcript stability in an allele-specific manner (Joosen *et al.*, 2009).

Conversely, a *trans*-eQTL represents a polymorphism at a different location than the position of the gene encoding the transcript being measured (Figure 1.2). For example, a *trans*-eQTL could represent the location of a regulator that controls the expression of the target gene, where this gene is potentially located on a different chromosome. A the *trans*-eQTL could be a polymorphism (i) located in the coding region of the regulator, not necessarily giving rise to a *cis*-eQTL, or (ii) in a motif integral for transcription factor

binding, causing a *cis*-eQTL at the regulator. Nevertheless, this potentially sets up a network where *cis* variation in regulatory factors, effect the expression of downstream target genes in *trans* (Hansen *et al.*, 2008).

Cis-eQTLs were found to have an additive influence on gene expression levels, meaning that the expression level in a hybrid will normally be intermediate between those of the two parents. This is thought to contribute to positive selection on *cis*-regulatory elements over long evolutionary time (Lemos *et al.*, 2008). *Trans*-eQTLs on the other hand show a tendency towards dominant regulation of their target genes, meaning that the expression level of the hybrid is similar to the one of the parental lines and significantly different from the other. Zhang *et al.* (2011) found that dominant *trans*-eQTLs are more likely to regulate multiple expression traits, and could be regulatory hotspots (see the next section).

1.4.2 *Trans*-eQTL hotspots

Numerous *trans*-eQTLs often cluster in hotspots, causing the genome-wide distribution of *trans*-eQTLs to significantly differ from that of gene density. A genomic locus is called a hotspot when more eQTLs map to the locus than expected by chance. A biologically meaningful *trans*-eQTL hotspot is assumed to contain a master transcriptional regulator controlling the expression of a group of genes that act in the same biological process or pathway (Druka *et al.*, 2010). To predict the biological relevance of hotspots and sometimes to predict the causal regulator, additional sources of information are often used. Such information sources include gene ontology (GO), gene co-expression, transcription factor binding sites (TFBS), transcription factor targets, ChIP-Seq and protein-protein interactions (Zhu *et al.*, 2008*b*). *Trans*-eQTL hotspots often have a directional bias, meaning that the same parental allele increases expression for most of the transcripts associated with a hotspot.

Trans-eQTL hotspots can contain up to thousands of genes. In such hotspots, gene expression can be affected by pleiotropic regulators in the hotspot locus, affecting many often unrelated phenotypes. When a mutation like this affect the expression of many genes in *trans*, large-effect *trans*-eQTL might be deleterious. In general *trans*-eQTLs are expected to have smaller effects on single transcripts than *cis*-eQTLs, possibly due to indirect regulatory mechanisms (Hansen *et al.*, 2008). An an example, in a genome-wide

eQTL analysis of an *Arabidopsis* RIL population derived from the parental accessions Landsberg *erecta* (Ler) and Cape Verde Islands (Cvi), Keurentjes *et al.* (2007) identified several regulatory *trans*-eQTL hotspots. The ERECTA locus was the most prominent hotspot, consisting of 176 *trans*-eQTLs. In Ler, this locus contains the mutated ERECTA gene (At2g26330) with a strong pleiotropic effect on the Ler phenotype. However, in Cvi a functional copy of the gene is present. The ERECTA protein is a membrane-bound leucine-rich repeat receptor-like serine (Ser)/threonine (Thr) kinase (LRR-RLK) (Torii *et al.*, 1996), known to regulate developmental processes, hormone signalling, and defense. Terpstra *et al.* (2010) combined monogenic mutant analysis with eQTL mapping in the Ler×Cvi RIL population to analyse the effect of ERECTA on down-stream gene expression. They linked mitogen-activated protein kinase (MAPK) signalling components and downstream WRKY transcription factors between ERECTA and the differentially expressed target genes.

It is thus important to realise master regulators do not necessarily have to be transcription factors. They could just as well be indirect transcript level regulators for genes within the same pathway (Brem *et al.*, 2002). Furthermore, identified master regulators might alternatively point toward ubiquitous *trans*-regulators that either controls the degradation process of unrelated transcripts or other processes affecting transcription in general (Holloway and Li, 2010).

To date only a few reported hotspots have been verified. Since *trans*-eQTLs generally show smaller effect sizes than *cis*-eQTLs, they are identified at a lower statistical power. It thus difficult to reliably detect *trans*-eQTL hotspots, which are as a result less consistent between studies than *cis*-eQTLs (Breitling *et al.*, 2008). It should be noted that not all *trans*-eQTL hotspots are master regulators. In order to truly define a regulator in quantitative genetics, more work on network-specific *trans*-eQTLs, the existence of potential feedback regulatory mechanisms, and the linking of *cis*-eQTLs via *trans*-regulatory connections (epistatic interactions between the loci), is needed (Kliebenstein, 2009). Breitling *et al.* (2008) speculates that the scarcity of plausible hotspots might indicate that most of heritable gene expression variation is effectively "buffered". Because it could result in systems failure, gene expression variation does not necessarily lead to downstream effects on other genes.

Instead of explaining *trans*-eQTL hotspots by master regulators, hotspots can also be

due to clusters of genes with highly correlated expression (Wang *et al.*, 2007; Breitling *et al.*, 2008). This is particularly possible if the false discovery rate (FDR) for individual eQTLs is very high. The joint response to an uncontrolled environmental factor is an example of a non-genetic mechanism that can produce strongly correlated clusters of functionally related genes.

1.4.3 Associating eQTLs with phenotypic QTLs

eQTLs can be used to search for associations between gene expression polymorphisms and phenotypic QTLs. Identification of all large-effect *cis*-eQTLs that underlie a phenotypic QTL, will provide a list of candidate genes that can be tested for causal linkage with the phenotype of interest (Kliebenstein, 2009). The number of such candidates in a QTL region will be influenced by the resolution of the genetic map.

Many of the QTL cloned in plants, before genome-wide eQTL analysis, were actually *cis*-eQTL. Salvi *et al.* (2007), through positional cloning and association mapping, identified the molecular basis of the major flowering-time QTL, *Vegetative to generative transition 1* (*Vgt1*), in maize. The QTL was narrowed down to a 2 kb non-coding region positioned 70 kb upstream of an AP2-like transcription factor, that have been shown to be involved in flowering-time control. *Vgt1* functions as a *cis*-acting regulatory element, which would show up as a *cis*-eQTL in a genetical genomics study, since expression levels of the downstream gene was significantly different in lines carrying the two different alleles.

Correlation analysis between gene expression profiles and the phenotype values across the individuals of a population, can be used to identify genes with high absolute correlation coefficients. A gene with a good correlation and with an eQTL overlapping the phenotypic QTL, is a strong candidate for causing the trait. As an indirect example, positional cloning was previously used to identify Rpg1, a gene that confers resistance in barley to the wheat stem rust pathogen *Puccinia graminis* f. sp. *Tritici* (*Pgt*). Druka *et al.* (2008) re-analysed the quantitative resistance phenotype data using 139 DH lines of the Steptoe×Morex reference barley mapping population. One of six identified QTL loci coincided with the major stem rust resistance locus *Rpg1*. Correlation analysis between phenotype values for rust infection and genome-wide gene expression values revealed that Rpg1 was in the top five best correlating candidate genes. When candidate genes are in-

ferred based on correlation, directionality of the observed associations has to be taken into account such that gene expression data are interpreted in the context of the underlying trait biology (Druka *et al.*, 2010). In the *Rpg1* example above, Druka *et al.* (2008) selected genes with positively correlated profiles as candidates. This however is based on the assumption that increased resistance should positively correlate with the amount of mRNA from the resistance gene.

Additional analyses could involve investigating the functions of correlated genes overlapping the phenotypic QTL, and also assessing whether the QTL have multiple coinciding *trans*-eQTLs. This may predict that the causal gene could be a *trans*-acting master regulator, especially if the coinciding eQTLs are mainly *trans* acting and functionally related. Such a master regulatory gene may or may not be represented on the array, but should be detected when RNA-based sequencing (RNA-seq) is used for gene expression profiling. The possibility that chance co-segregation is responsible for the correlation, rather common transcriptional regulation by a master regulator, is an important aspect that should be investigated with care (Druka *et al.*, 2010). The *sub1* locus in rice is an example of a true master regulatory locus. It controls the activity of an ethylene response factor with significant *trans* effects that confers submergence tolerance to rice (Fukao *et al.*, 2006; Xu *et al.*, 2006).

When eQTLs are used to filter candidate genes in a trait QTL interval, it is important to keep in mind that the polymorphism responsible for a phenotypic QTL will not necessarily change the expression level of a gene and thus will not be detected as an eQTL. For example, when the polymorphism is in the coding region of a gene, leading to variations in protein stability, enzymatic activity or post-translational modification, or when the polymorphism in the methylation level of the DNA, differential expression will not be evident. Also, when post-translational modifications are the predominant regulatory mechanisms for variation in a trait, using eQTLs to identify candidate genes will not be useful (Hansen *et al.*, 2008).

Harjes *et al.* (2008) and Balint-Kurti *et al.* (2010) are examples of two studies where maize microarrays were used to identify genes with *cis*-eQTLs as candidates for phenotypic QTL. In both studies, a subset of lines from the intermated B73×Mo17 (IBM) RIL population was used to perform gene expression profiling on 53K spotted-oligo microarrays. To help breeders produce maize grain with higher provitamin A levels, Harjes

et al. (2008) sudied the variation at the *lycopene epsilon cyclase* (*lcyE*) locus, known to alter flux down α-carotene versus β-carotene branches of the carotenoid pathway. Association analysis, linkage mapping, expression analysis, and transposon tagging was used to dissect the molecular basis of *lcyE*. Five polymorphisms were identified to be significantly associated with changes in flux between the lutein and zeaxanthin branches of the pathway. Using the eQTL data mentioned above, Harjes *et al.* (2008) concluded that besides for *cis*-regulation, several other regions (in *trans*) also contribute to expression level control of *lcyE*. Using low-cost markers, favorable *lcyE* alleles can now be selected by breeders to produce maize grain with higher provitamin A levels. Balint-Kurti *et al.* (2010) studied the connections between plant bacterial diversity and disease resistance in maize leaves. Six genomic loci were found to control epiphytic diversity. Interestingly, the identified loci significantly overlapped with loci controlling resistance to southern leaf blight (SLB). Using the eQTL data mentioned above, lists of candidate *cis*-eQTL genes were generated for each genome region identified as a QTL.

In the same way that mRNA abundance data is used across individuals in a population to map eQTLs, the parallel measurement of the abundance of thousands of proteins and metabolites similarly leads to the mapping of protein QTLs and metabolite QTLs. Such QTLs can also provide insight as to specific candidate genes controlling phenotypic variation. Keurentjes *et al.* (2008) used the *Arabidopsis* Ler×Cvi RIL population to perform a parallel genetic analysis between gene expression, enzyme activity and metabolite accumulation in primary carbohydrate metabolism. Correlation and QTL overlap analyses revealed connectivity between the three levels, as well as independent regulation at each level. An example of independent regulation at each level, is that post-transcriptional regulation of enzyme levels will not be detected when gene expression levels are measured. How transcript variation relates to other genomic or physiological levels is still largely unknown, and cross-phenotypic comparison studies need to take into account that the levels of heritability and epistasis underlying their genetic architecture may be very different (Kliebenstein, 2009). Furthermore, even though the combination of different types of genomic QTL could add value to studies like these, the technology required for proteomic and metabolomic approaches is currently not as advanced and accessible as for high-throughput whole genome transcriptional profiling (Delker and Quint, 2011).

1.5 Genome-wide eQTL mapping studies in plants

Large-scale global eQTL mapping studies on a variety of plants have been published over the past decade. These studies revealed that (i) the expression of large numbers of transcripts are genetically controlled, (ii) *cis*-eQTLs tend to have larger effects than *trans*-eQTLs, (iii) there tend to be more *trans*-acting than *cis*-acting polymorphisms, (iv) some genomic regions are called hotspots, which explain the expression variation of many transcripts and (v) sample sizes are usually much smaller than the number of expression traits (e-traits) tested (Mackay *et al.*, 2009). Tables 1.1 and 1.2 give an overview and a comparison of 16 genome-wide eQTL mapping studies. Crops and plants on which these studies were conducted are maize (Schadt *et al.*, 2003; Shi *et al.*, 2007; Swanson-Wagner *et al.*, 2009; Holloway *et al.*, 2011; Li *et al.*, 2013), eucalyptus (Kirst *et al.*, 2005), *Arabidopsis* (West *et al.*, 2007; Keurentjes *et al.*, 2007), wheat (Jordan *et al.*, 2007), barley (Potokina *et al.*, 2008; Chen *et al.*, 2010*b*; Moscou *et al.*, 2011), rice (Wang *et al.*, 2010*b*), cotton (Claverie *et al.*, 2012) and potato (Kloosterman *et al.*, 2012).

1.5.1 Expression traits, *cis*- vs *trans*-eQTLs and population size

To date whole genome microarrays for expression profiling were mostly used in global eQTL mapping studies, since already established protocols and data analysis methods can be used. RNA-seq as a large-scale gene expression platform will become more popular as analysis techniques advance and sequencing costs decrease. The average number of e-traits analysed per study in Tables 1.1 and 1.2, was 15,636 and an average of 74% of these e-traits were associated with at least one eQTL. This demonstrates that most transcripts seem to be under genetic control.

An annotated genome sequence is a valuable resource for accurate classification of eQTL as *cis* or *trans*, but this was not feasible for most crop species analysed (Tables 1.1 and 1.2). Various different ways of determining whether a gene is locally regulated were reported. In most cases, physical positions of the markers were used to anchor the genetic map to the best available physical map. West *et al.* (2007) classified an eQTL less than 3.5 cM from its target gene as *cis*-acting, and reported that using a 5 cM distance instead had minimal effect on the number of *cis*-eQTLs identified. Considering their current genetic resolution, Potokina *et al.* (2008) and Swanson-Wagner *et al.* (2009) considered

a 5 cM distance between an eQTL and the target gene as sufficiently close. Holloway *et al.* (2011) used a 10 cM distance, but mentioned that a 5 or 10 cM boundary would be appropriate for maize eQTL studies. Kloosterman *et al.* (2012) used the same linkage group as a criterion for identifying *cis*-eQTL. Keurentjes *et al.* (2008) and Wang *et al.* (2010*b*) calculated support intervals per eQTL and when the gene's position coincided with the support interval, classified it as *cis*-acting. Shi *et al.* (2007) and Drost *et al.* (2010) used the co-localisation of each eQTL peak with the genetic map marker bin containing the gene model. Lastly, Jordan *et al.* (2007) roughly compared gene locations that were determined from wheat–rice synteny, to corresponding eQTL locations in order to estimate whether eQTLs were *cis*- or *trans*-acting.

Cis-eQTLs generally explain a larger part of the expression level variation of their target genes than *trans*-eQTLs. Wang *et al.* (2010*b*) and Li *et al.* (2013) in their studies of 110 rice RILs and 105 maize RILs respectively, showed that 96% and 87% of the detected *trans*-eQTL explained $< 20\%$ of the variation, whereas 37% and 16% of the detected *cis*-eQTLs explained $< 20\%$ of the variation. Furthermore, West *et al.* (2007) when analysing 211 *Arabidopsis* RILs, found that 75% of the detected *trans*-eQTLs explained $< 10\%$ of the variation in e-traits. This indicates that *cis*-eQTLs tend to have larger effects than *trans*-eQTLs, and could be due to *cis* polymorphisms having a direct influence on the expression of a gene, contrary to *trans* polymorphisms. Also, a polymorphism in one of multiple *trans*-factors regulating a gene, is likely to result in only a small change in expression of the gene being regulated in *trans*.

Large differences between the *cis*- and *trans*-eQTL ratios are evident in genome-wide studies. The average ratio detected across the studies in Tables 1.1 and 1.2 was 40% *cis*-eQTL and 60% *trans*-eQTL. It remains to be tested whether the *cis*/*trans* ratio fluctuates due to differences in statistical power, or whether it is perhaps a true reflection of the genetic polymorphism present in the population (Kliebenstein, 2009). It is speculated that low-powered QTL mapping experiments will detect most of the larger-effect *cis*-eQTL. Thus increasing the number of individuals in the population (and degrees of replication), may lead to the detection of new smaller-effect *trans*-eQTLs. Figure 1.3 shows that the number of eQTLs differ considerably between studies. The correlation between population size and the number of eQTLs detected was 0.35 (p-value = 0.18). Even though this is not significant, the positive correlation may support the hypothesis

that low-powered QTL mapping experiments detect most of the larger-effect eQTL, and that increasing the number of individuals in the population may lead to the detection of new smaller-effect eQTLs.

It is important to note that increasing the number of individuals in a population will lead to a higher levels of recombination and thus to increased genetic resolution. This will reduce the size of detected eQTL regions. On the other hand, increasing the number of molecular markers might also reduce the size of detected eQTL regions since more statistical tests will be performed across the genome. However if the genetic resolution is not good enough, adding more markers will not add additional value.

Swanson-Wagner *et al.* (2009) were interested in the mechanisms responsible for heterosis, the superior performance of hybrid progeny relative to their inbred parents. They used hybrids between the maize inbred lines B73 and Mo17, which exhibited heterosis regardless of cross direction. Reciprocal hybrids were generated by crossing each B73×Mo17 RIL onto B73 (B73×RIL) and Mo17 (Mo17×RIL). Separate eQTL analyses were conducted within each cross type. More than 75% of the detected eQTL were *trans*-eQTLs and instead of showing additive gene action, 86% of these *trans*-eQTLs showed paternal dominance. Thus for the alleles at a specific marker, expression values in those heterozygotes with, say B73, as the paternal parental allele matched expression values in lines that were homozygous for the B73 allele. This result is consistent with imprinting, an epigenetic event by which certain genes can be expressed in a parent-specific fashion. Swanson-Wagner *et al.* (2009) hypothesise that at least some paternally dominant *trans*-eQTLs are small RNAs.

1.5.2 *Trans*-eQTL hotspots, enrichment analyses and networks

Twelve out of the 16 studies in Tables 1.1 and 1.2 used the frequency distribution of *trans*-eQTLs across the genome to identify hotspots. The majority of these studies permuted the eQTLs across bins (typically 2 cM intervals) of the genome 1,000 times, in order to capture the maximum number of eQTLs per genetic position per permutation. Subsequently, the 95^{th} percentile of the obtained distributions was then used as a confidence threshold for the occurrence of a hotspot. Potokina *et al.* (2008), Drost *et al.* (2010) and Li *et al.* (2013) also took gene density into account by further comparing the number of eQTLs linked to each bin with the number of genes identified in the same bin,

using a chi-squared test. At least seven studies used GO over-representation in order to show that the genes in some of the detected hotspots act in the same biological process or pathway. More detail on the functional enrichment strategies of five selected studies are given in the paragraph below.

West *et al.* (2007) obtained GO information from TAIR and used chi-squared tests (p-value < 0.001) to identify enriched GO-terms from the 17 identified *Arabidopsis* hotspots, with no success. Wang *et al.* (2010*b*) used TopGO, based on TIGR 5.0, for enrichment analysis of the 171 potential hotspots identified in rice. With a 0.01 p-value cut-off, they determined 21 functional terms enriched in 37 (22%) of the hotspots. Drost *et al.* (2010) identified *Arabidopsis* putative homologs using BLASTX (e-value $< 1e^{-5}$) for the populus target genes in the 7 leaf, 16 xylem and 11 root hotspots, and corresponding GO-terms were assigned. Most of the Fisher's exact tests (with Bonferroni correction) were successful and transcriptional networks were built on the basis of these hotspots. Li *et al.* (2013) used the Biological Networks Gene Ontology (BiNGO) plugin in Cytoscape based on the annotation information from AgriGO and MaizeCyc database, respectively, to identify GO and pathway enrichments of the target genes regulated by each of the 96 hotspots identified in maize. For 43% of these hotspots, enrichment for at least one GO term was evident. Thus, apart from some success stories, there also seems to be many false positive hotspots, with no biologically enriched function.

Drost *et al.* (2010) was the only study that attempted the construction of co-expression networks. Pairwise Pearson correlation coefficients between the genes with eQTLs in each $\sim$ 2 cM hotspot bin were calculated, and cases where at least 10 genes showed a correlation coefficient of $|r| > 0.80$ was called a network. Among the 97 leaf eQTL hotspot bins detected, 51 gene co-expression networks were constructed within 38 bins. The leaf co-expression networks consisted of a total of 1,678 distinct genes with 11 to 945 genes per network. Many neighboring bins produced redundant networks, and at least nine independent leaf co-expression networks were detected. Similar results were obtained for xylem and root. In addition to GO annotation enrichment per co-regulated network, transcription factor binding sites (TFBS) enrichment analyses were performed to infer the potential functional roles of transcriptional networks. Promoter sequences for all the genes on the microarray were extracted, and presence/absence of motifs in the plant *cis*-acting regulatory element sequence database (PLACE) database were determined

with Patmatch. For each motif, each co-expression network was tested for enrichment of genes carrying the motif using a Fisher's exact test. *Cis*-regulated genes belonging to a network, were identified as putative network regulators that potentially modifies downstream network functions. Drost *et al.* (2010) showed that the identified eQTL hotspots and the corresponding transcriptional networks were largely tissue-specific.

Further comparing the studies revealed that the number of *trans*-eQTL hotspots per study differs considerably, and so also the number of target genes per hotspot. Also, many hotspots were shown to have a directional bias, meaning that a significantly higher proportion of eQTLs with either positive or negative effect were observed. Specifically, Kirst *et al.* (2005), West *et al.* (2007), Potokina *et al.* (2008) and Li *et al.* (2013) used chi-squared tests to determine that 40%, 100%, 44%, and 78% of their identified hotspots respectively showed significant directional bias. This might reflect substantial regulatory differences between the parental lines.

1.5.3 Correspondence of eQTLs with phenotypic QTLs

Seven of the genome-wide eQTL studies in Tables 1.1 and 1.2 used eQTLs to identify potential candidate genes for different phenotypic traits. In most cases, new hypotheses to test in future work were constructed from these results.

Shi *et al.* (2007) studied the molecular basis for cell-wall digestibility in order to improve the feeding value of forage maize. One out of five eQTL hotspots co-localised with a cell wall digestibility related QTL cluster, implying that the gene(s) underlying these QTLs and eQTLs are identical. Jordan *et al.* (2007) based their eQTL study on the same population that was previously used for QTL mapping of agronomic and seed quality traits in wheat. They found that 17 out of 542 detected eQTLs (*cis*- and *trans*-eQTLs) corresponded to intervals that overlapped with QTLs for grain protein content and yield in wheat. Additionally, 28 eQTLs overlapped with QTLs for grain weight, maturity and several flour and dough quality traits. Potokina *et al.* (2008) used transcriptional variation in germinating barley grain to investigate the variation in malting quality phenotypes. They found that each of the analysed malting phenotypes had at least one significant QTL associated with at least one identified eQTL hotspot. Wang *et al.* (2010*b*) studied seedling vigor traits in a rice RIL population. They compared the genome-wide distributions of phenotypic QTLs and eQTLs, and calculated correlations

between the phenotype values of traits and expression levels of e-traits in the regions where eQTLs overlapped phenotypic QTLs. All three phenotypic QTLs for shoot dry weight overlapped with eQTL hotspots. In particular 93 e-traits with *trans*-eQTLs in the *SDW5-1* phenotypic QTL support interval showed significant correlations ($r = 0.3$–0.5). This suggests that genes with *cis*-eQTLs in the corresponding region, are potentially associated with early growth characteristics and regulate many genes with *trans*-eQTLs. Chen *et al.* (2010*b*) performed an eQTL analysis to study the quantitative resistance to barley leaf rust caused by *Puccinia hordei*. They identified 128 genes that were correlated with barley leaf rust resistance, of which 89 had an eQTL co-locating with the phenotypic QTL. Gene expression in the parental lines and conservation of synteny with rice, allowed them to prioritise six genes as candidates for leaf rust resistance. No eQTL hotspots co-located with any of the phenotypic QTL for leaf rust resistance. Moscou *et al.* (2011) focused on a new highly virulent race of stem rust in barley, known as TTKSK. In the example given on page 16, Druka *et al.* (2008) used a different race of *Pgt*, with a cloned resistance gene Rpg1. However, currently the *Rpg-TTKSK* locus on chromosome 5H is the only known locus that confers resistance to the aggressive *Pgt* race. In a comparison of eQTLs between pathogen-inoculated versus mock-inoculated, Moscou *et al.* (2011) revealed an inoculation-dependent expression polymorphism, *Actin depolymerising factor 3* (within the *Rpg-TTKSK* locus), as a candidate susceptibility gene based on a strong *cis*-eQTL with a magnified effect after inoculation with *Pgt* race TTKSK. Moreover, they identified a *trans*-eQTL hotspot on chromosome 2H that co-segregates with a quantitative resistance factor that acts as an enhancer of *Rpg*-TTKSK-mediated resistance in adult plants. Claverie *et al.* (2012) had access to phenotypic QTL results from a number of cotton fiber developmental stages. In the absence of a genome sequence for cotton, it was not possible to classify eQTLs as *cis* or *trans*. However, the occurrence of large hotspots clearly suggested the presence of *trans*-eQTLs. Although the data could not be used to identify *cis*-acting factors potentially causing the variation of fiber traits, eQTL hotspots overlapping with regions rich in phenotypic fiber QTL (meta-clusters) in at least 15 different cases were identified.

The studies mentioned above mainly used co-localisation of phenotypic QTLs with eQTLs and eQTL hotspots. Moscou *et al.* (2011) specifically focused on overlapping *cis*-eQTL candidates, and Wang *et al.* (2010*b*) and Chen *et al.* (2010*b*) combined co-

localisation of eQTLs with correlation analysis to identify candidate genes influencing the respective phenotypes.

1.6 Biological, technical and statistical considerations

A number of factors have an impact on the proportion of eQTLs that can be observed in a global eQTL study. Biological factors such as the assayed tissue, environmental conditions and the genotypic diversity present in the mapping population, have an influence on which genes are expressed and which have allelic variants. Furthermore, statistical factors such as population type and size, gene expression measurement accuracy including array limitations and mapping quality of RNA-seq data, the number of genes analysed and genetic map quality influence the mapping power and detection thresholds (Joosen *et al.*, 2009). Other technical limitations that could influence the interpretation of the results from an eQTL study, include the quality of functional annotations available for the transcripts being measured and whether the parental lines have genome sequences available that could reveal candidate causative polymorphisms.

1.6.1 Gene expression platform considerations

Gene expression information can be captured with a variety of techniques ranging from reverse transcription quantitative polymerase chain reaction (RT-qPCR), to DNA microarrays and more recently RNA-seq, which employs next-generation sequencing technologies. When choosing a gene expression platform for an eQTL study, it is important to consider genome representation, platform performance and the cost to capture the expression data of many individuals in a population. Fourteen out of the 16 genome-wide plant eQTL studies in Tables 1.1 and 1.2 used microarrays for population-wide gene expression profiling, mainly since it is a high-throughput method, an optimal design allows cost-effective use of the technology (Fu and Jansen, 2006), an annotated genome is not required, and data analysis approaches for microarrays are standardised.

A major limitation when using microarrays in an eQTL study, is that the analysis is limited to only the transcripts measured on the array. According to Cubillos *et al.* (2012), insufficient sensitivity and a lack of reproducibility are other drawbacks of microarray technologies. Furthermore, when using microarrays that employ short reporters

for each transcript, it could be that hybridisation differences are due to sequence polymorphisms rather than actual expression differences. For example when there are differences between the reference genome reporters and the subject genome, this may generate single feature polymorphisms (SFPs) and result in detection of false eQTLs. However generally only few genes are impacted, leaving the overall architecture not significantly affected (Kliebenstein, 2009). In this regard, Kloosterman *et al.* (2012) speculate that some of the *cis*-eQTLs they detected in a potato genome-wide study are likely to be false eQTLs due to the heterozygous nature of potato. Alberts *et al.* (2007) applied a novel statistical approach, which takes the individual reporter signals into account, to short-oligomer data from human and mouse Affymetrix microarrays that were used for eQTL mapping studies. They showed that even though many *cis*-eQTLs are falsely reported, this approach can successfully identify and eliminate these eQTLs. However, they also stated that when strong claims about *cis*-eQTLs are to be made, it is recommended to use additional methods to characterise polymorphisms, to re-sequence the reporter regions and to use alternative ways of gene expression profiling.

Spurious eQTLs can also be caused by technical confounding factors. These factors include systematic bias, such as technical variation in microarray manufacturing and variations introduced during sample preparation or expression measurements. Examples could be the use of a different batch of reagents or different room temperatures during two hybridisations. Kang *et al.* (2008*a*) propose a statistical method for eQTL mapping that corrects for the spurious associations caused by complex intersample correlation of expression measurements, provided that independent biological replicates are available. Using this method, Kang *et al.* (2008*a*) identified many more *cis* associations while eliminating most of the misleading *trans* associations.

RNA-seq provides a few advantages to transcriptome research, such as robust expression detection especially for lowly expressed genes, and the detection of all the transcripts not depending on whether they have reporters on an array (Li *et al.*, 2013). It is however important to implement quality control steps, for example the reads with low mapping quality or mapping ambiguity should be removed. Also, technical confounding factors could influence the results. Two novel elements that RNA-seq data makes possible are eQTL mapping using allele-specific expression and isoform-specific eQTL mapping (Sun and Hu, 2012).

1.6.2 Experimental design

There are three main categories of factors that influence the design of eQTL experiments: (i) type of population, (ii) population size and (iii) replication.

Populations of homozygous lines are mostly used in plant eQTL in studies. Two inbred lines are typically crossed to form a heterozygous but identical F_1 generation. These F_1 individuals are then crossed to form an F_2 generation. Homozygosity can be achieved by inbreeding, often by single seed descent to produce RILs or by production of DH lines from haploid F_1 plants. Eleven out of the 16 global eQTL studies in Tables 1.1 and 1.2 are based on RIL or DH populations. In these populations, only information regarding the additive effects of QTLs are provided since every genotype is homozygous. Also, the amount of recombination that has occurred in the production of the inbred lines, directly influence the accuracy with which eQTLs are located (Druka *et al.*, 2010). Although the specific parents inherently limit the available genetic variation, the phenotypic variation can be expanded beyond that of the parents through transgressive segregation. Transgressive segregation results from novel genotypic combinations due to recombination and independent assortment between genes (Kliebenstein, 2009).

Population sizes should be as large as possible, not only to increase the statistical power of the analysis, but also to increase the number of recombinants, which allow for better separation of genetic effects at distinct QTL locations.

In general, it is important to include an appropriate measure of replicate variation. The replicates of all genotypes should be unbiased, thus incorporating all the non-genetic factors that could cause lines to differ. These factors are (i) technical variation in sampling, preparing and assaying the samples, (ii) true biological environmental variation, which is captured by genetically identical full-sib lines, and (iii) environmental or maternal effects, thus epigenetic effects determining the phenotype of an organism (Druka *et al.*, 2010). It is generally agreed that the design should include biological rather than technical variation (Kerr and Churchill, 2001). To avoid unnecessary replication of slides, Fu and Jansen (2006) proposed a distant-pair design for two-colour microarrays.

There is a trade-off between the population size and the independent replication per line, and no absolute answer as to which is more important (Kliebenstein, 2009). Since the average transcript's heritability (the proportion of transcript abundance variance that is caused by genetic variation) in plants is roughly 60–65%, at least a two-fold replication

is necessary to maximise detection power (Keurentjes *et al.*, 2007; West *et al.*, 2007). Therefore decreasing replication will result in detection of eQTLs for transcripts with higher heritability (mostly *cis*-eQTLs). On the other hand, decreasing the population size will result in less recombination events and a limited ability to detect eQTLs with small effects since smaller sample sizes generally yield weaker estimates (Gilad *et al.*, 2008). Thus, decreasing either population size or biological replication will likely result in a decrease of the total number of eQTLs detected and an increase in the total percentage of *cis*-eQTLs versus *trans*-eQTLs.

Since the number of traits (transcript levels) measured, tend to be much larger than the number of individuals in an eQTL study, these studies are underpowered to detect and localise eQTLs (de Koning and Haley, 2005). For example in the 16 eQTL studies in Tables 1.1 and 1.2, the average number e-traits tested was 15,636 (ranging from 439 to 25,965) and the average number of individuals in a population was 132 (ranging from 39 to 360). In general, higher power allows the robust detection of more small-effect *trans*-eQTLs. Moreover, the large number of hypothesis tests required to associate a dense marker map with thousands of transcripts, makes it very difficult to control the FDR. As the cost of genome-wide expression profiling decrease, larger eQTL studies will be possible, which is hoped reduce these concerns (Mackay *et al.*, 2009).

Similar to the mapping approaches for phenotypic QTLs, approaches for eQTL mapping can be classified into linkage methods and association methods (Figure 1.1 on page 46). The main advantage with linkage mapping is that a small number of markers are sufficient for a genome-wide scan. eQTL-based association mapping is a more powerful approach for identification of common variants and for detection of eQTLs with small or medium effect, provided that the causal variants are in strong LD with the genotyped SNPs and that the genotyping is adequately dense (Gilad *et al.*, 2008). Association mapping also provides a fine-scale resolution contrary to linkage mapping which depend on the number of recombination events in a segregating population. A potential disadvantage of eQTL-based association mapping is the possibility of false positives due to population structure, however this can largely be corrected for by applying recently developed methods (see the section on page 5).

In contrast to animal studies, eQTL-based association mapping studies have not yet been applied to plants. In an outbred mouse population, Ghazalpour *et al.* (2008) applied

whole-genome association analysis to liver gene expression traits. They compared this analysis with eQTLs identified in previous studies of F_2 intercross mice and found that the mapping resolution was significantly greater in the outbred population. With high-density genome-wide genotyping currently available, association mapping will likely be the method of choice for future eQTL studies (Gilad *et al.*, 2008).

1.7 Network eQTL mapping

For certain biological processes, the contributing genes are well known. However, for most biological processes little is known about the regulation and interaction of the genes involved. Exploiting the data from genome-wide eQTL studies could lead to the identification of groups of genes potentially involved in the same biological processes, and consequently regulated by the same regulators. "Network eQTL" analysis allows the identification of genetic variation influencing entire processes, thereby revealing polymorphisms upstream in networks (Hansen *et al.*, 2008). Figure 1.4 summarises the two major approaches, classified as *a priori* and *a posteriori*.

In an *a priori* analysis, previous information such as pre-selected pathways is used to group genes into "networks". Thus the group of genes being tested must be known or at least predicted to be involved in a specific biological process (Hansen *et al.*, 2008). A network expression value is then calculated for each individual in a mapping population, usually by averaging across the expression values of the genes grouped as a network. The resulting network expression values are subsequently used as the trait in QTL analysis. The result is a single LOD profile per network that was studied (Kliebenstein *et al.*, 2006) (Figure 1.4).

Kliebenstein *et al.* (2006) mapped network eQTLs in a RIL population derived from accessions Bay-0 and Shahdara, for 18 known networks mainly involved in plant defense including glucosinolate and flavonol biosynthesis. Interestingly, a *cis*-eQTL for the transcription factor PAP1, known to regulate flavonol biosynthesis was found to co-locate with a network eQTL for the flavonol biosynthesis pathway. This result suggests that the *cis*-variation in the expression of PAP1 is also responsible for the flavonol biosynthesis network eQTL.

In an *a posteriori* analysis, eQTL data is typically utilised to generate novel "net-

works". Correlation of expression patterns or co-localisation of eQTL positions can be used to identify clusters of potentially co-regulated genes (Figure 1.4). Lan *et al.* (2006) was able to predict regulatory networks by combining the correlation results with eQTL mapping information. Similar to the *a priori* analysis, after averaging across the expression values of the grouped genes, network eQTLs can be mapped. The identified genetic loci can subsequently be searched for regulatory genes with *cis*-eQTLs that potentially regulate these networks (Sun *et al.*, 2007).

Keurentjes *et al.* (2007), used a hybrid *a posteriori* approach to an *Arabidopsis* eQTL study with a set of 175 flowering time genes. They utilised eQTL information to define connections and regulatory hubs within the flowering time network. Numerous unknown regulatory interactions related to flowering time were predicted, and many previously known regulators were confirmed.

According to Kliebenstein (2009) there are two concerns when using target genes from a *trans*-eQTL hotspot in an *a posteriori* analysis. Firstly, it could be difficult to interpret the biological meaning of massive hubs with thousands of genes, which is the case for some hotspots. Secondly, due to limited population sizes there is currently not sufficient recombination to accurately partition *trans*-eQTL hotspots, to ensure that all the transcripts are affected by the same genetic polymorphism.

1.8 Systems genetics of quantitative traits

1.8.1 Regulatory network reconstruction

Genetical genomics data can be used for regulatory network reconstruction. Instead of grouping genes with *trans*-eQTLs at an identical position, co-expressed genes across the individuals in a population can be identified and used as a powerful basis for regulatory network reconstruction.

Even though thousands of transcripts are genetically variable, they are not independent. This means that the expression levels of some transcripts may be correlated, which may suggest that these genes belong to a common regulatory network. When genes have higher correlations to each other than to the genes in the rest of the transcriptome, they can be grouped into co-expression clusters also called "modules". This can significantly reduce the dimension of the data, since the statistical information encoded in highly corre-

lated transcripts is redundant (Mackay *et al.*, 2009). To determine whether co-expression networks are biologically sound, enrichment analyses are generally performed. Such analyses utilise GO categories, KEGG pathways, protein-protein interactions, tissue-specific expression patterns or TFBS (Miller *et al.*, 2008; Ayroles *et al.*, 2009). Furthermore, functions of genes without annotations can be predicted based on "guilt by association" with well-annotated genes in the network (Tian *et al.*, 2008).

A module can be visualised graphically as a network, with nodes denoting transcripts and edges connecting nodes that are correlated based on co-expression. In order to build co-expression gene networks, one needs to calculate a pairwise correlation matrix with the full set of transcripts. From this, an adjacency matrix is constructed, which encodes whether and how a pair of nodes is connected (Zhang and Horvath, 2005). One option is to encode gene co-expression using binary information (connected=1, unconnected=0). In this case, if the correlation in transcript abundance for a pair of transcripts exceeds a threshold value, 1 is assigned, otherwise 0. Another option to encode gene co-expression is to use connection weights. These are normally values between 0 and 1, for example the absolute correlation coefficients between pairs of transcripts. More complicated connection weights in adjacency matrixes can be defined by adjacency functions. For each node (transcript), the connectivity (also known as degree) is defined as the sum of connection strengths with the other network transcripts. Hub nodes are those transcripts in modules with high connectivity (Langfelder and Horvath, 2008). Importantly, these networks do not represent direct interactions, but rather indirect statistical relationships. Information about DNA polymorphisms, such as knowledge about *cis* and *trans*-eQTLs can be used to specify the direction of flow of information in resulting networks (Mackay *et al.*, 2009).

The expression profile of a causal transcript is genetically correlated with the associated quantitative phenotypic trait. However, often hundreds of transcripts are identified as being correlated with a single phenotype (Ayroles *et al.*, 2009). Since these transcripts also correlate with each other, it is possible to identify co-expression modules that correlate with quantitative phenotypic traits. Module-based approaches can thus be used to uncover pathways and processes associated with phenotypic traits (Miller *et al.*, 2008; Ayroles *et al.*, 2009). However, causal relationships cannot be identified from a study of correlated transcripts alone. DNA sequence variation needs to be incorporated (Mackay *et al.*, 2009).

1.8.2 Module-based network analysis

Langfelder and Horvath (2008) published an R package for weighted gene co-expression network analysis (WGCNA) based on the framework proposed by Zhang and Horvath (2005). Zhang and Horvath (2005) use the scale-free topology criterion to determine the parameters of the adjacency function in the construction of co-expression networks. Approximate scale-free topology is a fundamental property of co-expression networks (Barabási and Bonabeau, 2003). It implies the presence of hub nodes that are connected to a large number of other nodes. These networks are robust with respect to the random deletion of nodes, however not to the targeted deletion of hub nodes. In order to detect gene modules of tightly co-regulated genes, Zhang and Horvath (2005) adopted the definition of Ravasz *et al.* (2002). Ravasz *et al.* (2002) describes modules as groups of nodes that have high topological overlap and use hierarchical clustering for module detection. After defining a dissimilarity measure between the genes to be used in clustering, modules are identified as branches of the hierarchical clustering tree.

Apart from identifying modules of highly correlated genes, each module identified by WGCNA is summarised with a module eigengene. A module eigengene is defined as the first principal component of a given module and can be considered a representative of the gene expression profiles in a module. WGCNA also provides functions that can relate modules to one another, as well as to external phenotypic traits, using eigengene network methodology. Lorenz *et al.* (2011) used microarray analysis to study root complementary DNA (cDNA) populations obtained from 12 genotype×treatment combinations in drought-stressed roots of loblolly pine (*P. taeda* L.). WGCNA was used to mine the 2445 differentially expressed genes for candidate regulatory genes. A scale-free network topology was predicted and 11 co-expression modules were identified, ranging in size from 34 to 938 genes. Furthermore, a number of central hub nodes were identified, some of which have previously been associated with osmotic stress. Miller *et al.* (2008) used WGCNA to study transcriptional networks in Alzheimer's disease. Twelve modules of genes with high topological overlap were identified. The module eigengenes of the 12 modules were correlated to relevant clinical traits using the Pearson correlation. Miller *et al.* (2008) also applied WGCNA to compare functional modules defined in Alzheimer's disease with those defined in normal aging. Two biologically relevant modules were conserved between the two conditions and several hub genes were identified in both aging and Alzheimer's

disease.

Modulated modularity clustering (MMC) is an example of another statistical method that has been developed to group genetically correlated transcripts into modules. This method, available as a web server, seeks community structure in graphical data. Community structure aims to group nodes into (potentially overlapping) sets, such that each set is densely connected internally. Thus MMC adjusts the connection strengths of edges in a weighted graph to maximise an objective function that quantifies community structure. This produces a final clustering with tightly-connected groups of genes (Stone and Ayroles, 2009). This approach was validated, by demonstrating that the clusters obtained through analyses of human and *Drosophila melanogaster* expression data are biologically meaningful.

1.8.3 Systems genetics: adding DNA sequence variation

To understand the connections between genotypes and phenotypes, systems genetics employs large-scale, high-throughput and comprehensive analysis. Instead of examining the effects of genes one by one, it investigates how gene networks interact to determine the traits of organisms. It aims to capture the flow of information from DNA to the organismal phenotype through RNA intermediates, proteins, metabolites and other molecular endophenotypes (Mackay *et al.*, 2009). Peidis *et al.* (2010) describes systems genetics as the analysis of gene co-expression within genetic populations. A systems genetics approach to dissecting quantitative phenotypic traits merges phenotype, genotype and gene expression data in order to prioritise mapped genes and identify gene networks associated with the phenotypic trait (Figure 1.5).

In 40 *Drosophila melanogaster* wild-derived inbred lines, Ayroles *et al.* (2009) used a systems genetics approach to study complex traits. Whole genome variation in transcript abundance for young males and females of each line, were assessed using Affymetrix arrays. In order to study the genetic variation in transcript abundance, analysis of variance was used to partition variation in expression between sexes, among lines, and the sex×line interaction for each expressed transcript. Several hundred transcripts and SFPs were identified as being associated with phenotypic variation in each of the quantitative traits. Phenotypes of P-element insertional mutations in or near candidate genes were tested and 70% of the mutants tested significantly affected the traits. After the pair-

wise correlations among the 10,096 variable transcripts were computed, 241 biologically plausible modules of highly interconnected genes were identified. Finally, 26 modules of correlated transcripts were identified to be associated with chill coma recovery time, 20 with fitness, 11 with starvation stress resistance, 10 with life span, and 9 each with locomotor reactivity and copulation latency.

Park *et al.* (2011) used a systems genetics approach to explore the genetics of conditional fear in mice. A hybrid mouse diversity panel with high mapping resolution was used to map 27 fear-related behavioral QTLs with a GWA approach. The gene expression measures of 25,697 transcripts were used to map eQTLs, also using association mapping, from hippocampus and striatum tissues respectively. WGCNA was used to identify 30 modules in hippocampus and 25 modules in the striatum. By correlating the resulting module eigengenes to behavioral phenotypes, groups of genes relating to aspects of conditional fear were identified. The context immobility phenotype showed the strongest correlations with two module eigengenes, $r = -0.43$ and $r = 0.4$ respectively, in the hippocampus. In a subsequent analysis, the module eigengenes were considered as quantitative traits and QTLs for groups of co-expressed genes were mapped. Potential key drivers influencing the expression of gene modules with relationships to fear-related phenotypes were identified. The Network Edge Orienting (NEO) software was used to fit a model that implicates a marker as causal for a phenotypic trait through expression of a gene (Aten *et al.*, 2008). This analysis revealed five genes with causal relationships for fear-related phenotypes. This study is an excellent example where the authors surely succeeded in bringing together all the datasets and analyses in Figure 1.5.

GeneNetwork (`http://www.genenetwork.org`) is a group of linked data sets and tools that can be used to explore systems genetics data in humans, mice, rats, *Drosophila*, barley and *Arabidopsis*. These population data sets are mostly linked with dense genetic maps that can be used to locate the genetic modifiers that cause differences in expression and phenotypes. Peidis *et al.* (2010) used GeneNetwork to predict a transcriptional role for the P2P-R gene in genetic reference panels of recombinant inbred strains of rat adipocytes and mouse eye, respectively. The results revealed that biological networks of 75 and 135 transcription-associated gene products for rat adipocytes and mouse eye respectively, are co-expressed with P2P-R in a genetically-defined way.

Community projects to determine whole-genome sequences, catalogue SNPs and

structural variants, and provide comprehensive phenotypic descriptions of thousands of individuals in linkage or association mapping populations are vital for the success of systems genetics approaches. These datasets in conjunction with sophisticated systems genetics techniques, available in a similar fashion to GeneNetwork mentioned above, will speed up the process of effectively linking causal molecular variants with organismal phenotypes.

1.9 *Cercospora zeina*-maize plant pathosystem

1.9.1 Maize as an important crop

Maize (*Zea mays* L.) is a major cereal crop that is grown widely throughout the world in diverse environments. Maize is an important staple food for southern and eastern Africa, while it is largely used as livestock feed and as a raw material for industrial products in developed countries. The $10,000$ years of domestication from its wild relative teosinte, makes this crop an excellent example of selection of desirable allelic diversity within a plant species (Buckler *et al.*, 2006). During initial stages of maize breeding, mass selection played a large role. Currently, structured breeding programmes are in place for F_1 hybrid development. Due to growing demands for food and fuel, global climate change and the potential for increased disease pressure, breeders are challenged to produce higher yielding and more resistant maize cultivars (Wei *et al.*, 2009*b*).

Maize is an important model species for biological research and knowledge gained from maize research can also be used to genetically improve its grass relatives such as sorghum, wheat and rice. Maize is diploid and the genome sequence of a public inbred line B73, which is used extensively in breeding programmes worldwide, was published in 2009 (Schnable *et al.*, 2009). Maize has 10 chromosomes and a genome size of approximately 2,300 Mb. It is thought that more than 80% of the maize genome sequence is composed of transposable elements (Wei *et al.*, 2009*a*). Springer *et al.* (2009) used comparative genomic hybridization to compare the genome structures of two maize inbred lines, B73 and Mo17. Their study confirmed that maize is a highly polymorphic species, although large genomic regions that have little or no variation were also identified. The extraordinary phenotypic diversity between maize inbred lines can therefore be ascribed to genome content variation and consequently large differences in transcript content.

1.9.2 GLS disease of maize

Grey leaf spot (GLS) is a devastating foliar disease of maize and growers worldwide spend millions on fungicides annually. GLS is caused by the fungal pathogens *Cercospora zeae-maydis* Tehon and E. Y. Daniels and *Cercospora zeina* Crous & U. Braun. Symptoms are necrotic lesions on the leaf surface. GLS has been documented in many sub-tropical and tropical regions of the world (Ward and Nowell, 1998). It is an economically significant disease to maize production in the eastern United States of America (USA) and one of the major constraints to maize production in sub-Saharan Africa where up to 70% yield losses have been reported (Ward *et al.*, 1999). In South Africa, the occurrence of GLS is most devastating for farmers in the KwaZulu-Natal province (Cedara, 1996). Resistant commercial hybrids are not readily available in Africa and other developing countries, especially to small-holder farmers. Even though hybrids with resistance are available for some areas of the world, other genetic characteristics such as yield or growing season length are often compromised in more resistant hybrids. Latterell and Rossi (1983) described GLS as "a disease on the move" and as it continues to expand its geographic distribution and severity, GLS has lived up to the 1983 prediction.

Two variants of *C. zeae-maydis* have been recognised as Types I and II, until Crous *et al.* (2006) renamed it as *C. zeae-maydis* and *C. zeina*. Meisel *et al.* (2009) later confirmed *C. zeina* to be the only causal agent of GLS disease on maize in southern Africa. *C. zeae-maydis* differs from *C. zeina* in that it has faster growth rate, the ability to produce the toxin cercosporin, longer conidiophores and broadly fusiform conidia (Wang *et al.*, 1998; Crous *et al.*, 2006). *C. zeae-maydis* has been documented in North and South America and China; and *C. zeina* in Africa, Brazil and the eastern USA.

GLS life cycle, maize symptoms and impact on yield

C. zeina is known to only infect maize. The fungi overwinter as stromata, a mixture of plant tissues and fungal mycelium, in infected maize residues on the soil surface. During early spring, the fungi start sporulating and conidia (asexual spores) are dispersed by wind or rain splash within and among newly planted maize fields. Lower leaves are usually the sites of primary infection. In humid conditions, hyphae emerge and grow across the leaf surface and penetrate the leaf mesophyll via stomata (Beckman and Payne, 1983; Lyimo *et al.*, 2013). After growing intercellularly for approximately three weeks

(Wisser *et al.*, 2011), the growth habit switches to necrotrophy, resulting in necrotic expanding lesions. Following lesion formation, conidiophores emerge through stomatal pores. Conidia are dispersed by wind or water to upper leaves or neighboring plants and secondary infection cycles are initiated (Ward *et al.*, 1999; Kim *et al.*, 2011). During favorable climatic conditions (moderate to high temperatures and high humidity) disease progress can be rapid, resulting in increasing lesion numbers on developing leaves higher in the canopy. In prolonged favorable conditions, developing lesions may coalesce, which ultimately results in necrosis of leaf tissue. The fungus can stay dormant in unfavorable climatic conditions and continue its development once conditions are favourable again (Thorson and Martinson, 1993).

Symptoms are typically first observed on the lower leaves, which gradually spread upwards on the plant during the season. Initial immature lesions appear as small tan spots with chlorotic borders, not readily distinguished from other foliar diseases of maize. However, more mature GLS lesions can be easily identified by characteristic rectangular shapes running within leaf margins, as the fungi is not able to penetrate sclerenchyma tissue in the leaf veins. These lesions are grey due to sporulation. Further lesion expansion leads to coalescing, which later result in the blighting of entire leaves. With severe blighting, stalks deteriorate and severe lodging may occur (Ward *et al.*, 1999). Susceptible genotypes commonly display numerous necrotic lesions, while moderately resistant genotypes exhibit fewer and smaller lesions as a result of prolonged latent and incubation periods, as well as reduced infection rates and sporulation capacity (Menkir and Ayodele, 2005). Symptoms of GLS caused by *C. zeae-maydis* and *C. zeina* are indistinguishable.

During the grain fill stages in maize, most of the photosynthate produced by the plant is deposited in the developing kernels and this takes priority over the photosynthate demands of the rest of the plant. Therefore loss of photosynthetic leaf area, associated with GLS disease, result in sugars being diverted from the stalks for grain filling, which predisposes plants to lodging. During severe disease pressure, the blighting and premature death of leaves, severely limits the production and translocation of photosynthate to developing kernels, resulting in yield loss (Ward *et al.*, 1999).

GLS control measures

The most effective control measures depend on understanding the epidemiology of the pathogen and the factors affecting disease development (Ward and Nowell, 1998). Integrated pest management (IPM) relies on a combination of all practical and environmentally sensitive practices, which may aid in prevention and reduction of the disease in a specific region. IPM practices for GLS control aimed at reduction of initial inoculum include (i) tillage practices, (ii) crop rotation and (iii) early harvesting, whereas IPM practices aimed at reducing the rate of disease development include (iv) choice of shorter-season hybrids, (v) early planting, (vi) chemical control, (vii) optimum plant density, (viii) irrigation, (ix) soil fertility and (x) host resistance. The use of resistant or tolerant cultivars is the most cost-effective means of managing GLS (Saghai Maroof *et al.*, 1996; Ward *et al.*, 1999). However, very few hybrids have adequate resistance to prevent yield losses due to GLS in commercial maize production.

GLS resistance is thought to be conferred by a small number of quantitative loci with additive effects (Clements *et al.*, 2000; Menkir and Ayodele, 2005), thus several genes for resistance need to be included in a hybrid to obtain a high level of resistance. One way of improving GLS resistance is to follow conventional polygenic breeding programs, which may take several years especially if other traits, such as yield, are also desirable. An alternative is to use quantitative trait locus (QTL) mapping, with high-density molecular marker maps, to identify markers closely related to resistance. These markers can be used in marker-assisted selection programs to accelerate the development of high-yielding GLS resistant hybrids (Saghai Maroof *et al.*, 1996). No GLS resistance genes have been cloned to date. There is an increased reliance on genetic resistance to ensure sustainable disease management in the long-term (Kim *et al.*, 2011). In order to reach this expectation, considerably more information is needed about the genetic architecture of GLS disease in maize.

1.10 Plant defense mechanisms against pathogens

Plants use combinations of structural characteristics and biochemical reactions to resist pathogen invasion and possess both pre-existing and inducible mechanisms for this purpose. As a result, pathogens initially need to avoid or overcome preformed morphological

barriers, secondary metabolites and antimicrobial proteins in order to invade a plant. Furthermore, once contact has been established, further defenses are induced consisting of the reinforcement of cell walls, the production of secondary metabolites (phytoalexins), the accumulation of reactive oxygen species (ROS) and the synthesis of defense-related proteins. The speed and magnitude with which these mechanisms are activated, their effectiveness against individual pathogens with different modes of attack as well as the plant's ability to detect the pathogen early, likely determine the degree to which a plant is susceptible or resistant (van Loon *et al.*, 2006).

Plant pathogens are often divided into two groups: (i) biotrophs feed on living host tissue; and (ii) necrotrophs kill host tissue and feed on the remnant. *C. zeae-maydis* and *C. zeina* are not linked to host cell death during early stages of infection when the fungus multiplies intercellularly, but are in fact associated with host tissue chlorosis and necrosis during later stages of infection. These fungi are therefore considered hemibiotrophs.

Overview of induced biochemical defenses

Jones and Dangl (2006) summarised the plant immune system as a multi-phase model. According to this model, the evolutionary arms race between host and pathogen results in an oscillation between compatible and incompatible states over time (Figure 1.6).

The basal defense system relies on the recognition of pathogen-associated molecular patterns (PAMPs) by plant pattern recognition receptors (PRRs). PAMPs are features common to many of pathogens such as lipopolysaccharides, chitins, glucans and flagellins. This recognition results in PAMP-triggered immunity (PTI), which serves as an early warning for the activation of defense-related genes. Adapted pathogens produce a suite of effector proteins, encoded by Avr (avirulence) genes, which are delivered directly into the plant cells to suppress these defenses. This results in effector-triggered susceptibility (ETS). As a countermeasure to Avr/effector proteins, plants have evolved to synthesise resistance (R) proteins. The most prominent class of R proteins is the nucleotide-binding site plus leucine rich repeat (NBS-LRR) protein, which can interact with Avr/effector genes to activate effector-triggered immunity (ETI). Both the detection of PAMPs by PRRs and the interaction of Avr-R, result in the activation of a signalling cascade that induces defense response genes. Biosynthesis pathways of phytohormones such as salicylic acid (SA), jasmonic acid (JA) and ethylene (ET) are initially activated. The resulting

phytohormones further activate the production of pathogenesis-related (PR) proteins and initiate various processes such as the hypersensitive response (HR), which is a form of programmed cell death, or systemic acquired resistance (SAR) to limit further invasion by the pathogen in distal parts of the plant (Ryals *et al.*, 1996; Heath, 2000).

The balance of hormonal crosstalk strongly influences the outcome of plant-pathogen interactions. Classically, SA signalling triggers resistance against biotrophic and hemibiotrophic pathogens, whereas a combination of JA and ET signalling activates resistance against necrotrophic pathogens (Glazebrook, 2005). However, other hormones such as abscisic acid (ABA), auxin, cytokinins (CKs), gibberellic acid (GA) and brassinosteroids also play a role in molding plant-pathogen interactions. These hormones may influence disease outcomes through their effect on SA or JA signalling (Robert-seilaniantz *et al.*, 2011). GA causes degradation of the DELLA protein growth repressors, which result in induced accumulation of ROS and SA, as well as reduced JA signalling (Navarro *et al.*, 2008). CKs contribute to resistance against biotrophs by enhancing the SA response through non-expressor of PR genes 1 (NPR1), a master regulator of SA defense signalling (Choi *et al.*, 2010). Brassinosteroid treatment enhances biotroph-hemibiotroph resistance (Nakashita *et al.*, 2003). Auxin signalling initiates the suppression of SA biosynthesis and signalling (Robert-seilaniantz *et al.*, 2007). ABA biosynthetic and signalling pathways promote disease susceptibility to several plant pathogens by suppressing SA-dependent signalling mechanisms (Audenaert *et al.*, 2002). However, plant hormones interact in complex networks and in many cases there is evidence for both positive and negative interactions.

PR proteins generally possess antifungal or antimicrobial activity through hydrolytic activities on cell walls, contact toxicity and an involvement in defense signalling. Most PR proteins are induced through the action of the signalling compounds SA, JA or ET (van Loon *et al.*, 2006). Seventeen recognised families of PR proteins exist. Well known PR protein families include β-1,3-glucanases (PR-2), chitinases (PR-3), thaumatin-like proteins (PR-5), proteinase inhibitors (PR-6), peroxidases (PR-9), defensins (PR-12) and thionins (PR-13). Although minor levels of PR proteins may be present in healthy plants, attack by pathogens, wounding or stress generally induce transcription of genes that encode PR proteins (Agrios, 2005).

Plants also defend themselves via the production of secondary metabolites such as

phenolics and phytoalexins that are toxic to pathogens. Plant phenolics include several structurally diverse classes of natural products from the shikimate-phenylpropanoids-flavonoids pathways. Plants need phenolic compounds for pigmentation, growth, reproduction, pathogen resistance and for many other functions (Lattanzio *et al.*, 2006). Terpenoids (terpenes) occur in all plants and represent the largest class of secondary metabolites. Monoterpenoids are the primary components of essential oils, which are highly volatile compounds with important aromatic qualities (Freeman and Beattie, 2008). According to Cowan (1999), terpenes may have the ability to disrupt microbial membranes, which may explain their antimicrobial properties.

Qualitative versus quantitative resistance

Two general categories of genetic control of disease resistance in plants are: (i) qualitative resistance, conditioned by a single gene using gene-for-gene recognition mechanisms and (ii) quantitative resistance, conditioned by multiple genes of partial effect, i.e. controlled by multiple genetic factors. Resistance genes (R-genes) normally refer to genes that confer qualitative effects and QTL refer to loci or genes that confer quantitative disease resistance. Although qualitative and quantitative resistance can be considered different, the overlap between the extremes has raised the question of whether the two types of resistance are conditioned by the same genetic mechanisms (Poland *et al.*, 2009).

Qualitative resistance is usually accompanied by an oxidative burst, taking place in early plant defense mechanisms. Oxidative burst is the rapid release of reactive oxygen species (ROS), which is mainly superoxide (O_2^-) and hydrogen peroxide (H_2O_2). According to Bolwell (1999), ROS has numerous roles including direct killing of the pathogen, involvement in structural changes in the cell wall, the induction of defense gene expression as well as promotion of programmed cell death (PCD) during HR. R-genes often lack durability, since a single loss-of-function mutation in an Avr/effector protein enable the pathogen to escape recognition by its corresponding R-protein (van der Biezen and Jones, 1998). However, R-genes are expected to participate in coevolutionary arms races where plant specificity and pathogen virulence continually adapt in response to each other (Bergelson *et al.*, 2001). NBS-LRR proteins, the major class of R-proteins, serve as complex intracellular receptors that have both perception and signaling roles in activating defenses (Steinbrenner *et al.*, 2012). NBS domains are involved in signalling, and

include several highly conserved and strictly ordered motifs (Tan and Wu, 2012). LRRs are highly adaptable receptor domains and may be involved in direct protein-protein interactions with Avr/effector proteins of the pathogen. The solvent-exposed amino acid residues of LRRs often evolve at unusually fast rates, suggesting that they have evolved to detect variation in pathogen-derived ligands (Bergelson *et al.*, 2001). Some R-genes encode receptor-like protein kinases (RLKs), with an extracellular domain that is involved in signal recognition, a transmembrane domain and a cytoplasmic serine-theonine kinase domain for initiating a signal transduction cascade in the cell. Many R-proteins are activated indirectly by Avr/effectors, and not by direct recognition. This "guard hypothesis" implies that some R-proteins bind to other plant proteins, which are targeted and modified by the pathogen (de Wit, 2002; Marathe and Dinesh-Kumar, 2003). It appears that in plant genomes, most R-genes exist as clustered gene families of varying sizes that are thought to assist in rapid R-gene evolution. An example of this is the tomato Pto locus, where an NBS-LRR gene Prf lies within the cluster of five kinase genes (Hulbert *et al.*, 2001). Furthermore, R-gene clusters often reside in mega-clusters where smaller clusters are localised within a few million base pairs of one another. According to Young (2000), the organisation of R-gene clusters emphasise a tension between diversifying and conservative selection. Clustering of R-genes are thought to be due to tandem duplications (duplicated genes adjacent to the original) or ectopic duplications (duplicated genes translocated to distal locations in the genome resulting in mixed clusters) followed by local rearrangements and gene conversion (Marone *et al.*, 2013).

Quantitative resistance is often associated with resistance to necrotrophic pathogens, whereas qualitative resistance is generally associated with resistance to biotrophic pathogens (Balint-Kurti and Johal, 2009). The latter corresponds with activation of SA-dependent signalling pathways and hypersensitive cell death. This is effective, since HR deprives the pathogens of a food source. In contrast, necrotrophic pathogens benefit from host cell death and therefore effective defense would rather be to activate a different set of defense responses by JA and ET signalling (Glazebrook, 2005). Interestingly, Lincoln *et al.* (2002) demonstrated that transgenes encoding anti-apoptotic proteins from mammals could provide resistance to necrotrophs in plants, since these transgenes evidently interfere with activation of PCD in the host. Quantitative resistance provides non-race-specific, intermediate levels of resistance and as a result is more durable. The

mechanisms of quantitative resistance are more difficult to characterise and it may vary for different plant-pathogen interactions (Balint-Kurti and Johal, 2009). Poland *et al.* (2009) stated that it is unlikely that the model proposed by Jones and Dangl (2006) (Figure 1.6) accounts for all known forms of quantitative resistance and that it has been primarily constructed based on observations of biotrophic pathogens for which R-gene-mediated recognition is effective.

Identifying the genes (and ultimately the genetic polymorphisms) underlying QTLs that confer quantitative disease resistance is currently a major challenge. Since DNA variations impact complex diseases through the perturbations they cause to transcriptional, protein and metabolite networks, these molecular phenotypes are intermediate to the phenotypic effect. Therefore, the integration of biological networks (e.g. transcriptional networks) with DNA variation and phenotypic data has the potential to aid in identification of the associations between DNA variation and quantitative disease resistance, as well as to characterise parts of the molecular networks that drive quantitative resistance (Sieberts and Schadt, 2007).

1.11 Future perspectives

A general concern with current eQTL studies is that populations are normally sampled at one developmental stage in a single environment. The broad use of this data assumes stable genetic variation, which is not necessarily the case. Kliebenstein (2009) also noted that eQTL studies are mostly conducted in the conditions of interest, for example pathogen-infected plants are used to provide information about pathogen resistance. Whether it is necessary to compare an analysis like this with a similar analysis on uninfected plants in order to make strong conclusions remain to be seen. Druka *et al.* (2008) started showing that genetic variation, specifically *cis*-variation, in a defined population allowed the transfer of eQTL information to experiments conducted in different conditions, tissues, and years. Such tests of association to multiple phenotypes for the same genotypes are necessary in the future, for a clearer understanding of pleiotropy (Mackay *et al.*, 2009).

Currently genome scans in large populations allow the detection of multiple QTL regions. However, for each QTL, candidate causal sequences need to be identified and in-

dependently verified. The promise and expectation offered by next-generation sequencing technologies in this regard is substantial. Accessibility to full DNA genome sequences, fine-scale genotyped mapping populations, phenotypic descriptions of thousands of individuals in these populations and high-throughput RNA sequencing technologies for gene expression profiling, will speed up the process of identifying causal polymorphisms, potentially to the point of simultaneous detection and localisation of QTLs (Mackay *et al.*, 2009).

Certainly also in the near future, the combination of GWA studies with classical linkage and eQTL mapping strategies, studied under different environmental conditions, is hoped to expand the repertoire of interacting *cis*-acting and *trans*-acting variants. This will improve our understanding of how variation within regulatory elements affects gene expression, which will allow the prediction of mechanisms by which these elements shape phenotypic diversity in natural populations (Cubillos *et al.*, 2012).

Interestingly, the focus of general genotype–phenotype association studies is predicted to shift. Instead of assessing multilocus genotypes, the challenge will become to obtain multidimensional phenotypes for large numbers of individuals (Mackay *et al.*, 2009). Thus the development of high-throughput methods for automated phenotyping will be highly advantageous. Examples of two recent studies in this regard include the development of a nondestructive imaging and analysis system for automated phenotyping and trait ranking of root system architecture in rice (Iyer-Pascuzzi *et al.*, 2010), and the development of a system to automatically measure plant characteristics of tall pepper plants in the greenhouse (van der Heijden *et al.*, 2012).

Systems genetics integrates the questions and methods of systems biology with those of genetics, to interrelate genotype and phenotype in quantitative traits (Nadeau and Dudley, 2011). The incorporation of expression patterns of genes and gene modules, contributes in this regard to elucidate the complex molecular networks underlying phenotypic traits. It is important to note that not all functional molecular polymorphisms affecting phenotypic traits will result in differential gene expression. With technologies advancing and costs decreasing, it will be possible to conduct systems genetics analyses on larger samples, more environmental conditions, more developmental time points and more tissues. A more complete picture of the effects of genetic perturbations on whole organisms will be seen once information from various sources, including proteins and

metabolites, as well as epigenetic modifications, are added. Furthermore, sizable and interactive databases to manage the different types of data and new statistical methodology to infer significant biological networks will be needed. Finally, all organisms will become model organisms, which will enable us to understand the genetic basis of many phenotypic traits, including ecological specialisations and adaptations (Mackay *et al.*, 2009).

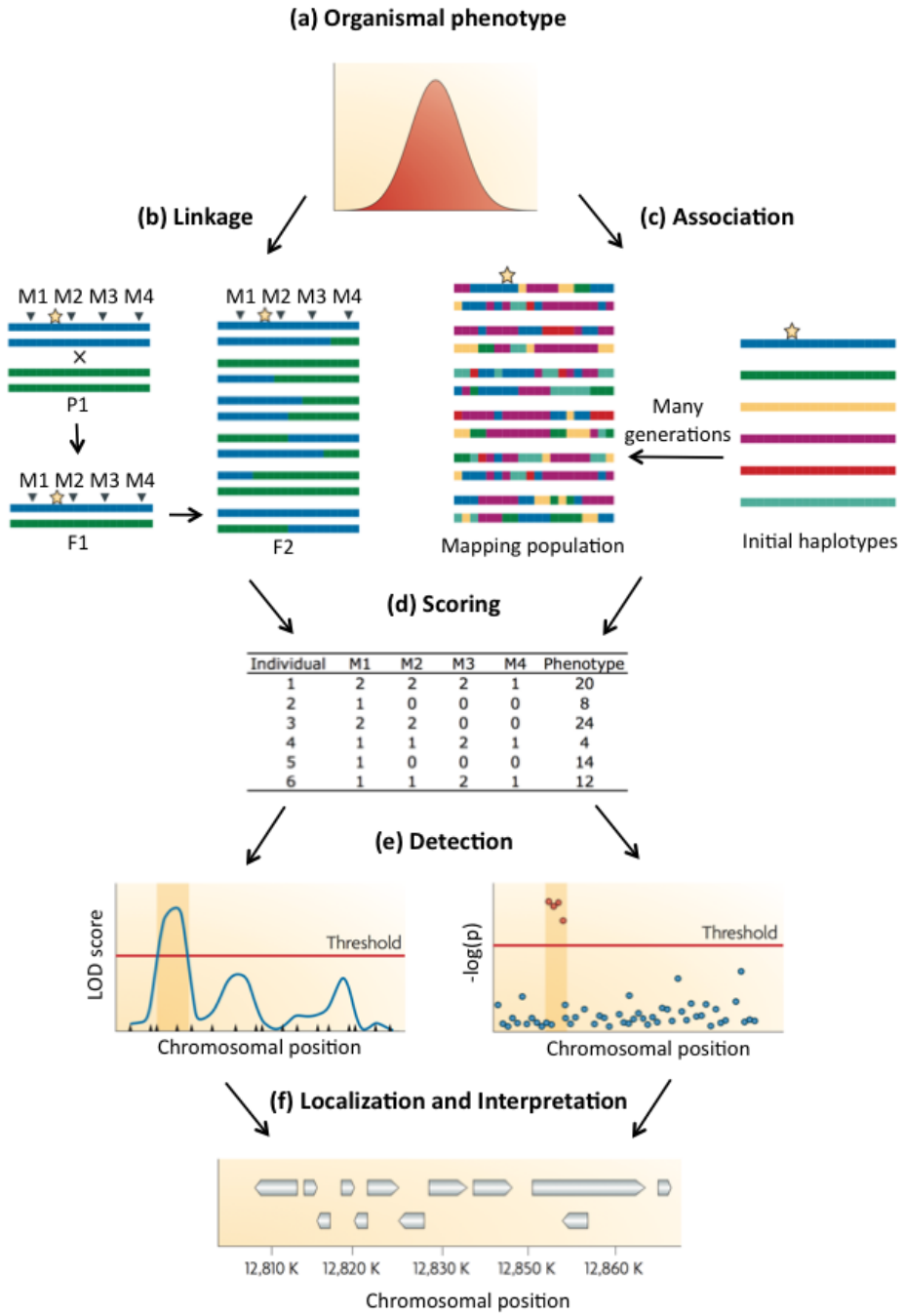


Figure 1.1: QTL mapping. Adapted from Mackay *et al.* (2009). **(a)** QTL mapping aims to discover the genetic basis of an organismal phenotype with a quantitative distribution in trait values. **(b)** Linkage-based analyses use related individuals to identify segregating markers linked to the phenotype. M1, M2, M3 and M4 are markers that distinguish the two parental lines. The yellow star marks the position of a causal locus. **(c)** Association mapping is based on historical recombination that has effectively shuffled the initial haplotypes, in order to uncouple all but the most tightly linked markers from the causal locus. **(d)** In both approaches, phenotypes and marker genotypes are scored using the mapping population. **(e)** A QTL is detected (the marker is linked to the causal locus) if there is a mean difference in the trait phenotypes between marker genotype classes. **(f)** The causal locus is usually mapped to a smaller genomic region, due to smaller haplotype blocks, for association mapping compared to linkage-based studies. The QTL region allow the identification of candidate genes for future study.

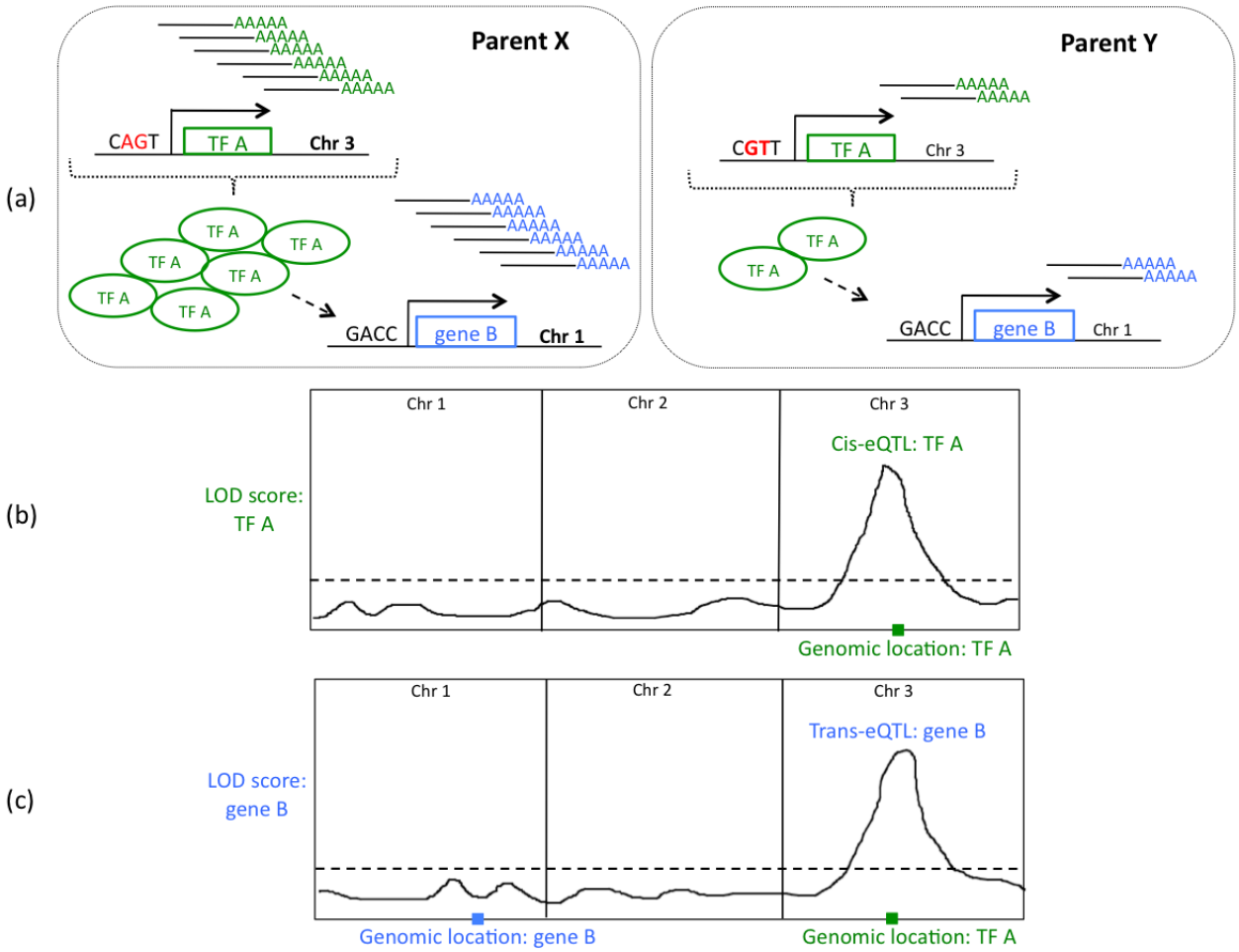


Figure 1.2: An example of *cis*- versus *trans*-eQTLs. Adapted from Hansen *et al.* (2008). **(a)** Expression of transcription factor A (TF A) and gene B, a regulatory target of TF A, in the parental lines of a segregating population. The gene expression values of TF A and gene B across all individuals in the population were used to map eQTLs for TF A and gene B, given in (b) and (c). The protein level of TF A is indicated by the number of green ovals. An expression polymorphism for TF A is observed, which in turn cause an expression polymorphism of gene B between parental lines X and Y. **(b)** An eQTL for TF A is present on chromosome 3. Since the genomic location of the eQTL coincide with the position of TF A, this is a *cis*-eQTL. The *cis*-eQTL is due to a polymorphism in the promoter of TF A, marked in red in (a), causing the expression polymorphism of TF A between parental lines X and Y. **(c)** An eQTL for gene B is also present on chromosome 3. Since the genomic location of the eQTL does not coincide with the position of gene B, this is a *trans*-eQTL. However, the eQTL for gene B coincide with the position of TF A. The *trans*-eQTL is due to a polymorphism in the promoter of TF A, marked in red in (a), causing an expression polymorphism of TF A, which in turn cause an expression polymorphism of gene B between parental lines X and Y.

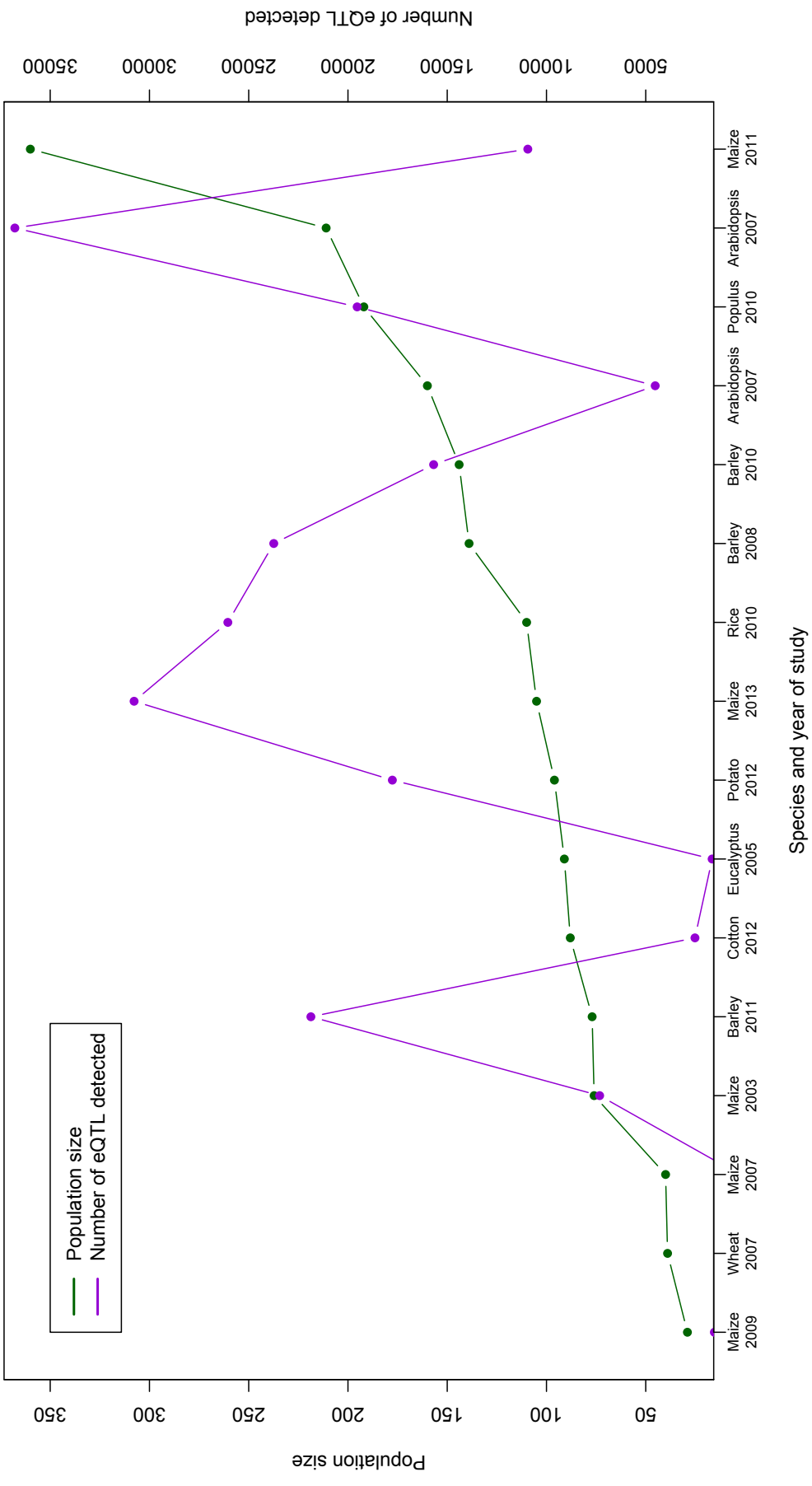


Figure 1.3: Population size (y-axis on the left side) versus number of eQTL detected (y-axis on the right side) in sixteen global eQTL mapping studies on various crops (x-axis), published over the past decade.

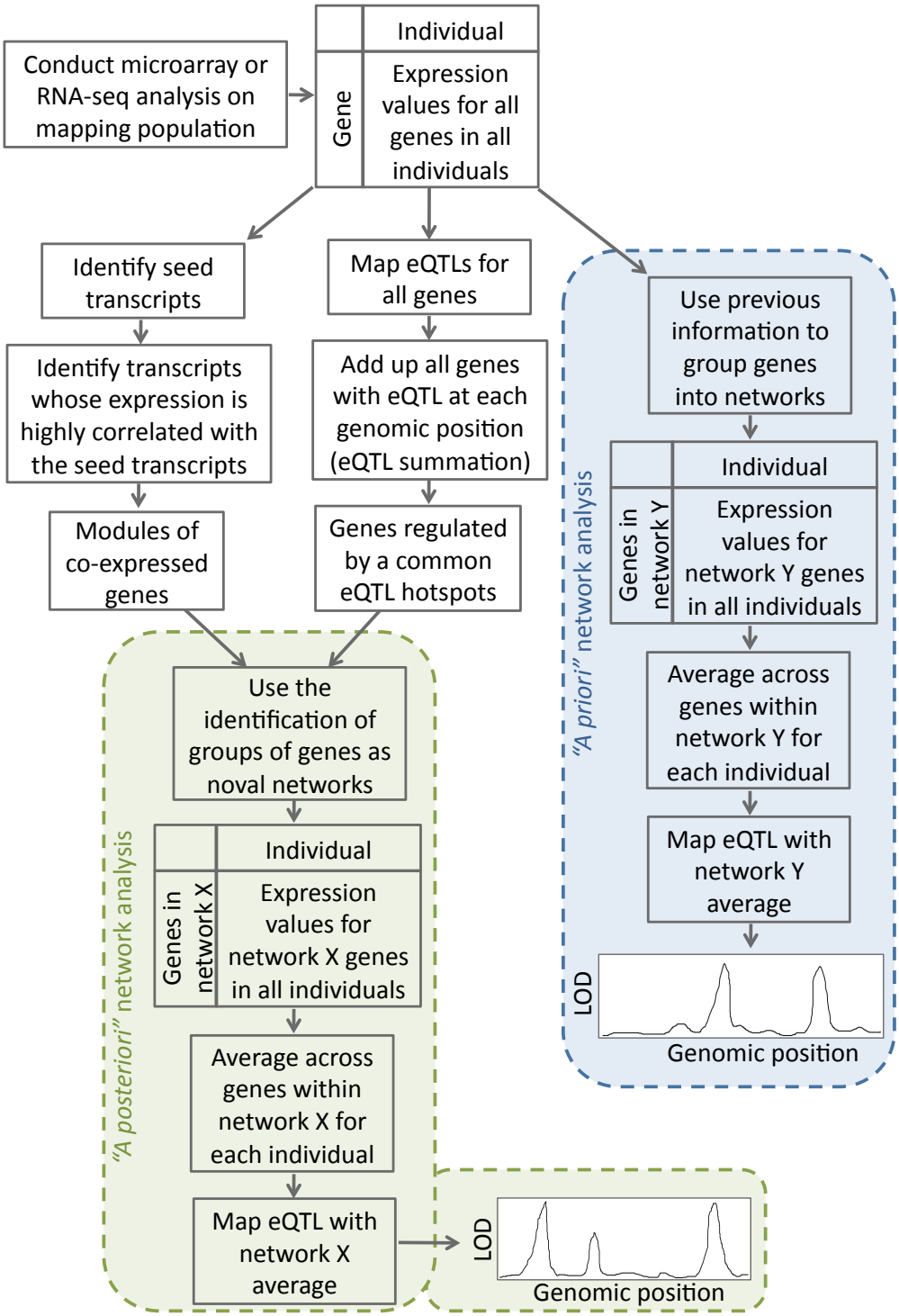


Figure 1.4: Network eQTL analysis. Adapted from Kliebenstein *et al.* (2006). A flowchart to distinguish between the *a priori* and *a posteriori* network analysis approaches.

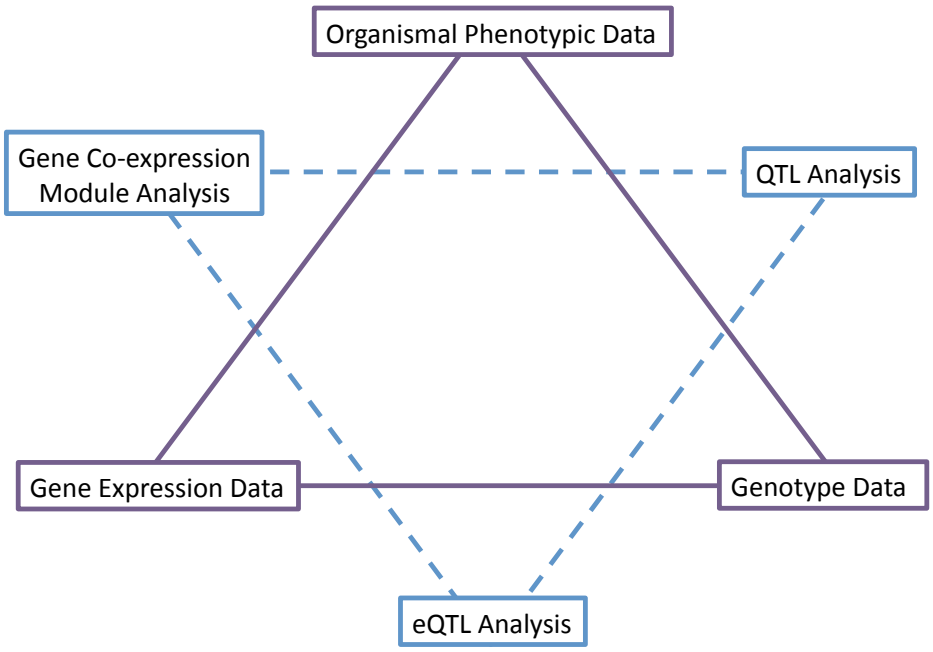


Figure 1.5: Systems genetics approach to dissecting a quantitative phenotypic trait. Adapted from Park *et al.* (2011). Data from organismal phenotype analysis can be integrated with genotype data to map phenotypic QTLs. Organismal phenotypes can also be compared to gene co-expression modules. Gene expression data and genotype data can be used together to map eQTLs. All three datasets can be merged to prioritise mapped genes and identify gene networks associated with organismal phenotypes.

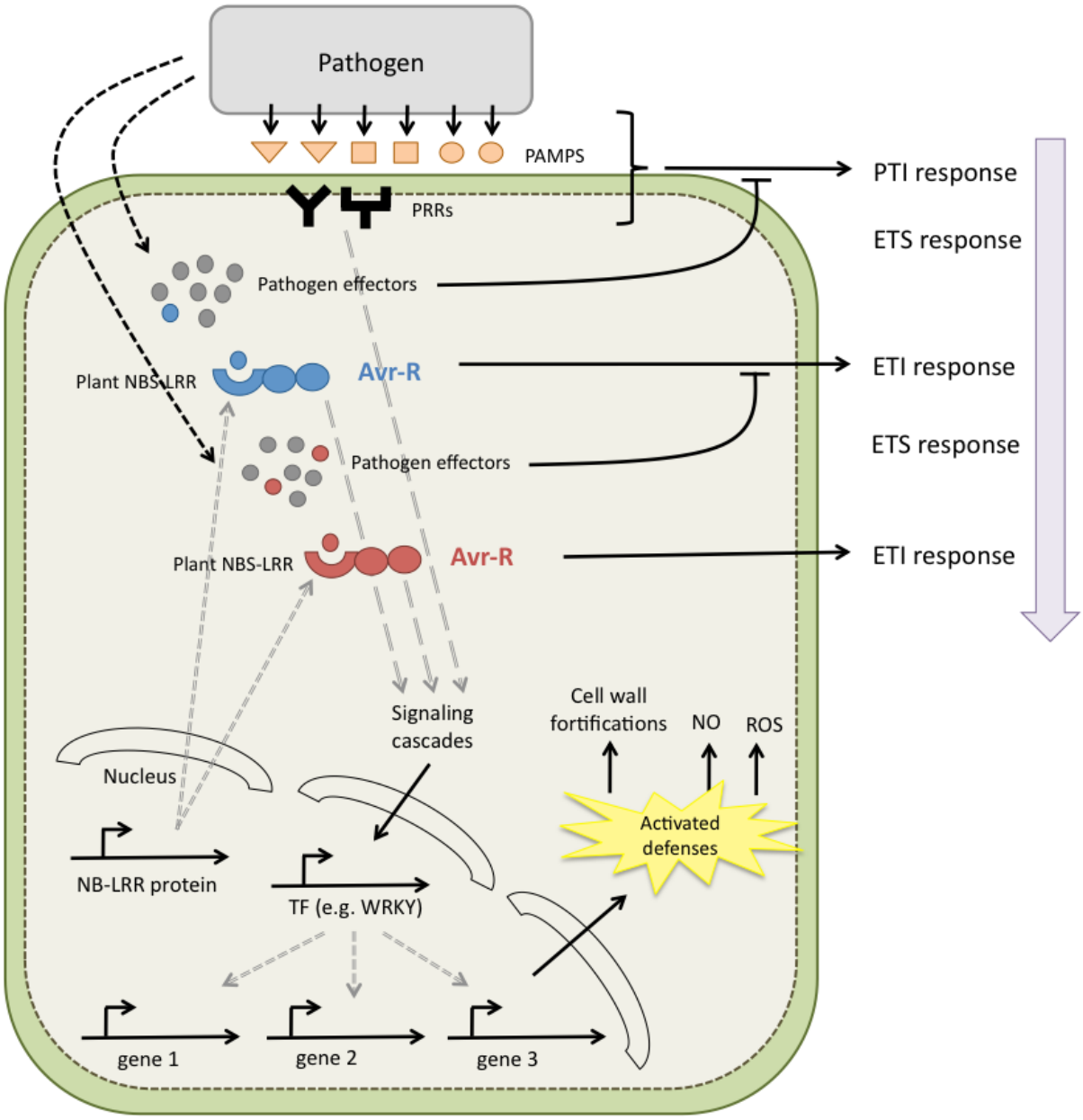


Figure 1.6: Model depicting plant responses to pathogen infection. Adapted from Dangl and Jones (2001) and Abramovitch *et al.* (2006). Plants detect pathogen-associated molecular patterns (PAMPs) via plasma membrane-localised pattern recognition receptors (PRRs) to trigger PAMP-triggered immunity (PTI). Successful pathogens deliver effectors that interfere with PTI, resulting in effector-triggered susceptibility (ETS). When an effector is recognised by an NBS-LRR protein, activation of effector-triggered immunity (ETI) follows. ETI is an amplified version of PTI that triggers specific transcription factors and consequently various signalling cascades. ETI often leads to the induction of hypersensitive response (HR). Phenotypes that are associated with HR include cell wall fortifications and the production of reactive oxygen species (ROS) and nitrogen species (NO). Natural selection drives pathogens to acquire additional effectors that suppress ETI. Selection favours new plant NBS-LRR alleles that can recognise one of the newly acquired effectors, resulting again in ETI.

Table 1.1: Comparison of global eQTL mapping studies on crop species

Publication	Plant species	Parental lines	Population trait	Population type	Population size	Tissue sampled	Marker type
Schadt *et al.* (2003)	Maize	Stiff stalk x Lancaster	None	F3	76	Ear leave	SSRs
Kirst et al. (2005)	Eucalyptus	E. grandis x F1 hybrid	None	Pseudobackcross	91	Differentiating xylem	AFLPs
West et al. (2007)	Arabidopsis	Bay x Sha	None	RIL	211	6 Week old plants	SFPs
Keurentjes *et al.* (2007)	Arabidopsis	Ler x Cvi	None	RIL	160	Seedlings	PCR-based
Shi *et al.* (2007)	Maize	Flint x Flint	Cell wall digestibility	RIL	40	5 Week old stems	NA
Jordan *et al.* (2007)	Wheat	RL4452 × AC Domain	Agronomic and seed quality traits	DH	39	Developing seed	SSRs
Potokina *et al.* (2008)	Barley	Steptoe x Morex	Malting-quality	DH	139	Germinating embryos	SNPs and TDMs
Swanson-Wagner *et al.* (2009)	Maize	B73 x Mo17[a]	Heterosis	Reciprocal hybrids[a]	29	Seedlings	PCR-based
Wang *et al.* (2010)	Rice	Zhenshan 97 x Minghui 63	Seedling vigor traits	RIL	110	Rice shoots	SFPs
Drost *et al.* (2010)	Populus	P. trichocarpa × P. deltoides	None	Pseudobackcross	192	Xylem, Leaf, Roots	SSRs and SFPs
Chen *et al.* (2010)	Barley	Steptoe x Morex	Leaf rust	DH	144	Seedlings	SNPs and TDMs
Moscou *et al.* (2011)	Barley	Q21861 x SM89010	Stem rust (INOC, MOC)	DH	77	Seedlings: inoc, moc	TDMs
Holloway *et al.* (2011)	Maize	B73 x Mo17	None	DH	360	Roots	SNPs
Claverie *et al.* (2012)	Cotton	G. hirsutum x G. barbadense	Fiber developmental stages	RIL	88	2 Fiber development stages	SSRs and AFLPs
Kloosterman *et al.* (2012)	Potato	C x E	None	Diploid backcross	96	Leave and Tuber	SNPs
Li *et al.* (2013)	Maize	B73 x Mo17	None	RIL	105	Shoot apices	SNPs from RNA-seq

[a] RILs were derived from a B73×Mo17 cross. Hybrids were generated by crossing each RIL onto B73 and Mo17. Separate eQTL analyses were conducted within each cross type: B73×RIL, Mo17×RIL, and RIL.

Table 1.2: Comparison of global eQTL mapping studies on crop species (continued)

Publication	Gene expression platform	Number of e-traits	Number of genes with eQTL	Number of eQTL	eQTL method	eQTL threshold	*Cis*-eQTL percentage	Number of *trans*-eQTL hotspots
Schadt *et al.* (2003)	Agilent	18,805	6,481	7,322	Mixture model	LOD >3	34%	NA
Kirst et al. (2005)	cDNA microarray	2,608	1,067	1,655	CIM	LR >11,13,16	NA	18; 5[b]
West et al. (2007)	Affy GeneChip	22,746	15,664	36,781	CIM	LR >12.06	32%	17
Keurentjes *et al.* (2007)	DNA microarray	24,065	4,066	4,523	Linear model	p <5.29e^{-5}	46%	NA
Shi *et al.* (2007)	Forage quality array	439	89	271	IM	LOD >2.4	0%	5
Jordan *et al.* (2007)	Affy GeneChip	1,455	NA	542[c]	CIM	Single high peak[c]	42%	2
Potokina *et al.* (2008)	Affy GeneChip	15,967	12,987	23,738	CIM	LOD >2.87	29-39%	9
Swanson-Wagner *et al.* (2009)	cDNA microarray	NA	1,474; 1,078; 1,128	1,904; 1,334; 1,387	Least squares linear regression	Permutation p-value <0.05	10%	NA
Wang *et al.* (2010)	Affy GeneChip	25,965	16,372	26,051	CIM	LOD >3.12	17%	171
Drost *et al.* (2010)	NimbleGen	NA	30,313; 13,403; 9,137	36,071; 13,403; 9,137	CIM	LOD >2.89; 2.92; 2.93	8-10%	16; 7; 11
Chen *et al.* (2010)	Agilent	15,208	9,557	15,685	Linear model	p <0.001	NA	3
Moscou *et al.* (2011)	Affy GeneChip	22,792	13,919; 15,468	20119, 23616	CIM	LOD >3.138; 3.142	NA	6; 6
Holloway *et al.* (2011)	Agilent	NA	NA	10,941	Haley-Knott regression	KS p-value <1e^{-5}	90%	NA
Claverie *et al.* (2012)	cDNA-AFLP and qRT-PCR	3,263; 1,201	2,220; 803	3,665; 1,375	SMA and CIM	LOD >3.5	NA	21; 12
Kloosterman *et al.* (2012)	Agilent	19,590	14,834	17,764	Haley-Knott regression	LOD >4.35	85%	5
Li *et al.* (2013)	RNA-seq	22,242	19,304	30,774	CIM	LOD >4.17	37%	96

[b] Eighteen hotspots were detected using the F_1 hybrid paternal map, and 5 using the *E. grandis* maternal map.
[c] Reporters with a single highly significant LOD peak, which was at least double the LOD score of any other peak.

Chapter 2

Maize microarray annotation database

Nanette Coetzer[1], Alexander A Myburg[2], D.K. Berger[3*]

[1] Bioinformatics and Computational Biology Unit, Department of Biochemistry, University of Pretoria, Private Bag X20, 0028, South Africa.
[2] Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private Bag X20, 0028, South Africa.
[3] Department of Plant Science, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private Bag X20, 0028, South Africa.

* Corresponding author

Email addresses:
NC: Nanette.Coetzer@gmail.com
AAM: Zander.Myburg@up.ac.za
DKB: Dave.Berger@fabi.up.ac.za

2.1 Note

The content of this chapter has been published in the journal Plant Methods. To be consistent with the thesis layout, the figures and tables are given at the end of the chapter and the references are included in the Bibliography section at the end of the thesis. Additional files that are too large to display are available in the electronic Appendix

as well as from the online version of the manuscript (`http://www.plantmethods.com/content/7/1/31/additional`).

2.2 Authors' contributions

NC built the database, analysed the data, and drafted the manuscript. DKB initiated the study, contributed to the strategy, database design and analysis, and helped to draft the manuscript. ZM contributed to the strategy and database design and helped to edit the manuscript. All authors have read and approved the final manuscript.

2.3 Abstract

2.3.1 Background

Microarray technology has matured over the past fifteen years into a cost-effective solution with established data analysis protocols for global gene expression profiling. The Agilent-016047 maize 44K microarray was custom-designed from EST sequences, but only reporter sequences with EST accession numbers are publicly available. The following information is lacking: (a) reporter - gene model match, (b) number of reporters per gene model, (c) potential for cross hybridisation, (d) sense/antisense orientation of reporters, (e) position of reporter on B73 genome sequence (for eQTL studies), and (f) functional annotations of genes represented by reporters. To address this, we developed a strategy to annotate the Agilent-016047 maize microarray, and built a publicly accessible annotation database.

2.3.2 Description

Genomic annotation of the 42,034 reporters on the Agilent-016047 maize microarray was based on BLASTN results of the 60-mer reporter sequences and their corresponding ESTs against the maize B73 reference genome (RefGen) v2 "Working Gene Set" (WGS) predicted transcripts and the genome sequence. The agreement between the EST, WGS transcript and gDNA BLASTN results were used to assign the reporters into six genomic annotation groups. These annotation groups were: (i) "annotation by sense gene model" (23,668 reporters), (ii) "annotation by antisense gene model" (4,330); (iii) "annotation

by gDNA" without a WGS transcript hit (1,549); (iv) "annotation by EST", in which case the EST from which the reporter was designed, but not the reporter itself, has a WGS transcript hit (3,390); (v) "ambiguous annotation" (2,608); and (vi) "inconclusive annotation" (6,489). Functional annotations of reporters were obtained by BLASTX and Blast2GO analysis of corresponding WGS transcripts against GenBank.

The annotations are available in the Maize Microarray Annotation Database `http://MaizeArrayAnnot.bi.up.ac.za/`, as well as through a GBrowse annotation file that can be uploaded to the MaizeGDB genome browser as a custom track.

The database was used to re-annotate lists of differentially expressed genes reported in case studies of published work using the Agilent-016047 maize microarray. Up to 85% of reporters in each list could be annotated with confidence by a single gene model, however up to 10% of reporters had ambiguous annotations. Overall, more than 57% of reporters gave a measurable signal in tissues as diverse as anthers and leaves.

2.3.3 Conclusions

The Maize Microarray Annotation Database will assist users of the Agilent-016047 maize microarray in (i) refining gene lists for global expression analysis, and (ii) confirming the annotation of candidate genes before functional studies.

2.4 Background

Currently, there are several maize microarray platforms available, including an Affymetrix short oligonucleotide array (Kirst *et al.*, 2006), a Nimblegen 50-mer array (Sekhon *et al.*, 2011), a 70-mer array from the University of Arizona Maize Oligonucleotide Array project (Galbraith and Edwards, 2010) and the 60-mer Agilent-016047 Maize 4×44K microarray (Wang *et al.*, 2010*a*).

The Agilent microarray platform (`http://www.agilent.com`) is a mature technology that yields high quality gene expression data, which can be readily analyzed using established statistical tools (Coetzer *et al.*, 2010). The Agilent-016047 Maize 4×44K microarray was custom-designed by the Walbot laboratory, with 42,034 in situ synthesised 60-mer oligonucleotide reporters (excluding controls) (Ma *et al.*, 2008). Currently, the Agilent "e-array" tool (`https://earray.chem.agilent.com/earray/`) only provides the

60-mer sequences and expressed sequence tag (EST) accession numbers from which the reporters were designed, without detailed or up-to-date annotations. There was therefore a need to develop a strategy for annotation, and thereby build a database of annotations for this maize microarray, as well as similar custom arrays.

The maize B73 genome sequence was released in November 2009 (Schnable *et al.*, 2009), and this provided the opportunity to locate the reporters on the genome sequence, and provide functional annotations. Each reporter is intended to report the expression of a single gene unambiguously. However, since the reporters were designed from ESTs from different maize lines before a reference genome sequence was available (Ma *et al.*, 2008), redundancy on the array, as well as imperfect reporter matches were expected.

Version 1 and Version 2 Agilent 22K arrays (Ma *et al.*, 2006, 2007) were precursors for the 44K Agilent-016047 array. Version 1 was designed from the December 2003 maize EST assembly of MaizeGDB and was made up of 21,782 reporters. More than 80% of these reporters were also included in Version 2 plus $\sim 3,000$ new reporters, designed from maize sequences in GenBank. Of the 20,963 gene features on Version 2, $\sim 13,000$ were sense strand reporters and $\sim 5,000$ antisense strand reporters.

For the Agilent-016047 44K array, an updated set of 60-mer reporters were designed using Picky 2.0 (Chou *et al.*, 2004). The reporter set mainly consists of validated reporters from the two precursor maize arrays described above and validated reporters from anther expressed genes detected using a spotted 70-mer array format (containing reporters to about 35,000 maize genes) (Ma *et al.*, 2006) (`http://www.maizearray.org`). Additional gene reporters were based on release 16.0 of the TIGR Maize Gene Index as well as cDNA or EST sequences from GenBank (that were at the time not yet in the TIGR Maize Gene Index assembly) (Ma *et al.*, 2008). According to Ma *et al.* (2008), the 42,034 maize gene reporters represent $\sim$ 39,000 sense transcripts including a subset of genes with multiple reporters, and ~ 500 antisense transcripts. In addition to the 42,034 maize gene reporters, the array also contains internal quantitative "spike-in" controls of non-maize sequences, which were not annotated in this study.

The aims of this study were to annotate the reporter set of the Agilent-016047 microarray by: (i) locating each reporter on the maize B73 genome sequence; (ii) associating each reporter to the transcript of a single gene, if possible; and (iii) assigning functional annotations to the gene represented by each reporter. Our results revealed that we

could not associate all of the reporters with a single transcript with high confidence, and therefore we built a database `http://MaizeArrayAnnot.bi.up.ac.za/`, which provides confidence scores of the genomic positions and functional annotations of reporters on the Agilent-016047 Maize array. Our annotation strategy provides guidelines for annotation of custom-designed microarray slides where partial EST information is available, and this resource will therefore be useful to maize researchers, and other researchers using custom arrays.

2.5 Construction and content

2.5.1 Data sources

The gene list for the Agilent Maize Gene Expression Microarray 4 × 44K (design ID 016047) was downloaded from Agilent's eArray tool (`https://earray.chem.agilent.com/earray/`) containing a reporter ID, a 60-mer reporter sequence and an EST accession number for each of the 42,034 reporters on the microarray. EST sequence information for 34% of the reporters was available on GenBank, and BioPython (Cock *et al.*, 2009) was used to extract sequence and other relevant information from individual GenBank files. For an additional 31% of the reporters, EST sequences were obtained from the Walbot laboratory. For the remaining 35% of the reporters, no EST sequences were available, since these are likely to be derived from proprietary sources. The cDNA sequences (in FASTA format), their transcript start and end positions on the B73 RefGen v2 genome sequence as well as InterPro and GO annotations for genes, were downloaded from the maizesequence.org file transfer protocol (FTP) site (`http://ftp.maizesequence.org/current/`). Only the protein coding transcripts in the B73 RefGen v2 Working Gene Set (WGS) were used (88,611 cDNAs representing 63,331 genes). We chose to use the WGS and not the Filtered Gene Set (FGS) since it was more inclusive of transcripts that could have been used in the reporter design. The FGS (63,540 transcripts; 39,656 genes) is a subset of the WGS in which transcripts that are "probable pseudogene", "possible transposon", "contamination" or "low confidence" have been filtered out.

The maize B73 RefGen v2 genome sequence (sequences of all 10 chromosomes, in FASTA format) was downloaded from the maizesequence.org FTP site (`http://ftp.maizesequence.org/current/`). Lastly, the maize core bin markers (Wei *et al.*, 2009*b*)

and corresponding B73 RefGen v2 base pair positions were retrieved from MaizeGDB (Sen *et al.*, 2009). All sets of data were downloaded in December 2010/January 2011.

2.5.2 Genomic annotation

Figure 2.1 outlines the strategy that was followed to obtain genomic annotations for each reporter on the Agilent-016047 microarray. All nucleotide sequences were searched against target datasets using the BLASTN algorithm version 2.2.18 (Altschul *et al.*, 1990). For BLASTN searches of the 60-mer reporter sequences against ESTs, the WGS transcripts and genomic DNA (gDNA) (B73 RefGen v2), the word size parameter was set to 23 and gaps were not allowed. This cut-off was chosen based on a study that showed that matches of $\geq$ 23 contiguous nucleotides yielded hybridisation signals under stringent conditions in more than 90% of a set of Agilent reporters (Poulsen *et al.*, 2008). Thus, the identity out of 60, rather than the E-value, was used as the measure of similarity for BLASTN searches with the reporters. We also carried out BLASTN searches with EST sequences, and in these cases E-values were used. All BLAST results were stored in a relational database. The parameters used for the BLAST searches are shown in Table 2.1.

Three sets of BLASTN hits were stored for each reporter (Figure 2.1). Firstly, BLASTN of the reporter was implemented against the genome sequence (B73 RefGen v2), and the top BLASTN hit (if $\geq$ 23 contiguous matches), or multiple BLASTN hits (if $\geq$ 23 contiguous matches and identity $\geq$ 55/60) was called the "reporter-gDNA result". The reporters were also searched against the genome sequence using "exonerate" (Slater and Birney, 2005) to detect reporters that spanned introns. The parameters for exonerate are shown in Table 2.2. In cases where reporters had positive exonerate matches (and 23 contiguous matches) to the genome sequence, this result was recorded as the "reporter-gDNA result". Secondly, BLASTN of the reporter was implemented against the WGS transcripts, and the top BLASTN hit (if $\geq$ 23 contiguous matches), or multiple BLASTN hits (if $\geq$ 23 contiguous matches and identity $\geq$ 55/60) was called a "reporter-WGS transcript result". Thirdly, after confirming that the reporter matched its corresponding EST listed in the Agilent eArray database ($\geq$ 23 contiguous matches), the top BLASTN hit (if E-value $\leq 1e^{-10}$), or multiple BLASTN hits (if E-value $\leq 1e^{-10}$) of the EST against the WGS transcript dataset was called an "EST-WGS transcript result" (BLASTN parameters in Table 2.1). These BLASTN cut-offs were selected based on a previous study

where the same cut-offs were used to align ESTs to predicted maize cDNAs (Emrich *et al.*, 2007).

The ESTs were also searched against the gDNA using exonerate (parameters in Table 2.2), and matches with a normalised score of at least 3 (calculated by dividing the exonerate raw score by the query EST sequence length) (Donmez *et al.*, 2009) were recorded as the "EST-gDNA result".

The next step was to determine if there was agreement between the reporter-gDNA result and the reporter-WGS transcript result, in other words whether the WGS transcript that the reporter matched was derived from the same gDNA position that the reporter matched. This was recorded as the reporter-gDNA/WGS agreement result (Figure 2.1). Similarly, we tested whether the EST-gDNA result and EST-WGS transcript result were in agreement, and recorded this as the EST-gDNA/WGS agreement result (Figure 2.1).

Finally, the reporters were placed into one of six annotation groups, informed by sequence matching and agreement results described above and whether the genomic position of the reporter overlapped with one or more gene models in the sense or antisense direction (Figure 2.1). The annotation groups were:

(i) Annotated by sense gene model: Reporters that match a WGS transcript and genomic location of the same gene model (single reporter-gDNA/WGS agreement result);

(ii) Annotation by antisense gene model: Reporters that match a transcript and genomic location of the same gene model, but align to the antisense direction of the transcript (single reporter-gDNA/WGS agreement result);

(iii) Annotation by gDNA: Reporters that match a unique location on the maize B73 genome, but this location is not currently annotated as a gene model (single genomic result);

(iv) Annotation by EST: Reporters that do not match a WGS transcript, but that are derived from an EST that matches a WGS transcript and its genomic location (single EST-gDNA/WGS agreement result);

(v) Ambiguous annotation: Reporters with more than one sense gene model, antisense gene model or EST result (More than one reporter-gDNA/WGS transcript agreement or EST-gDNA/WGS transcript agreement result);

(vi) Inconclusive annotation: Reporters that match more than one transcript, but not the genomic location of the corresponding gene models. Reporters that match more than

one genomic location, but no corresponding transcripts. Reporters with no valid hits.

2.5.3 Functional annotation

The "reporter-WGS transcript result" for each reporter (described above) was used to assign a functional annotation to each reporter. The functional annotations for each of the 88,611 cDNA sequences in the WGS of the B73 RefGen v2 genome sequence were obtained by BLASTX (Altschul *et al.*, 1990) searches (with default parameter settings; Table 2.1) against the National Center for Biotechnology Information (NCBI) non-redundant peptide database (nr). The top three hits (and corresponding statistics) were stored in a relational database. Blast2GO (Conesa *et al.*, 2005) was used to associate each WGS transcript (and therefore the corresponding reporters) with GO terms, using default settings.

2.5.4 Database and web interface

The Maize Microarray Annotation Database interface was written using Turbogears (Ramm *et al.*, 2006), a Python web application framework. A central MySQL database is used to store sequence and annotation information. SQLAlchemy (Copeland, 2008), an object relational mapper for Python and toolkit for SQL, is implemented within the Maize Microarray Annotation Database when a user queries the database.

2.5.5 Integration with the MaizeGDB genome browser

An annotation file with the genomic positions for each reporter that could be matched to the genome was generated (Additional file 3 available in the electronic Appendix). This can be uploaded to the MaizeGDB genome browser (Sen *et al.*, 2010) and viewed as an annotation track in the context of the B73 RefGen v2 genome sequence.

2.5.6 Reporters with expression in maize leaf material

Reporters "with measurable signal" were identified as those with a signal to noise ratio (SNR) > 3 in at least one of fifty Agilent-016047 microarrays hybridised with cDNA from maize leaves of a segregating population (data not shown).

2.6 Utility and Discussion

2.6.1 Genomic annotation groups

Table 2.3 shows the breakdown of the annotation groups of the 42,034 reporters on the maize Agilent-016047 microarray, as determined by our strategy outlined in Figure 2.1. Importantly, 27,998 reporters (67%) were annotated by gene model in the sense or antisense direction, which means that they correspond to a transcript with a defined gDNA position. Approximately half of the reporters in this group mapped to UTR regions, and the rest to coding regions. A number of these reporters (1,554) were shown using exonerate software (Slater and Birney, 2005) to span introns.

The reporters annotated by gene model in the sense or antisense direction represent 46.7% of the genes in the B73 maize FGS. Within this group, there were 4,330 reporters that aligned to the antisense direction of a gene model (Table 2.3). Natural antisense transcripts (NATs) contain sequences complementary to the sense transcripts of protein-coding genes (Jin *et al.*, 2008). Between 7 and 30% of genes in animal and plant genomes encode overlapping *cis*-NATs (Jin *et al.*, 2008). Many NATs are conserved, implying regulatory functions for these transcripts in gene expression. According to Ma *et al.* (2006), 14.3% of the pollen transcriptome consists of detectable antisense transcripts. It should be borne in mind that, in some cases, a reporter with an antisense annotation could in fact correspond to a sense transcript if the EST from which it was designed was incorrectly oriented or the genomic annotation was in the wrong strand. These errors are expected to be corrected in future annotation versions of the maize B73 genome sequence.

The 1,549 reporters "annotated by gDNA" (Table 2.3) represent reporters that had significant matches to the maize genome sequence but did not match current transcripts in the WGS of the maize B73 RefGen v2. These reporters will possibly be linked to gene models in future versions of the B73 genome due to improvements in gene prediction algorithms or availability of RNA-seq data from different tissues of maize B73 plants. Although these reporters are not currently associated with functional annotations, their placement on the B73 genome sequence is useful for eQTL studies in maize. This category of reporters showed a lower proportion (64%) with measurable expression in maize leaves compared to reporters annotated by sense gene model (88%; Table 2.3).

The 3,390 reporters "annotated by EST" (Table 2.3) were derived from ESTs that

showed sequence similarity to a WGS transcript from B73 (E-value $\leq 1e^{-10}$), however the reporter itself did not have a significant hit to the WGS transcript. The reporters on the maize Agilent-016047 microarray were designed from ESTs from various maize lines (Figure 2.2). Reporters "annotated by EST" are most likely derived from maize lines other than B73, although the source is not known for all reporters since this information could be retrieved for only 34% of the reporters (Figure 2.2 (a)). The region of the transcript that corresponds to the reporter is therefore predicted to be divergent between B73 and the line from which the reporter was derived.

The 2,608 "ambiguous" reporters (Table 2.3) each represent more than one gene model, which are mostly members of the same gene family. Interpretation of expression data from these reporters should be done with caution, as it is possible that the signal is due to cross hybridisation from more than one family member. As an example, reporter A_92_P037799 represents four members of the cytochrome P450 gene family on chromosomes 2, 3, 6 and 8, as shown in the multiple sequence alignment (Figure 2.3).

There were 6,489 reporters with "inconclusive annotation" (Table 2.3), and thus interpretation of expression data from these reporters should be made with caution. This group contained a relatively low proportion of reporters with signal in a maize leaf microarray experiment conducted in our laboratory, namely 61% compared to 88% of reporters "annotated by sense gene model" (Table 2.3). Re-sequencing of six maize lines from China identified several hundred genes that were not present in B73, but could be annotated as plant proteins (Lai *et al.*, 2010), and therefore it is possible that reporters with "inconclusive annotation" may represent transcribed genes from other maize lines. A subset of the reporters with inconclusive annotation and no hits against the B73 genome sequence had EST sequences available (1,727) and these were searched against GenBank using BLASTX. Only 892 had significant hits (E-value $\leq 1e^{-10}$) and 553 matched plant proteins.

Prior to our work, the reporters on the Agilent-016047 maize array could be visualised in the context of the B73 maize genome sequence at MaizeGDB (`http://gbrowse.maizegdb.org/cgi-bin/gbrowse/maize_v2/`) based on the Walbot laboratory annotations. However, there are several limitations of this annotation track, namely: (i) the positions given are based on RefGen v1, whereas the sequence is RefGen v2; (ii) the positions are based on MegaBLAST hits to the gDNA, but no matches to transcripts

are given; (iii) the reporters are named using a unique identifier (UID) which is different from the Agilent e-array ID; and (iv) three confidence categories are given, however some reporters have up to 500 hits. Therefore we have produced an updated annotation track that is compatible with MaizeGDB (Additional file 3 available in the electronic Appendix) that reports the positions of all reporters on the array except those with inconclusive annotation or annotation by EST. An example of three reporters that match one gene model is shown in Figure 2.4.

2.6.2 Maize Microarray Annotation Database

The Maize Microarray Annotation Database has an interactive web interface `http://MaizeArrayAnnot.bi.up.ac.za/`, providing the user with three main functionalities namely "Search Agilent slide", "BLAST sequences" and "Get sequences from GenBank" (Figure 2.5). Most users are likely to use the "search Agilent slide" function, since they would be interested in downloading annotations for a list of reporters that are differentially expressed in a microarray experiment. In order to search the Agilent slide, the user can provide Reporter IDs, EST Accession numbers or gene names. The outputs from a query are reporter information, EST information, gene information, genomic and functional annotation information as well as the evidence for the annotation results. The following can be downloaded: DNA sequences (reporter, EST or WGS transcript sequences in FASTA format), a table with all annotation information, and/or multiple sequence alignments. Searching by WGS gene name makes it possible to see whether there is more than one reporter for a gene. On average, there are ~ 1.6 reporters per gene. Users can also retrieve nucleotide sequences by submitting GenBank accession numbers for ESTs, or BLAST sequence(s) against the Agilent slide to identify which reporters represent the query sequence best.

2.6.3 Case studies

Table 2.4 gives a selection of five publications in which the Agilent-016047 array has been used, with an indication of how many reporters gave a measurable signal according to the authors. Lack of a signal may be due to tissue specific expression, genotype differences, or poor reporter design. Ma *et al.* (2008) used this microarray to study the expression profiles of maize anther and pollen ontogeny. They found that more than

24,000 different transcript types were expressed, and that each anther stage expressed $\sim 10,000$ constitutive and $\sim 10,000$ or more transcripts restricted to one or a few stages in anther development. Casati and Walbot (2008) measured transcriptome changes between RNA interference (RNAi) transgenic maize lines and a ultraviolet B (UV-B) tolerant B73 control line, using this Agilent slide. Approximately 26,000 reporters showed expression in adult maize leaves. Skibbe *et al.* (2009) hypothesised that Mutator transposon activity reprograms the transcriptomes of developing maize anthers. About 35,000 reporters had signals > 2.6 times the standard deviation of the background (i.e. 99.5% confidence interval), and they concluded that Mu transposition activated by transcriptionally active *MuDR* results in a 25% change in the transcriptome. Wang *et al.* (2010*a*) hypothesised that the male sterile 8 mutation (ms8) of maize disrupts the temporal progression of the transcriptome. They found that fertile anthers exhibit an unexpectedly high transcript complexity; there were 27,400 constitutively expressed transcripts, 2,143 stage-specific transcripts and 2,484 transcripts that were expressed at two stages, giving $\sim 32,000$ transcripts in total that were expressed over a 90-h period. Lastly, Rajhi *et al.* (2011) used this array and laser microdissection to identify transcripts expressed in maize root cortical cells during lysigenous aerenchyma formation.

We analysed expression data from hybridisation of maize leaf cDNA from a segregating population to fifty Agilent-016047 arrays to assess the number of reporters with measurable signal in our hands. The data showed that $\sim 32,000$ reporters had a consistent signal to noise ratio (SNR) greater than 3, whereas $\sim 10,000$ reporters were deemed non-hybridising to leaf transcripts (Table 2.4). These six studies demonstrate that in all cases a large proportion of the reporters on the Agilent-016047 arrays give measurable signals in tissues as diverse as anthers, leaves and roots.

The questions are, however, how many genes are represented by these reporters and how much confidence is there in their annotations? To address these questions, we extracted tables of differentially expressed reporters reported in these studies and annotated the reporters using our Maize Microarray Annotation Database (Table 2.5). Most of the data tables have the majority of reporters annotated with high confidence by a single sense or antisense gene model ($59-86\%$ of reporters in each data table, Table 2.5). However, $3-9\%$ of reporters have ambiguous annotations, and thus their hybridization signals could be due to cross-hybridisation between gene family members (Table 2.5). This is

of particular relevance in the data table S4 from Ma *et al.* (2008) which was a selection of reporters corresponding to Zinc finger-related proteins, where 9% of reporters were "ambiguous". Each data table contained reporters with inconclusive annotations. The data table which appears to be the exception is the study of gene expression in anthers of the *ms8* mutant in which only 38% of reporters were annotated by sense gene model, and this table had a higher proportion of antisense, EST and inconclusive annotation reporters (14%, 13% and 25%, respectively). This may reflect a difference in the biology of this experiment compared to the other experiments.

We suggest that annotation of reporters with the Maize Microarray Annotation Database can be useful for refining lists of "differentially expressed" reporters for subsequent global analyses (e.g. GO enrichment using tools such as MADIBA (Law *et al.*, 2008)). In addition, the database is also essential to confirm the annotation of candidate genes identified from a microarray experiment before detailed functional analyses (e.g. gene knockouts) are carried out. To this end, we have provided, as Additional files 8, 9, 10, 11, 12, 13, 14 and 15 (available in the electronic Appendix), our annotations of the data tables from the case studies listed in Table 2.5.

The importance of correct annotation of microarrays is illustrated by the study of Gertz *et al.* (2009) who performed a similar analysis on the 44K Agilent human expression arrays and found that many reporters had inconclusive annotations. Out of 42,683 reporters, 25,505 (60%) were considered "fully valid" according to their analyses. In another study, an Agilent mouse 44K array was re-annotated resulting in improved annotations for more than 10,000 reporters on the array (Gaj *et al.*, 2007). Furthermore, gene models are constantly being updated as new experimental and annotation data accumulates. Therefore re-annotation of reporters is required as illustrated by a study in which a dozen mammalian GeneChip arrays were re-annotated (Dai *et al.*, 2005). This would be of particular importance in maize where the genome sequence was recently released (Schnable *et al.*, 2009) and is currently only at version 2 of annotation.

2.7 Conclusions

A reporter-by-reporter validation of the 4 × 44 K Agilent-016047 maize microarray was performed. In total, 71% of the reporters correspond to a transcript with a defined

gDNA position and represent 46.7% of the genes in the B73 FGS. All results have been included in a database `http://MaizeArrayAnnot.bi.up.ac.za/`, which provides confidence scores of the genomic positions and functional annotations of reporters on the Agilent-016047 Maize array. The database facilitates interpretation of maize gene expression data. Scientists embarking on expression profiling in maize are likely to find this array an attractive option, since the combination of our annotation database with established analysis methods (Smyth, 2004) facilitates data interpretation. In addition, our strategy can be applied when annotating any custom-designed array from a species for which the genome sequence is available.

2.8 Availability and requirements

The Maize Microarray Annotation Database is publicly available at `http://MaizeArrayAnnot.bi.up.ac.za/`.

2.9 Acknowledgements

We thank the Walbot laboratory for kindly providing EST sequence information and their initial annotations. We thank an anonymous reviewer for constructive comment. The financial assistance of the Technology Innovation Agency (TIA)(South Africa) and the National Research Foundation (NRF)(South Africa) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to TIA or the NRF.

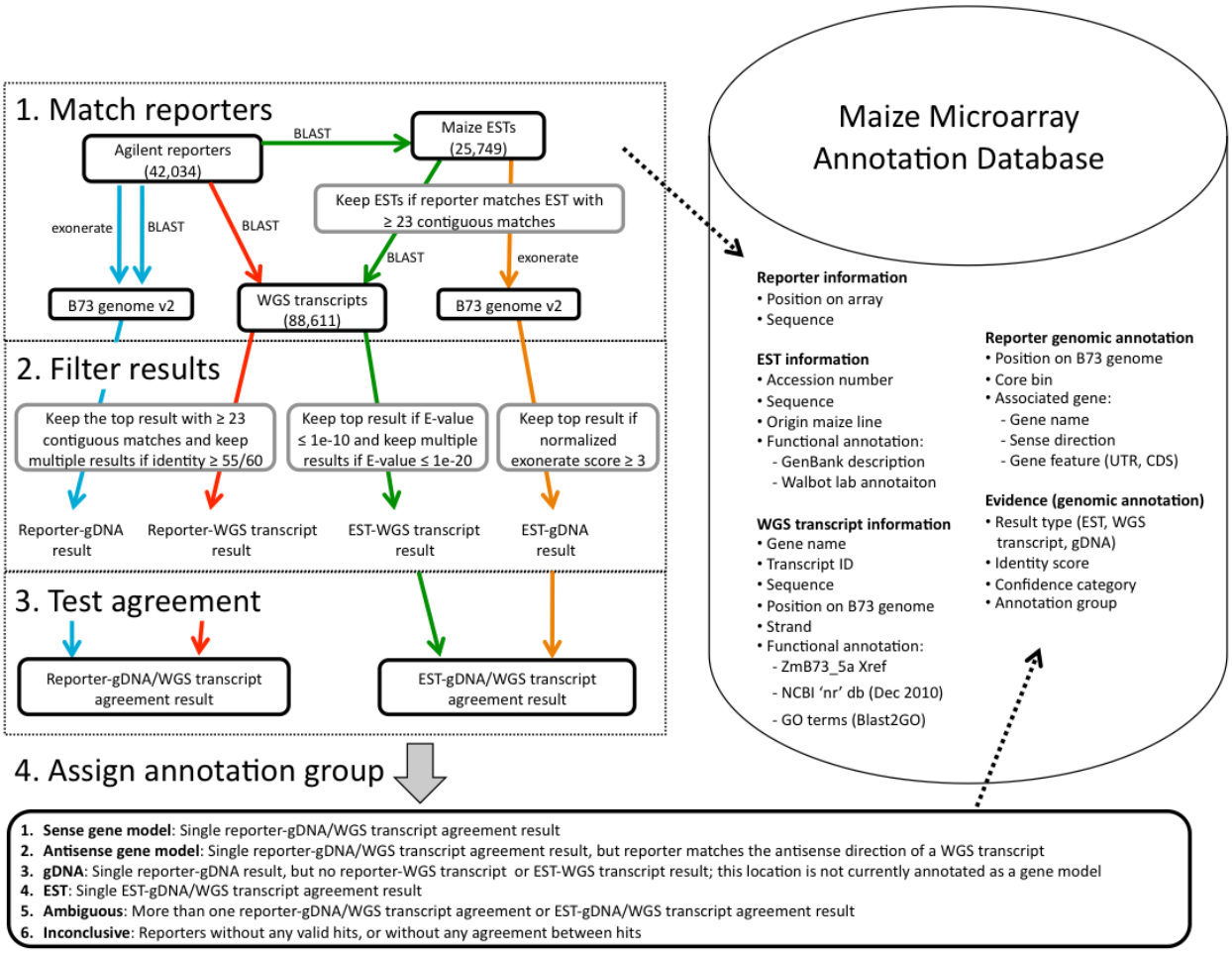


Figure 2.1: Strategy followed to assign genomic and functional annotations to the reporters on the Agilent-016047 maize microarray. Using BLASTN and exonerate software, the 42,034 60-mer reporters were matched to available EST sequences, the maize B73 RefGen v2 genome and the WGS predicted transcripts. BLASTN and exonerate results were filtered and compared to test agreement between EST, WGS transcript and gDNA hits. Based on the agreement analysis, one of six genomic annotation groups was assigned for each reporter. Functional annotations of reporters were based on the functional annotations of their corresponding WGS transcripts. The data has been made accessible from the Maize Microarray Annotation Database `http://MaizeArrayAnnot.bi.up.ac.za/`.

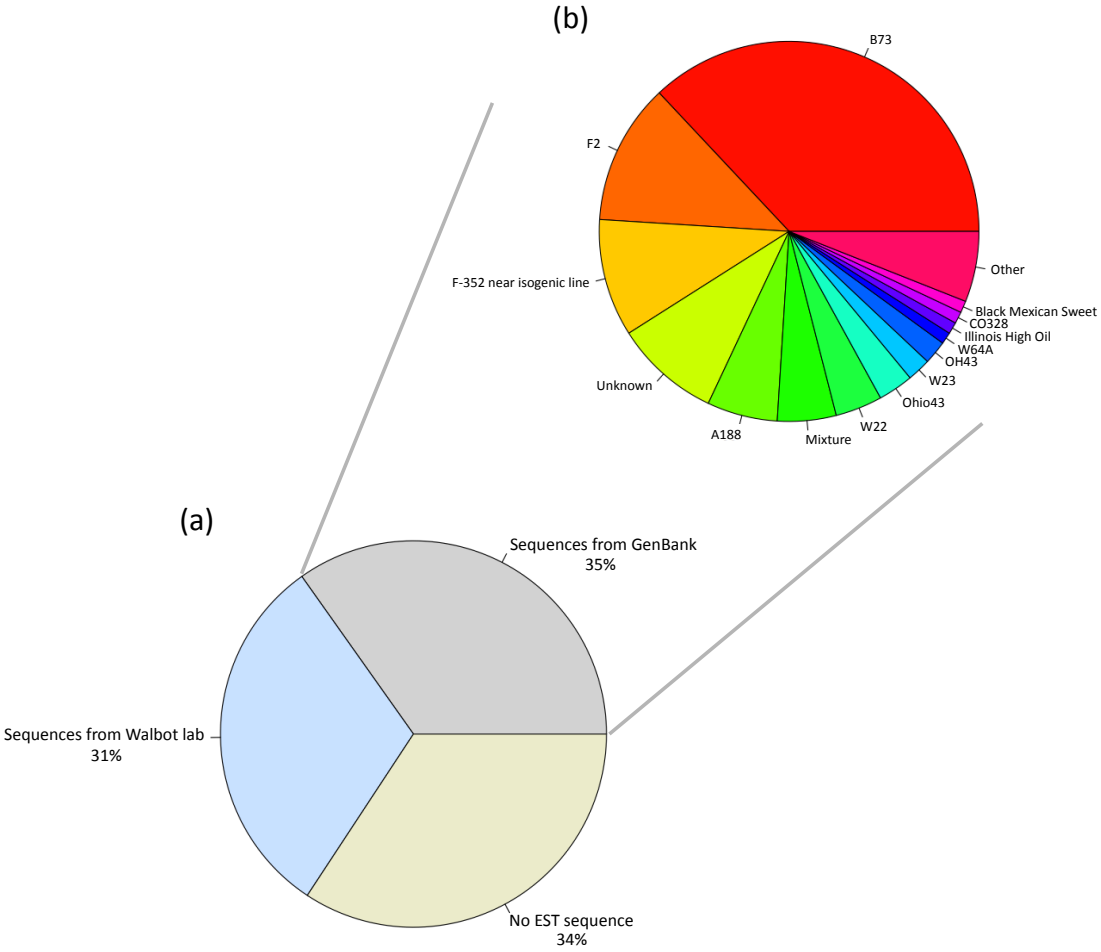


Figure 2.2: Sources of maize ESTs. **(a)** ESTs (39, 174) from which reporters on the Agilent-016047 microarray were designed. **(b)** Sources of ESTs with GenBank annotations (13, 640).

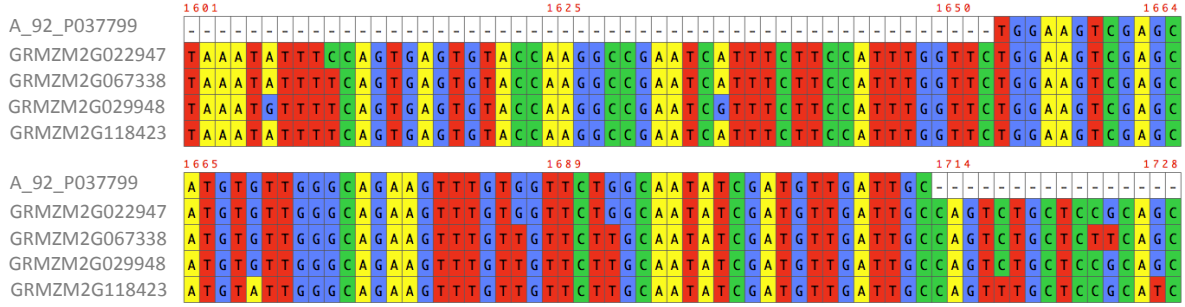


Figure 2.3: Example of a reporter in the "ambiguous" annotation group. Multiple alignment of reporter A_92_P037799 with corresponding parts of four maize cytochrome P450 cDNAs.

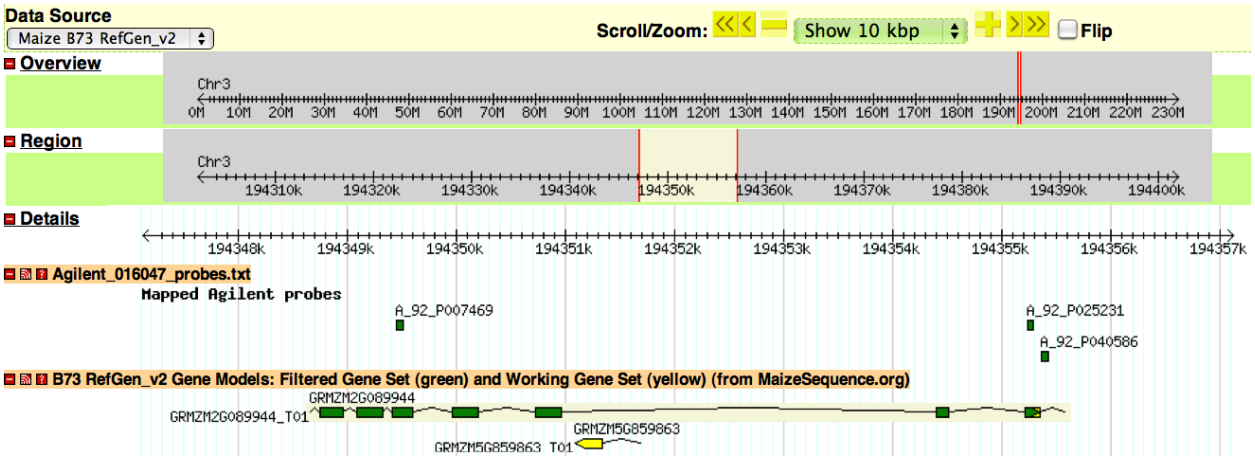
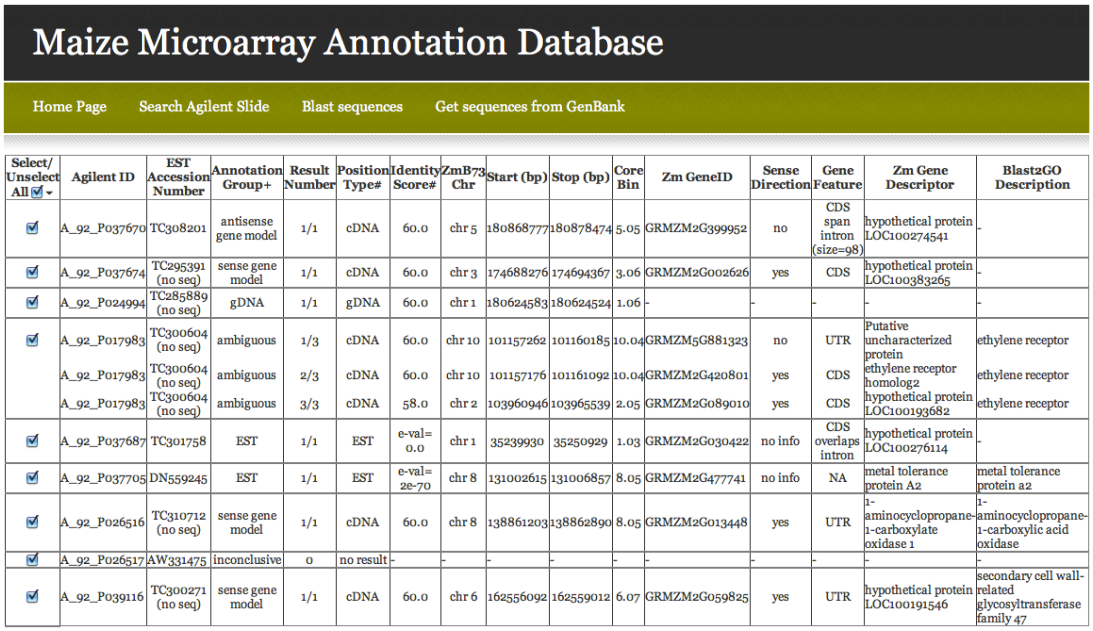


Figure 2.4: Screenshot of the B73 RefGen v2 genome browser at MaizeGDB. Three Agilent reporters (A_92_P007469, A_92_P025231, A_92_P040586) are linked to gene model GRMZM2G089944 on chromosome 3.



Maize Microarray Annotation Database

Home Page | Search Agilent Slide | Blast sequences | Get sequences from GenBank

Select/ Unselect All	Agilent ID	EST Accession Number	Annotation Group+	Result Number	Position Type#	Identity Score#	ZmB73 Chr	Start (bp)	Stop (bp)	Core Bin	Zm GeneID	Sense Direction	Gene Feature	Zm Gene Descriptor	Blast2GO Description
☑	A_92_P037670	TC308201	antisense gene model	1/1	cDNA	60.0	chr 5	180868777	180878474	5.05	GRMZM2G399952	no	CDS span intron (size=98)	hypothetical protein LOC100274541	-
☑	A_92_P037674	TC295391 (no seq)	sense gene model	1/1	cDNA	60.0	chr 3	174688276	174694367	3.06	GRMZM2G002626	yes	CDS	hypothetical protein LOC100383265	-
☑	A_92_P024994	TC285889 (no seq)	gDNA	1/1	gDNA	60.0	chr 1	180624583	180624524	1.06	-	-	-	-	-
☑	A_92_P017983	TC300604 (no seq)	ambiguous	1/3	cDNA	60.0	chr 10	101157262	101160185	10.04	GRMZM5G881323	no	UTR	Putative uncharacterized protein	ethylene receptor
	A_92_P017983	TC300604 (no seq)	ambiguous	2/3	cDNA	60.0	chr 10	101157176	101161092	10.04	GRMZM2G420801	yes	CDS	ethylene receptor homolog2	ethylene receptor
	A_92_P017983	TC300604 (no seq)	ambiguous	3/3	cDNA	58.0	chr 2	103960946	103965539	2.05	GRMZM2G089010	yes	CDS	hypothetical protein LOC100193682	ethylene receptor
☑	A_92_P037687	TC301758	EST	1/1	EST	e-val= 0.0	chr 1	35239930	35250929	1.03	GRMZM2G030422	no info	CDS overlaps intron	hypothetical protein LOC100276114	-
☑	A_92_P037705	DN559245	EST	1/1	EST	e-val= 2e-70	chr 8	131002615	131006857	8.05	GRMZM2G477741	no info	NA	metal tolerance protein A2	metal tolerance protein a2
☑	A_92_P026516	TC310712 (no seq)	sense gene model	1/1	cDNA	60.0	chr 8	138861203	138862890	8.05	GRMZM2G013448	yes	UTR	1-aminocyclopropane-1-carboxylate oxidase 1	1-aminocyclopropane-1-carboxylic acid oxidase
☑	A_92_P026517	AW331475	inconclusive	0	no result	-	-	-	-	-	-	-	-	-	-
☑	A_92_P039116	TC300271 (no seq)	sense gene model	1/1	cDNA	60.0	chr 6	162556092	162559012	6.07	GRMZM2G059825	yes	UTR	hypothetical protein LOC100191546	secondary cell wall-related glycosyltransferase family 47

Figure 2.5: Screenshot of the Maize Microarray Annotation Database. The Maize Microarray Annotation Database enables users to retrieve reporter-specific and global information regarding the reporters on the Agilent-016047 microarray.

Table 2.1: BLAST parameters used for annotation of the Agilent 016047 maize microarray.

BLAST parameter	reporters vs (i) ESTs; (ii) WGS transcripts; (iii) gDNA	ESTs vs WGS transcripts	WGS transcripts vs GenBank (functional annotation)
Program name	BLASTN	BLASTN	BLASTX
Word size	23	11 (default)	3 (default)
Filter query sequence for low-complexity subsequences	TRUE (default)	TRUE (default)	TRUE (default)
Gapped alignment	FALSE	TRUE (default)	TRUE (default)
Initial penalty for opening a gap	NA (no gaps allowed)	5 (default)	11 (default)
Penalty for each gap character	NA (no gaps allowed)	2 (default)	1 (default)
Threshold expectation value for keeping alignments	10 (default); own filtering on identity out of 60	10^{-10}	10^{-10}
Protein similarity matrix	Not used in BLASTN	Not used in BLASTN	BLOSUM62 (default)

Table 2.2: Parameters used for exonerate analysis of the Agilent 016047 maize microarray reporters and ESTs against the B73 maize genome sequence.

Exonerate parameter	Value	Description
alignment model	est2genome	This model is similar to the affine:local model, but it also includes intron modelling on the target sequence to allow alignment of spliced to unspliced coding sequences for both forward and reversed genes
bestn	1	Report the best N results for each query (only results scoring better than the score threshold will be reported)
minintron	30 (default)	Minimum intron length limit
maxintron	100000	Maximum intron length limit
raw score	100 (default minimum score)	The sum of transition scores used in the dynamic programming. The maximum score depends on the query sequence length.
gapopen	-12 (default)	This is the gap open penalty
gapextend	-4 (default)	This is the gap extension penalty
mismatch	-9 (default)	Mismatch penalty

Table 2.3: Number of reporters placed in the genomic annotation groups of the maize Agilent-016047 microarray.

Annotation groups	Number of reporters [a]	Reporters with Signal/Noise > 3 (maize leaves)[b]	Proportion of reporters with Signal/Noise > 3[c]
sense gene model	23668 (56.3%)	20752	88%
antisense gene model	4330 (10.3%)	2470	57%
gDNA	1549 (3.7%)	985	64%
EST	3390 (8.1%)	1920	60%
ambiguous	2608 (6.2%)	2208	84%
inconclusive	6489 (15.4%)	4038	61%
Total	42034 (100%)	32373	77%

[a] The number and percentage of reporters in each annotation group.
[b] Hybridization to fifty Agilent-016047 arrays by maize leaf cDNA from a segregating population.
[c] Percentage calculated from the number of reporters with signal/noise > 3 in maize leaves divided by the total number of reporters in each annotation group.

Table 2.4: List of studies using the Agilent-016047 maize microarray.

Publication	Maize Tissue	Signal	No Signal	Criteria to define a "Signal"
Ma *et al.* (2008)	anther and pollen	>24K	<18K	3 out of 4 hybridization signals > background (99% confidence)
Casati *et al.* (2008)	adult leaves	26K	16K	median reporter intensity > background
Skibbe *et al.* (2009)	developing anthers	30K	14K	reporter intensity > 2.6 X SD of background
Wang *et al.* (2010*a*)	fertile anthers	32K	10K	not specified
Rajhi *et al.* (2011)	roots	no info	no info	not specified
Current study	adult leaves	32K	10K	signal/noise > 3

Table 2.5: Case studies using the Maize Microarray Annotation Database.

Reference	Description *	Table (see Reference)	Total	Genomic Annotation Groups #					
				GMS	GMA	G	E	AM	I
Ma *et* al. (2008)	a	Table S3	2285	1614 (70.6%)	141 (6.2%)	54 (2.4%)	108 (4.7%)	148 (6.5%)	220 (9.6%)
Ma *et al.* (2008)	b	Table S4	281	209 (74.4%)	11 (3.9%)	7 (2.5%)	13 (4.6%)	27 (9.6%)	14 (5.0%)
Casati *et al.* (2008)	c	Table S1	2092	1373 (65.6%)	142 (6.8%)	58 (2.8%)	111 (5.3%)	134 (6.4%)	274 (13.1%)
Skibbe *et al.* (2009)	d	Table S1	449	329 (73.3%)	6 (1.3%)	18 (4.0%)	16 (3.6%)	35 (7.8%)	45 (10.0%)
Skibbe *et al.* (2009)	e	Table S1	399	279 (69.9%)	15 (3.8%)	16 (4.0%)	18 (4.5%)	23 (5.8%)	48 (12.0%)
Wang *et al.* (2010*a*)	f	Table S3	416	159 (38.2%)	58 (13.9%)	22 (5.3%)	57 (13.7%)	14 (3.4%)	106 (25.5%)
Rajhi et al. (2011)	g	Table S2	239	177 (74.1%)	15 (6.3%)	5 (2.1%)	10 (4.2%)	12 (5.0%)	20 (8.4%)
Rajhi et al. (2011)	h	Table S3	336	278 (82.7%)	7 (2.1%)	8 (2.4%)	14 (4.2%)	15 (4.5%)	14 (4.2%)

* Description
a = Transcripts differentially expressed between meiotic and post-meiotic stages
b = Zinc finger-related proteins
c = Transcripts that are expressed differentially between mbd101 and chc101 RNAi transgenic plants and WT non transgenic siblings under control and/or UV-B conditions.
d = Up-regulated genes between Mu-active vs inactive lines (Mitotic stage)
e = Down-regulated genes between Mu-active vs inactive lines (Mitotic stage)
f = 1.0 mm stage-specific genes of ms8 anthers expressed at later stages in normal anthers
g = Genes up-regulated in maize root cortex during aerenchyma formation
h = Genes down-regulated in maize root cortex during aerenchyma formation

# Genomic annotations:
GMS = Annotation by sense gene model
GMA = Annotation by antisense gene model
G = Annotation by gDNA (Genomic position)
E = Annotation by EST
AM = Ambiguous annotation
I = Inconclusive annotation

Chapter 3

Gene co-expression network analysis of a maize RIL population exposed to GLS disease

3.1 Introduction

Cercospora zeina Crous & U. Braun causes grey leaf spot (GLS), a yield-limiting disease on maize in South Africa. *C. zeina* is considered a hemibiotroph, since it first establishes a biotrophic interaction with its host and later switches to a destructive necrotrophic lifestyle.

Wisser *et al.* (2006) reviewed the genetic architecture of disease resistance in maize and listed 50 publications on the mapping of maize disease resistance loci. These papers reported the locations of 437 quantitative trait loci (QTL) for disease, 17 resistance genes (R-genes), and 25 R-gene analogs. Eight years later, these numbers are expected to be significantly higher. In maize, the majority of disease resistance deployed in elite varieties in the field is quantitative in nature.

QTLs for resistance to GLS have been reported from several studies in the USA where *C. zeae-maydis* Tehon and E. Y. Daniels is the causal agent (Bubeck *et al.*, 1993; Saghai Maroof *et al.*, 1996; Clements *et al.*, 2000; Balint-Kurti *et al.*, 2008). Only a few studies were published in South Africa, China and Brazil where *C. zeina* and not *C. zeae-maydis* was isolated from the locations sampled (Lehmensiek *et al.*, 2001; Juliatti *et al.*, 2009; Liu and Xu, 2013; Berger *et al.*, 2014). Hotspots of GLS QTLs were identified on chromosomes

one, two, four, five and seven (Berger *et al.*, 2014). Association mapping in a panel of 253 diverse inbred maize lines led to the identification of single-nucleotide polymorphisms (SNPs) in a glutathione S-transferase gene that were correlated with resistance to GLS (Wisser *et al.*, 2011).

However, no GLS resistance genes have been cloned to date and this is currently a major challenge. Since DNA variations impact complex diseases through the perturbations they cause to transcriptional, protein and metabolite networks, these molecular phenotypes are intermediate to the phenotypic effect. Therefore, studying the coordinated expression of gene expression profiles in a segregating population can shed light on co-regulation of genes assumed to be part of common pathways. The subsequent identification of co-expression modules that correlate with GLS disease severity can lead to the identification of mechanisms or pathways that play a role in the defense response of maize to *C. zeina* infection.

The aim of the work reported in this chapter was to combine genome-wide gene expression profiles as well as GLS severity scores across the individuals in the CML444×SC Malawi maize recombinant inbred line (RIL) population in a weighted gene co-expression network analysis (WGCNA) to identify gene co-expression modules relating to *C. zeina* disease severity. Hypotheses of driver/hub genes as regulators and of biological processes associated with the GLS disease response were identified.

3.1.1 Network biology and scale free networks

Network biology deals with networks that summarise complex biological systems as components (nodes) and interactions (edges) between them. It thus offers a quantitative description of complex biological processes. Specifically, gene co-expression networks are increasingly used to investigate the functionality of genes on the system-level (Zhang and Horvath, 2005). Nodes and edges of a network can be used to model pairwise interactions. When analysing complex interactions, such as gene co-expression networks, intuitive network concepts such as modules (sub-networks) and connectivity have proved useful. Since it is thought that the coordinated co-expression of genes encode interacting proteins, insight into the underlying cellular processes can be gained when gene co-expression patterns are studied (Eisen *et al.*, 1998). In such networks, nodes represent gene expression profiles and are connected if the corresponding genes are significantly

co-expressed across samples.

The degree or connectivity of a node, denoted by k, is the number of links it has to other nodes (Figure 3.1). In directed networks one can distinguish the in-degree, the number of directed edges that point toward the node, and the out-degree, the number of directed edges that start at the node. Node degrees characterise individual nodes, whereas a degree distribution can be used to quantify the diversity of the whole network. The degree distribution $P(k)$ gives the fraction of nodes that have degree k and is obtained by counting the number of nodes that have $k = 1, 2, 3\ldots$ edges and dividing it by the total number of nodes (Figure 3.1).

Based on the degree distribution of all the nodes in a network, it is possible to distinguish between different classes of networks. The first class of networks is characterised by a $P(k)$, that peaks at an average and decays exponentially for large k. An example is a random network, such as the random graph model of Erdös and Rényi (1960), which follows a peaked (Poisson) distribution, indicating that most nodes have approximately the same number of links (Figure 3.2A). By contrast, degree distributions may follow a well-defined functional form $P(k) = k^{-\gamma}$ called a power law, where the degree exponent γ is usually in the range $2 < \gamma < 3$ (Albert and Barabási, 2002). This function indicates that there is a high diversity of node degrees and no typical node in the network that could be used to characterise the rest of the nodes. The absence of a typical degree (or typical scale) is why these networks are described as "scale-free" (Albert, 2005). The topology of such a network is dominated by a few highly connected hub nodes, holding together numerous other less connected nodes in the network (Barabási and Bonabeau, 2003) (Figure 3.2B).

One example of a scale-free network is the network of scientific papers, connected by citations. The most cited articles in the scientific literature stimulate more researchers to read and cite these same articles (Barabási and Bonabeau, 2003). A variety of complex systems, including most biological networks, share this important property of hub nodes. When a scale-free network emerges, new nodes are preferentially attached to already established nodes (destined to be hub nodes). In contrast with random networks, scale-free networks display a very high degree of tolerance against random failures. For example, although key components occasionally malfunction in complex communication networks, local failures rarely lead to the loss of the global information-carrying ability of the

network (Albert *et al.*, 2000). However, removal of hub nodes, which play a vital role in maintaining the network's connectivity, comes at a high price in that these networks are extremely vulnerable to attacks (Albert and Barabasi, 2000). According to Zhang and Horvath (2005), most gene co-expression networks exhibit a scale-free topology at least approximately. One would expect to find hub nodes in a biologically meaningful co-expression network.

3.1.2 Steps in gene co-expression network analysis

Gene co-expression network analyses arose from the merging of network theory and gene expression data analysis techniques. In general, firstly, gene expression profiles are used to create a pairwise Pearson correlation matrix of all genes across the individuals in a population (or across treatments). The correlations are then transformed to connection strengths in the form of an adjacency matrix, from which a gene co-expression network can be constructed. The identification of sub-networks, which are clusters of densely interconnected genes, is a natural next step. To determine whether a co-expression module is biologically meaningful, one can use functional enrichment and gene ontology information. Ultimately, the identified modules can be related to one another and to external trait information.

Langfelder and Horvath (2008) published an R software package, called Weighted Gene Co-expression Network Analysis (WGCNA), for performing various aspects of weighted co-expression network analysis. The package includes functions for network construction, module detection, gene selection, visualisation, and interfacing with external software. The remainder of this section revisits the various steps in a gene co-expression network analysis, with a technical emphasis to illustrate the steps. Most of the methodology mentioned below have been adopted in WGCNA.

Gene co-expression similarity

The first step in gene co-expression network analysis, is to construct a $n \times n$ similarity matrix $S = [s_{ij}]$ (a matrix of values that express the similarity between two data points) between all gene expression profile pairs, across experiments or samples (Figure 3.3 (a)). The standard similarity or co-expression measure used in gene expression cluster analyses, is the absolute value of the Pearson correlation coefficient, $s_{ij} = |cor(i, j)|$. To preserve

the sign of the correlation, one could use $s_{ij} = \frac{1+cor(i,j)}{2}$. For each pair of genes i and j, the similarity measure denoted by s_{ij} must be a value in $[0, 1]$.

Adjacency functions

Any network is based on an adjacency matrix, which encodes the connection strengths between pairs of nodes. To transform the similarity matrix into an n × n adjacency matrix $A = [a_{ij}]$, one needs to define an adjacency function (Figure 3.3 (b)). By convention, the diagonal elements of A are set to 0 and a_{ij} must be a value in $[0, 1]$. Depending on the choice of adjacency function, the resulting network will be unweighted (hard thresholding) for example the *signum* function, or weighted (soft thresholding) for example the *sigmoid* and *power* functions:

$$a_{ij} = signum(s_{ij,}\tau) \equiv \begin{cases} 1 & if\ s_{ij} \geq \tau \\ 0 & if\ s_{ij} < \tau \end{cases} \tag{3.1}$$

$$a_{ij} = sigmoid(s_{ij}, \alpha, \tau_0) \equiv \frac{1}{1 + e^{-\alpha(s_{ij} - \tau_0)}} \tag{3.2}$$

$$a_{ij} = power(s_{ij}, \beta) \equiv |s_{ij}|^{\beta} \tag{3.3}$$

The *signum* adjacency function depends on the parameter τ, the *sigmoid* function on α (slope parameter) and τ_0 (shift parameter) and the *power* function on β. Drawbacks of hard thresholding are loss of information and sensitivity to the threshold choice (Carter *et al.*, 2004). An increased value of τ may lead to fewer node connections and thus less noise in the network. However, for a too large value of τ, not enough nodes will be connected to detect network modules. Since binary information is used in equation 3.1 (connected=1, unconnected=0), the node connectivity equals the number of direct neighbors. It is biologically more meaningful to use continuous rather than binary information to encode gene co-expression. However, a disadvantage with soft thresholding is that it is not clear how to define the directly linked neighbors of a node; one can only threshold the connection strengths. Node connectivity in the case of soft thresholding is defined as the row sum of the adjacency matrix (k_i in Figure 3.3 (b)).

Since metabolic networks have been found to display approximate scale-free topology (Jeong *et al.*, 2000), Zhang and Horvath (2005) proposed the scale-free topology criterion

for choosing the parameters of an adjacency function. A defining property of scale-free networks (see section 3.1.1) is that the probability that a node is connected with k other nodes (the degree distribution $p(k)$ of a network) decays as a power law, $p(k) \sim k^{-\gamma}$. The value of γ determines the properties of the system. For smaller values of γ, the role of the hubs in the network are more important. To visually inspect whether approximate scale-free topology is satisfied for a given network, one can plot $log(p(k))$ versus $log(k)$. The model fitting index R^2 of the linear model that regresses $log(p(k))$ on $log(k)$ can be used as a testing measure. If R^2 approaches 1, indicating a straight line, the scale-free topology criterion is satisfied. Only adjacency function parameter values that give rise to networks that satisfy approximate scale-free topology at least approximately, i.e. $R^2 > 0.8$, should be considered. There is a natural trade-off between maximizing scale-free topology model fit (R^2) and maintaining a high mean number of connections in the network, i.e. the mean connectivity.

Identifying gene co-expression modules

There are various different methods for detecting subsets of nodes that are tightly connected to each other, i.e. modules. Intuitive views of modularity assume the existence of a set of modules with varying sizes, potentially separated from other modules. In contrast, Ravasz *et al.* (2002) found that the metabolic network, being a scale-free network, has an inherent self-similar property: there are many highly integrated small modules, which group into a few larger modules, which in turn can be integrated into even larger modules. They illustrated that the hierarchical organisation of modularity revealed biologically meaningful modules. Zhang and Horvath (2005) have used a dissimilarity measure in conjunction with a hierarchical clustering method for module identification, which was applied in WGCNA. Specifically, the topological overlap dissimilarity measure was employed and modules were defined as groups of nodes with high topological overlap.

The topological overlap matrix (TOM), $\Omega = [\omega_{ij}]$, provides a similarity measure which reflects the relative interconnectedness of pairs of nodes:

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{min\{k_i, k_j\} + 1 - a_{ij}} \tag{3.4}$$

where $k_i = \sum_u a_{iu}$ (node connectivity) and $l_{ij} = \sum_u a_{iu}a_{uj}$ (the number of nodes to which both nodes i and j are connected) (Figure 3.3 (c)). When $\omega_{ij} = 1$, the node with fewer

connections meet two conditions: (i) all of its neighbors are also neighbors of the other node and (ii) it is connected to the other node. In contrast, when $\omega_{ij} = 0$, the two nodes i and j are not connected and do not share any neighbors. TOM is a similarity measure since it is non-negative, symmetric and ω_{ij} is a value in $[0, 1]$. TOM can be transformed into a dissimilarity measure, by subtracting it from one (Figure 3.3 (d)):

$$d_{ij}^{\omega} = 1 - \omega_{ij} \tag{3.5}$$

WGCNA uses average linkage hierarchical clustering (Figure 3.3 (e)) based on TOM-based dissimilarity, d_{ij}^{ω}, to identify gene co-expression modules. Gene modules are identified to correspond to branches of the resulting hierarchical clustering dendrogram (Figure 3.3 (f)). Selecting a height cut-off to cut the tree branches is a judgement call that can be guided by inspection of the TOM plot; a color-coded heatmap of the values of the TOM-based dissimilarity matrix (see Figure 3.4 for an example TOM plot). A height cut-off value can be chosen such that some of the resulting branches correspond to dark squares along the diagonal (Zhang and Horvath, 2005).

Relating modules to external traits

The gene expression profiles of a module can be summarised with a weighted average expression profile. Such a profile is called a module eigengene (ME), and is defined as the first principal component of a given module (Figure 3.3 (f)). Finding a biologically significant module depends on the research question under consideration. For example, when one wants to identify modules relating to a phenotypic trait T, a significance measure between the module eigengene and the trait can be defined by calculating an eigengene significance score (ES) of module q (Figure 3.3 (g)):

$$ES^{(q)} = |cor(ME^{(q)}, T)| \tag{3.6}$$

where $ME^{(q)}$ is the module eigengene of module q. Modules with high trait significance may represent pathways associated with the phenotypic trait (Langfelder and Horvath, 2008). In a similar way, a gene significance (GS) score can also be calculated for each gene, to quantify the association of that gene with the external trait:

$$GS_i = |cor(x_i, T)| \tag{3.7}$$

where x_i is the profile of node i. In addition, a module membership (MM) score can be calculated for each gene to quantify the association of that gene to a specific module:

$$MM_i^{(q)} = |cor(ME^{(q)}, x_i)| \tag{3.8}$$

where $ME^{(q)}$ is the module eigengene of module q and x_i is the profile of node i. The MM measure can also be used to find intramodular hub genes, i.e. genes with a high correlation to the module eigengene. In modules related to a trait of interest, genes with high GS values often also have high MM values. Such genes are natural candidates for further validation.

3.1.3 Application of gene co-expression networks

Different types of experimental designs can be used in a gene co-expression network analysis study. For example, a correlation matrix can be calculated from gene expression profiles either throughout a time course, or across population-based samples. Both designs can be used to gain a better understanding of the processes involved in a biological system. As an example of the first mentioned design, Adhikari *et al.* (2012) performed expression profiling of cultivated cucumber (*Cucumis sativus*) over a time course of infection with the downy mildew pathogen *Pseudoperonospora cubensis* to identify genes, pathways, and systems that are altered during a compatible interaction. Through co-expression network analyses, modules of temporal-specific transcriptional networks were identified, which provided a basis for connecting transcription factors with defense response genes. With this design however, causal interactions between transcript levels and phenotypic traits cannot be studied, since these samples and resulting modules do not involve genetic variation. On the other hand, in population-based studies where global transcript levels and phenotypic traits are measured, one could incorporate QTL and eQTL mapping (which depend on genetic variation) with gene co-expression network analysis (which does not necessarily depend on genetic variation) in order to study the connections between genotypes and phenotypes.

Gene co-expression network analysis has recently been used to discover specific genes

and pathways affecting various biological systems. Azuaje *et al.* (2013) used a gene co-expression network analysis to establish a robust association between Col5a2 (a relatively uncharacterised gene) and ischemic heart disease. The analysis was based on microarray data originating from a mouse model of myocardial infarction (MI). After Spearman co-expression coefficients were calculated among all pairs of genes, a weighted gene co-expression network was generated. Candidate clinically relevant clusters were detected by applying A-CODE (association-centered community detection algorithm). Motivated by their results, Azuaje *et al.* (2013) further assessed the potential relevance of Col5a2 in MI by estimating its disease discriminatory capability in previously generated microarray datasets. Col5a2 was identified as a potential novel candidate marker for the identification and treatment of ischemic heart disease. In a different study, Kugler *et al.* (2013) performed a QTL-dependent analysis of a gene co-expression network associated with Fusarium head blight (caused by *F. graminearum* Schwabe) resistance in bread wheat (*Triticum aestivum* L.). A combination of a network-driven approach, using WGCNA, and differential gene expression analysis identified genes and pathways associated with two well-validated and highly reproducible QTLs, *Fhb1* and *Qfhs.ifa-5A*. RNA-seq data from near-isogenic lines (NILs), harboring either the resistant or the susceptible allele for *Fhb1* and *Qfhs.ifa-5A*, was utilised. Genes involved in the biosynthesis and metabolism of riboflavin were found more abundant after infection in lines harboring *Qfhs.ifa-5A*. Furthermore, G-protein coupled receptor kinases and biosynthesis genes for jasmonate and ethylene earlier induced for NILs harboring *Fhb1* were identified.

A variety of other studies that also employed co-expression networks, obtained less specific results. However, valuable insight and new hypotheses regarding candidate genes and pathways governing their respective biological systems were extracted. Villa-Vialaneix *et al.* (2013) conducted a case study on mammalian species. They used a gene co-expression network to reveal biological functions underlying eQTLs. Key genes and gene clusters that were related to muscle pH were highlighted as a potential focus for forthcoming biological experiments. Zhang *et al.* (2010) used gene co-expression network analysis to identify a set of genes that are potential prognostic biomarkers for chronic lymphocytic leukemia. In plants, Shaik and Ramakrishna (2013) used WGCNA to detect consensus modules in *Arabidopsis* and rice, based on differentially expressed genes common to drought (abiotic) and bacterial (biotic) stress. They identified 9 and 4 modules

in rice and *Arabidopsis*, respectively, with either conserved responsive profiles in both stresses or reversed expression status. Wang *et al.* (2012*b*) used co-expression network analysis to identify cell-wall related genes in *Arabidopsis*.

Approximately one third of the studies mentioned above used WGCNA to construct gene co-expression networks, thus many other tools and algorithms also exist for this purpose. Open access databases and tools for the investigation of gene co-expression networks in plants are currently available for *Arabidopsis*. Researchers can either calculate gene-to-gene correlation coefficients using their own expression data sets, or retrieve correlation coefficient data from public databases (Aoki *et al.*, 2007). *Arabidopsis* co-expression tool (ACT) (Jen *et al.*, 2006), ATTED-II (Obayashi *et al.*, 2007, 2009), Genevestigator (Zimmermann *et al.*, 2004) and the Botany Array Resource (BAR) (Toufighi *et al.*, 2005) are databases of gene co-expression data implemented with gene expression visualising tools. Single or multigene queries can be submitted to these tools. As an example, the ACT database stores pre-calculated co-expression results for 21,800 genes based on data from over 300 arrays. When a query gene is submitted, ACT ranks the genes across these microarray datasets according to how closely their expression follows the expression of the query gene. An extra tool within ACT called Clique Finder can identify groups of genes, which are consistently co-expressed with each other across a user-defined co-expression list (Manfield *et al.*, 2006). Additionally, the AraNet probabilistic functional gene network of *Arabidopsis* can be searched (Lee *et al.*, 2010). The network was constructed from 24 distinct types of gene-gene associations from multiple organisms consisting of > 50 million individual experimental or computational observations, where each observation was scored for its ability to correctly reconstruct shared membership in *Arabidopsis* biological processes. AraNet provides a resource for plant gene function identification and genetic dissection of plant traits. An AraNet search will return all of the genes in the netowrk that are directly connected to an input gene (or set of genes), ranked according to their log-likelihood scores. Also worthy of mentions is CSB.DB, a comprehensive systems biology database consisting of bio-statistical analyses on gene expression data in association with additional biochemical and physiological knowledge (Steinhauser *et al.*, 2004). For grapevine (*Vitis vinifera*), a gene co-expression database called VTCdb, recently became available (Sweetman *et al.*, 2013). It stores over 800 publicly available microarray datasets that were selected to construct global co-expression networks. The database is

also equipped with functional enrichment and visualisation capabilities. From the variety of studies mentioned above and the different tools available for gene co-expression network analysis, it is evident that researchers currently consider this a valuable method to uncover genes and pathways associated with different biological systems.

3.2 Aims and objectives

A major aim of this chapter was to establish whether coordinated responses to *C. zeina* infection under field conditions were evident in a maize RIL population by exploiting genome-wide gene expression profiles. The ultimate goal was to identify genes and pathways that respond to *C. zeina* infection in a susceptible or resistant response, in this particular maize RIL population derived from two sub-tropical inbred lines that have been bred for maize growing conditions in southern Africa.

Specific objectives were to employ a correlation analysis based on WGCNA, in order to: (i) identify gene co-expression modules; (ii) identify the modules relating to *C. zeina* disease severity; (iii) determine enriched functional categories within the identified gene co-expression modules; and (iv) identify intramodular hub genes (drivers) for the biologically relevant modules.

3.3 Materials and methods

3.3.1 Germplasm and field trials

A recombinant inbred line population (RIL, F7:S6) derived from a cross between subtropical white dent inbred lines CML444 and SC Malawi was used (Messmer *et al.*, 2009). A total of 145 RILs were planted at Baynesfield Estate in KwaZulu-Natal Province, South Africa in December 2008. The plot design was a randomised block with three replicates. Each replicate of a RIL was a row of 10 plants. GLS disease severity was scored on a per plant basis using a $1-9$ scale, where 1 and 9 represent no GLS disease and full GLS susceptibility, respectively (Munkvold *et al.*, 2001). The GLS disease severity scores for each of the three replicate rows were averaged and recorded in order to be used as input for subsequent analysis. GLS disease severity data was collected at 92, 99, 109 and 116 days after planting (DAP). The maturity of the RILs was between R1 and R4 over this

period.

3.3.2 RNA extraction and microarray analysis

RNA was extracted from three biological repeats of 100 selected RILs at 103 DAP, sampled at Baynesfield in March 2009. A Bioanalyser RNA quality control assay was run on all 300 samples. The samples were of good quality, with an average RNA integrity number (Schroeder *et al.*, 2006) of 7.7. The RNA from the three biological repeats was pooled to create 100 samples. Each biological repeat consisted of leaf pieces from two different plants in a RIL row. Fifty RIL microarrays were processed according to the "distant pair" experimental design, based on the method described in Fu and Jansen (2006). The choice of the 100 RILs and the microarray design were determined with a customised computer script based on genotyping results, such that samples with dissimilar genomes were paired. RNA was amplified from all 100 pools, labeled with either Cy3 or Cy5 and hybridised to Agilent 4×44K maize arrays. Analysis of spike-in controls indicated that labeling and hybridisation worked well and background for both channels was low.

Normalisation of the expression data was performed in the R-based software package *limma* (Smyth, 2004), with a weighting of zero for flagged spots. This meant that data from flagged spots did not influence the normalisation of other spots, although all the data was transformed at each step. Approximately $15,000$ reporters were flagged in less than 10% of the arrays, whereas roughly $19,000$ reporters were flagged in more than 90% of arrays. Background correction was performed using the *normexp* method (offset = 50) (Ritchie *et al.*, 2007). The *loess* method was used for normalisation within arrays and *Aquantile* for normalisation between arrays (Yang and Thome, 2003).

After normalisation, there were 50 datasets of M and A values representing expression data from 100 RILs. Each M and A value had been calculated from hybridisdation of pairs of RILs to each array. Therefore, back-conversion was required to obtain separate expression values per reporter for each of the 100 RILs. The back-conversion of the normalised data was performed, using the final M and A values, by solving simultaneously for R (Red intensity; Cy5 channel) and G (Green intensity; Cy3 channel) from the formulas $A = \frac{1}{2} log_2(RG)$ and $M = log_2(\frac{R}{G})$, to yield $G = \sqrt{\frac{2^{2A}}{2^M}}$ and $R = \sqrt{2^{2A}2^M}$. Out of the $42,034$ maize gene reporters on the Agilent arrays, after removal of flagged reporters, back-converted intensity expression profiles for $30,280$ reporters across the 100 RILs were

obtained.

3.3.3 Network construction and module identification

WGCNA was performed as described previously (Langfelder and Horvath, 2008) with an R-script modified for this analysis (available in the electronic Appendix). The input data matrix consisted of 100 RILs (columns) and 30,280 microarray reporters (rows). This matrix was log_{10}-transformed relative to the reporter gene expression means (row means; across the RILs), prior to the WGCNA analysis. The reason for the log-transformation, instead of using the raw back-converted expression values, was that the choice of parameter for the power adjacency function (based on the scale-free topology criterion) was more intuitive for the log-transformed data.

Before network construction, a filtering step in the WGCNA analysis removed reporters with zero variance as well as reporters with more than 50% missing entries. After filtering, 19,281 reporters, representing 14,201 maize gene models according to the annotations in Chapter 2 (Coetzer *et al.*, 2011), remained in the data set. To assess whether any samples were outliers, which could disqualify the scale-free topology criterion, average linkage hierarchical clustering with Euclidean distance was used to draw the sample dendogram in Figure 3.5. By inspection, no obvious outlying branches were observed in the dendogram and therefore no RILs were removed from the analysis. When there seem to be outliers, a constant-height cut can be implemented on the dendogram in order to remove branches with less than a pre-specified number of objects.

The next aim was to choose a soft thresholding power β, in order to calculate the adjacency matrix $a_{ij} = cor(x_i, x_j)^{\beta}$ (the *power* adjacency function was used; equation 3.3). The function *pickSoftThreshold* from the R package *WGCNA* was used to guide the selection of a proper soft-thresholding power by applying the approximate scale-free topology criterion. Figure 3.6 shows the scale-free topology fit index (R^2) as well as the mean connectivity for various soft-thresholding powers. Zhang and Horvath (2005) recommended only considering parameter values that would lead to a network satisfying scale-free topology at least approximately, i.e. $R^2 > 0.8$. Furthermore, the mean connectivity need to be as high as possible so that the network contains enough information for module detection. The best estimate soft-thresholding power for this dataset was 12, since it was the lowest power for which the scale-free topology fit index was above 0.8.

To minimize effects of noise and spurious associations, the adjacency matrix was transformed into a topological overlap matrix TOM (equation 3.4 on page 79). The resulting TOM was converted into a dissimilarity matrix, $1 - TOM$ (equation 3.5 on page 80), which was used in hierarchical clustering to produce a dendrogram of genes. Since densely interconnected and co-expressed genes were grouped together in branches of the dendogram, the *Dynamic Tree Cut* method was used (with default parameters) to detect clusters by branch cutting. Lastly, modules with similar expression profiles were merged and colours were assigned to the final modules in order to distinguish between the modules (Figure 3.9 on page 122).

3.3.4 Relating modules to GLS disease and biological interpretation

The *moduleEigengenes* function from the *WGCNA* R package was used to calculate a summary profile, called a module eigengene, for each gene co-expression module. Pearson correlation was used to determine the correlation between the module eigengenes and the GLS severity profile (equation 3.6). For each correlation, an associated p-value was also calculated to indicate the significance of the corresponding Pearson correlation coefficient. The GLS severity profile consisted of the GLS severity scores from 1 (indicating a resistant RIL) to 9 (indicating a susceptible RIL). Therefore, a positive correlation indicated that higher expression related to higher GLS severity scores (i.e. GLS susceptibility). For convenience, a profile called "GLS swop" was also calculated. Ten minus the normal GLS severity scores yielded swopped scores (also between 1 and 9). In this case, a positive correlation indicated that higher expression related to GLS resistance. The normal GLS severity profile was used throughout this chapter, unless "GLS swop" was specifically mentioned.

Gene-wise correlations were also calculated to quantify the association of each gene with the trait (GLS disease severity) and with each module (the module eigengene). These correlations were presented as GS scores (equation 3.7) and MM scores (equation 3.8), respectively.

Out of the 19,281 reporters that were used in WGCNA network construction, 17,829 (92%) were annotated with a maize gene ID according to the Maize Microarray Annotation Database (Chapter 2). A *Z. mays* annotation file, which was released as part of

Phytozome version 7.0 (`http://www.phytozome.net`), was downloaded from their FTP site. The file included the best *Arabidopsis* TAIR10 and rice BLAST hits for each maize gene. For 84% and 87%, respectively, of the 17,829 Agilent reporters, *Arabidopsis* and rice best BLAST hits were recorded. The resulting *Arabidopsis* and rice hit descriptions together with the BLAST2GO description for each gene (which was extracted from the Maize Microarray Annotation Database), were used to formulate a final functional annotation per reporter.

BiNGO (Maere *et al.*, 2005) was used to identify enriched GO-terms (`http://www.geneontology.org`) in order to determine whether genes in the same co-expression modules were involved in the same biological processes. Groups of "best BLAST hit" *Arabidopsis* IDs (from Phytozome) corresponding to maize genes in the same modules, were used as input to the BiNGO analyses. Default BiNGO parameters were used and the reference set corresponded to the $19,281$ reporters whose expression profiles were included in the WGCNA analysis. As an alternative to using the full GO hierarchy, BiNGO provides several GOSlim ontologies that are organism-specific slimmed-down versions of the full GO hierarchy. GOSlim ontologies generally give a broad overview of the ontology content without the detail of the specific fine-grained terms. In cases where the full GO hierarchy did not produce significantly enriched GO-terms, additional BiNGO analyses based on the plant GOSlim ontology were performed. All GO enrichment output Tables list the enriched GO-terms from the three categories (i.e. biological process, molecular function and cellular component) together in one table, sorted by significance. In addition, MapMan was used to functionally classify genes into predefined bins (Thimm *et al.*, 2004). The MapMan ontology comprises a set of 34 tree-structured bins, describing a variety of cellular processes.

Node and edge files were generated by the WGCNA package in R and imported to Cytoscape version 2.8.2 (Cline *et al.*, 2007; Smoot *et al.*, 2011). Cytoscape was used to visualise the co-expression modules as a network and to calculate topology parameters, such as the node degree (the number of edges connected to a specific node). Candidate module hub/driver genes were identified by ranking the reporters within each co-expression module by their node degree. A hub gene is expected to have a similar gene expression profile than the module eigengene, since it is highly co-expressed with many of the genes in a module. Thus, MM scores can also be used to find intramodular hub genes, i.e.

genes with a high correlation to the module eigengene. In modules related to a trait of interest, genes with high GS scores often also have high MM scores. Such genes are natural candidates for further validation.

3.4 Results and discussion

3.4.1 The maize RIL population exposed to GLS disease

A maize RIL population derived from a cross between the subtropical parental lines CML444 and SC Malawi had previously been shown to segregate for quantitative resistance to GLS disease over several seasons and field sites in South Africa (Berger *et* al., manuscript in preparation). This population was grown at the Baynesfield Estate (KwaZulu-Natal, South Africa) over the 2008/2009 summer season and was scored for GLS using a $1-9$ scale (as used in Munkvold *et al.*, 2001). Typical GLS disease symptoms were observed with lesions first developing on lower leaves and progressing to higher leaves. The maturity of the RILs was between growth stages R1 and R4 when GLS disease was scored (stages R2 to R6 are associated with grain filling and maturity; `http://maizedoctor.cimmyt.org`). Figure 3.7 on page 121 gives four photos of samples that were harvested, including a representative resistant and susceptible RIL. Differences in the number of lesions on the earleaf were visible between 92, 99, 109 and 116 DAP, with most of the RILs showing intermediate levels of disease severity at each of the four ratings ($r1-r4$ in Figure 3.8).

Global gene expression profiling using the Agilent 44K microarray was carried out on earleaf samples collected from 100 RILs in March 2009, at 103 DAP (between ratings 2 and 3). Expression profiles for 30,280 microarray reporters across the 100 RILs were obtained after filtering, normalization and back-conversion. The aim was to determine if there were any patterns of co-expression of genes across the RIL population at this stage of disease pressure, and whether these correlated with disease severity.

A weighted average of the GLS severity scores across the four ratings was calculated for the 2008/2009 season, depending on the number of days between the GLS rating and the day of RNA sampling. The resulting scores were summarised in a weighted average boxplot (*WA* in Figure 3.8), which shows that some RILs exhibit less disease than the two parents and some RILs more disease than the two parents. CML444 and SC Malawi

scored 3.7 and 5.7, respectively. These values were used as the "GLS severity scores" in subsequent correlation analyses to the above-mentioned gene expression data.

3.4.2 Co-expression module identification and relation to GLS disease

The WGCNA input data matrix consisted of 100 samples and 19,281 microarray reporters, after removing reporters with zero variance and more than 50% missing entries. Almost half of the reporters ($8,665/19,281 = 45\%$) were assigned to 42 co-expression modules, with a minimum module size of 30 (default value) (Figure 3.9). A full list of the microarray reporters and its functional annotations per co-expression module are available in the electronic Appendix. When changing the parameter to 5 instead of 30, 62% of the reporters were assigned to 269 co-expression modules. The larger modules did not differ remarkably from those detected previously, however many additional smaller modules were identified. A decision was made to continue with the parameter set to a minimum module size of 30.

The next step was to determine if any of the identified co-expression modules significantly correlated with the GLS severity scores across the individuals in the RIL population. The module eigengenes of eight co-expression modules significantly correlated (p-value < 0.05) to the GLS severity profile (Table 3.1). The greenyellow module, consisting of 185 reporters, had the strongest positive correlation (0.71) and the turquoise module, consisting of $1,564$ reporters, had the strongest negative correlation (-0.31) to the GLS severity profile (Table 3.1). A positive correlation indicates that higher module eigengene expression values were associated with higher disease severity scores (more susceptibile RILs), i.e. "H" in Table 3.1, whereas a negative correlation indicates that higher module eigengene expression values were associated with lower disease severity scores (more resistant RILs), i.e. "L" in Table 3.1. Figure 3.10 illustrates this by showing the module eigengene expression values across the RILs, sorted by GLS severity scores, for the greenyellow and turquoise modules, respectively.

A co-expression module eigengene dendogram and adjacency heatmap was generated with the *plotEigengeneNetworks* function in the WGCNA R package. Input to the clustering was the module eigengenes as well as the GLS severity profile as either "GLS" (higher scores indicate susceptibility) or "GLS swop" (higher scores indicate resistance).

The aims were to identify groups of correlated module eigengenes termed meta-modules and to relate the GLS severity profile to the meta-modules (Figure 3.11). Branches of the dendrogram as well as squares of red color along the diagonal of the adjacency map, group eigengenes together that are positively correlated (meta-modules). Interestingly, the three modules with a significant negative correlation (p-value < 0.05) to the GLS severity profile (turquoise, darkred and yellow in Table 3.1) together with the green module formed a meta-module, which also included "GLS swop". However, the mutual correlations of the eigengenes in this meta-module were stronger than their correlations with "GLS swop". As expected, the greenyellow module, with a strong positive correlation to the GLS severity profile, clustered tightly with "GLS". Interestingly, each of the remaining modules with a significant positive correlation to the GLS severity profile (paleturquoise, blue, yellowgreen and magenta in Table 3.1), was part of a different meta-module.

Figure 3.12 gives the module-trait associations for numerous field trials, where GLS was scored for this population at different locations and during different seasons. The first trait called "B_09_GLS" was the focus of this study, since these GLS scores corresponded to the same GLS infected plants from which RNA was extracted for the gene expression study (Baynesfield, season 2008/2009). The patterns in Figure 3.12 confirmed that the same modules generally correlated to GLS resistance and susceptibility across different field trials. The greenyellow and turquoise modules, previously identified as the strongest positive and negative correlating modules to GLS severity (Table 3.1), respectively, were well-correlated with GLS throughout the field trials. For 94% of the field trials, the greenyellow module had a significant positive correlation to the GLS scores (p-value $<$ 0.05) and for 88% of the field trials this module was the top correlating module to GLS susceptibiltiy. For 82% of the field trials, the turquoise module had a significant negative correlation to the GLS scores (p-value < 0.05) and for 71% of the field trials this module was the top correlating module to GLS resistance.

3.4.3 Interpretation of GLS-related co-expression modules

Co-expression modules are suggestive of coordinated regulation of genes. Therefore, as an initial step, GO enrichment was used to assess the functional significance of each module that significantly correlated with GLS severity. Considering the genes in a co-expression

module, two types of gene lists that can be useful: (i) the genes that highly correlate with the module eigengene (the MM score in equation 3.8; also potential hub genes); and (ii) the genes that highly correlate with the trait (the GS score in equation 3.7). Genes that highly correlate with the trait and also highly correlate with the module eigengene could potentially be global regulators of processes relating to the trait.

Five co-expression modules had significant positive correlations (p-value < 0.05) with the GLS severity scores. These modules correlated with GLS susceptibility, since higher module eigengene expression values were associated with higher disease severity scores. The greenyellow (185 genes), paleturquoise (41 genes), blue (1521 genes), yellowgreen (35 genes) and magenta (266 genes) modules had correlation coefficients of 0.71, 0.31, 0.22, 0.21 and 0.20, respectively (Table 3.1). Three co-expression modules had significant negative correlations (p-value < 0.05) with the GLS severity scores. These modules correlated with GLS resistance, since higher module eigengene expression values were associated with lower disease severity scores. The turquoise (1564 genes), darkred (63 genes) and yellow (1170 genes) modules had correlation coefficients of -0.31, -0.24 and -0.23 respectively (Table 3.1). Figure 3.11 provides an overview of the relationships between the modules and the GLS severity traits.

GO enrichment and overview of the greenyellow module

The greenyellow module had an exceptionally strong correlation with GLS susceptibility (the correlation coefficient of 0.71 was significantly higher than that expected by chance; according to a permutation test the expected maximum correlation coefficient was 0.31). Table 3.2 shows that there were a variety of over-represented GO-terms for the 185 reporters in the greenyellow module. Enriched GO-terms in the biological process category included (i) secondary metabolic process, with more specific terms diterpenoid and gibberellin metabolic process; (ii) lipid metabolic process; (iii) catabolic process, including cellular nitrogen compound catabolic process and heterocycle catabolic process; and (iv) response to stress: specifically response to organic substance, response to chitin and gibberellic acid-mediated signalling pathway. In the molecular function category, catalytic activity was highly enriched with more specific terms such as aspartic-type endopeptidase and lipase activity, as well as C4-dicarboxylate transmembrane transporter and malate transmembrane transporter activity. Therefore, since the genes in the greenyellow module

were co-expressed, their expression profiles were highly correlated with GLS severity and the module were enriched for a variety of functional categories, it could be concluded that the susceptible interaction (the presence of lesions) resulted in coordinated transcriptional responses.

Gibberellins (GAs) form a large family of phytohormones that are important for many aspects of plant growth and development, as well as for the discernment of environmental stimuli (Hedden and Kamiya, 1997). Examples of GAs are tetracyclic diterpenoids, products of the terpenoid pathway (together with terpenes), terpene-derived compounds and steroids. GAs were first identified from the necrotrophic fungus *Gibberella fujikuroi* (Yabuta and Sumiki, 1983), which causes super-elongated "bakanae" rice. The pathogen produces GAs, which causes rice seedlings to become spindly and lodge (Grennan, 2006). The necrotroph benefits by later extracting nutrients from the dead host cells. Yang *et al.* (2008) found that GAs negatively regulate rice basal disease resistance against bacterial blight (*Xanthomonas oryzae* pv. *oyrzae*), a biotrophic pathogen, and rice blast fungus (*Magnaporthe oryzae*), a hemibiotrophic pathogen of rice. DELLA proteins are a family of transcriptional repressors of GA responses and their accumulation implicates resistance to necrotrophs and susceptibility to virulent biotrophs, partly by altering the relative strength of JA and SA signalling (Navarro *et al.*, 2008). As a result, GA-activated degradation of DELLA proteins leads to negative regulation of defense against necrotrophs.

Seven enzymes have been identified to be involved in GA biosynthesis, of which four are present in the greenyellow module: ent-copalyl diphosphate synthase (A_92_P008699), ent-kaurene synthase (A_92_P005300), ent-kaurene oxidase (A_92_P020405) and GA 2-oxidase (A_92_P016737). These genes were responsible for the enriched GO-terms "gibberellin metabolic process" and "gibberellin biosynthetic process", as well as for a few GO-terms lower down in Table 3.2, such as "gibberellic acid mediated signalling pathway" and "cellular response to gibberellin stimulus". A BLASTX analysis of the 60-mer reporter sequences against the non-redundant (nr) database confirmed that the four above-mentioned genes were not fungal genes, since only plant hits were obtained. Only one reporter in the list of 19,281 reporters represented a DELLA protein, which did not have a significant correlation to GLS severity (0.1). It could be that other DELLA reporters (i) were not present on the microarray, (ii) were not included in the WGCNA input dataset of 19,281 due to missing data or (iii) were mis-annotated

or annotated as "protein with unknown function". However, one can speculate that up-regulation of GA and consequent down-regulation of DELLA proteins is associated with maize susceptibiltiy to *C. zeina* infection.

Thirty-seven genes in the greenyellow module (20%) were annotated with the enriched GO-term "response to stimulus" and four of these genes were also annotated with the term "response to chitin" (Table 3.2): a WRKY family transcription factor (A_92_P018873), a MYB family transcription factor (A_92_P020962), a zinc finger (AN1-like) protein (A_92_P041535) and an immediate-early fungal elicitor protein CMPG1 (A_92_P019503). Chitin is a major component of fungal cell walls and a general elicitor of plant defense responses (Boller, 1995). Fungal infection induces the expression of chitinases (chitin-degrading enzymes) in plant cells as well as numerous downstream defense response genes. Four genes encoding chitinase-like proteins were present in the greenyellow module and all four were highly correlated to GLS susceptibility (GS scores varied from 0.51 to 0.66). Interestingly, out of the thirteen chitinases in the full set of 19,281 reporters, eight were significantly correlated with GLS susceptibility and two with GLS resistance. One reason for the general strong correlation of chitinases to GLS susceptibility in this study, could be that more chitinases were being made in susceptible plants due to a more severe fungal infection. Furthermore, it has been reported that one race-specific effector AVR4 of the tomato pathogen *Cladosporium fulvum* (a biotrophic fungus), which is a chitin-binding protein, can protect fungi against plant chitinases (van den Burg *et al.*, 2004). One can speculate that *C. zeina* could also be protected against plant chitinases, due to some form of chitin-binding effector. It is further possible that R-genes evolved to recognise this effector, but that these are more effective in resistant maize lines.

Two other highly enriched GO-terms in Table 3.2, was "carboxylic acid transmembrane transporter activity" and "organic acid transmembrane transporter activity". These terms were due to five reporters: a carnitine acylcarnitine translocase (A_92_P040413) involved in metabolite transport at the mitochondrial membrane; an amino acid permease 6 (A_92_P033915) and a lysine histidine transporter 1 (A_92_P017759) involved in amino acid transport; a tonoplast dicarboxylate transporter (A_92_P012461) and a general dicarboxylate transporter (A_92_P019164) involved in ion transport. The latter two reporters were responsible for additional GO-terms lower down in Table 3.2 includ-

ing "malate transport", "dicarboxylic acid transport" and "C4-dicarboxylate transport". Physiological processes in plants strongly depend on solute and water fluxes across the plasma membrane, tonoplast and other endomembranes. Therefore, membrane transporters, such as those mentioned above, have been associated with various processes such as stomatal closure, hormone signalling, membrane excitability, cellular osmoregulation, growth regulation, and anionic nutrition (Tavares *et al.*, 2011).

Another group of enriched GO-terms that linked with the previous mentioned group included "small molecule metabolic process" and "monocarboxylic acid metabolic process". Genes involved in these processes were two 3-ketoacyl-CoA synthases (A_92_P018659, A_92_P032943) and a 3-ketoacyl-CoA thiolase (A_92_P012775) involved in lipid metabolism, an iso-kaurene synthase (A_92_P005300) involved in terpenoid secondary metabolism, an ent-kaurene synthase (A_92_P008699) and a gibberellin 2-oxidase (A_92_P016737) involved in gibberellin hormone metabolism, a malate synthase (A_92_P005391) involved in gluconeogenesis, a cytochrome P450 (A_92_P020405) which is part of the miscellaneous enzyme families and a phospholipase (A_92_P040918) involved protein storage.

Figure 3.13 gives an overview of the functional categories that are associated with the 185 reporters in the greenyellow module. According to this figure, abundant functional categories in the greenyellow module included calcium signalling, regulation of transcription, protein degradation and transport-related genes. The greenyellow module contained many metabolism-related genes, of which lipid and secondary metabolism were the largest categories. Furthermore, numerous biotic and abiotic stress-related genes as well as genes encoding enzymes in the miscellaneous enzyme families were present. Genes in the latter mentioned categories were particularly strongly correlated to GLS susceptibility.

Driver genes in the greenyellow module with high GLS severity correlation

Figure 3.14 shows that the bulk of potential driver genes in the greenyellow module, were also the genes that best correlated with GLS severity. Table 3.3 lists the top 35 genes that best correlated with the greenyellow module eigengene (potential driver genes) and Table 3.4 the top 35 genes that best correlated with GLS susceptibility. More than half of the entries (54%) in the two tables were identical. Furthermore, five out of the top 10 driver genes for the greenyellow module were also in the top 10 GLS severity-correlating

genes. These included an F-box kelch-repeat protein skip11-like (A_92_P037621), carnitine acylcarnitine translocase (A_92_P040413), heptahelical transmembrane protein receptor (A_92_P035171), nodulin MtN3 family protein (A_92_P015826) and a NAC transcription factor (A_92_P032766). These reporters are represented by the 5 largest yellow nodes in Figure 3.14 and are the best candidate regulators of processes relating to GLS susceptibility in the greenyelow module.

An F-box domain is a motif that binds to the Skp1 family of proteins, resulting in the formation of the Skp1 Cullin F-box (SCF) E3 ubiquitin ligase complex. The E3 ligase is involved in the degradation of a specific target protein by polyubiquitination. The kelch-repeat domain is a motif typically involved in protein-protein interactions. In *Arabidopsis*, a kelch repeat-containing F-box protein SON1, acts in the defense response independent of SA and SAR (Kim, 2002). In most of the known phytohormones (e.g. auxin, GA, JA, SA and strigolactone), the signals are mediated by the components of E3 ligase-substrate complexes (Takahara *et al.*, 2013). In gibberellin signalling, the DELLA proteins are targeted for degradation by the F-box proteins SLY1 in *Arabidopsis* and GID2 in rice (Dill *et al.*, 2004; Gomi *et al.*, 2004). It can be hypothesised that the F-box kelch-repeat protein skip11-like gene in Tables 3.3 and 3.4 assist in GA signalling (an enriched process of the greenyellow module) by targeting DELLA proteins for degradation. This gene was part of the protein degradation category in Figure 3.13.

Carnitine-acylcarnitine translocase is an enzyme responsible for transporting both carnitine-fatty acid complexes and carnitine into and out of the mitochondria, across the inner mitochondrial membrane, and is involved in fatty acid degradation and energy metabolism. Yang *et al.* (2012) found that a carnitine-acylcarnitine carrier protein, MoCrc1, is essential for pathogenicity in rice blast fungus (*M. oryzae*). This gene seems to play a vital role in appressorium-mediated infection, where generation of appressorial turgor is needed for penetration. They showed that deletion of this gene severely reduced appressorium turgor generation, appressorial penetration, and development of infection hyphae. However, according to BLASTN searches against the *C. zeae maydis* genome sequence, using the Joint Genome Institute (JGI) Genome Portal, it appears that this reporter (A_92_P040413) does not represent the presence of fungal mRNA (no hits); whereas a 60/60 BLASTN match against the *Zea mays* genome sequence was obtained. This reporter contributed to a few enriched GO-terms in Table 3.2, including "carboxylic

acid transmembrane transporter activity", "small molecule metabolic process" and "response to stimulus". It was part of the "transport" functional category in Figure 3.13. The reason that high expression of this maize gene correlates with GLS susceptibility is unclear.

Hsieh and Goodman (2005) studied heptahelical transmembrane proteins (HHP) in *Arabidopsis*. They found the expression of the HHP gene family to be differentially regulated by plant hormones, i.e. levels of HHP1 mRNA were increased by treatments with ABA and GA, whereas levels of HHP2 mRNA were increased by ABA and benzyladenine treatments. Kim *et al.* (2002) showed that the expression of a rice heptahelical plasma membrane-localized (MLO) gene was strongly induced by a fungal elicitor as well as by plant defense signalling molecules. They reported that it functions as a negative regulator of broad-spectrum disease resistance and leaf cell death; and further showed that MLO mediates defense modulation via direct Ca^{2+}-dependent interaction with calmodulin. According to Panstruga (2005), specific isoforms of the family of heptahelical MLO proteins in barley is required for successful host-cell invasion by the biotroph powdery mildew species, *Blumeria graminis* f. sp. *hordei*. Powdery mildew fungi appear to manipulate plant heptahelical MLO to regulate vesicle-associated processes at the plant cell periphery for successful pathogenesis. Expression levels of the HHP receptor in Tables 3.3 and 3.4 was high in plants with (i) high expression levels of GA-related genes (corresponding to the above-mentioned result from Hsieh and Goodman, 2005) as well as (ii) high levels of calcium/calmodulin signalling (corresponding to the result from Kim *et al.*, 2002). GA and calcium signalling were both over-represented processes in the greenyellow module. This gene appears to be associated with maize susceptiblity to *C. zeina* infection. One can speculate that *C. zeina* also manipulates maize HHP receptor for successful pathogenesis.

Nodulins are organ-specific plant proteins induced during symbiotic nitrogen fixation. Apart from genes involved in root nodule development, this gene family also includes recombination activation genes (RAGs) as well as specific sugar efflux transporters essential for plant nectar production, and plant seed and pollen development. The Pfam annotation for this nodulin MtN3 family protein is "sugar efflux transporter for intercellular exchange". Although the molecular function of these proteins is largely unknown, they are mostly transmembrane proteins and generally mediate glucose transport. Chen *et al.*

(2010*a*) identified a new class of sugar transporters, named SWEETs, which mediates glucose transport. Of the eleven nodulin MtN3 family proteins that were present in the WGCNA data set, Blast2GO annotated seven as bidirectional sugar transporter SWEET-like proteins. Two of these significantly correlated with GLS susceptibility and with GLS resistance. Two other nodulin MtN3 family proteins, not annotated as SWEET-like proteins by Blast2GO (including A_92_P015826 mentioned above), strongly correlated to GLS susceptibility (with correlation coefficients of 0.66 and 0.61) and belonged to the greenyellow module. However, these two genes were annotated with similar GO-terms than the previously mentioned SWEET-like proteins (including "carbohydrate transport" and "plasma membrane") and likely have similar functionality. Chen *et al.* (2010*a*) showed that bacterial symbionts as well as fungal and bacterial pathogens can induce the expression of different plant SWEET genes. This indicates that the sugar efflux function of SWEET transporters is likely targeted by pathogens for nutritional gain, which could also be the case for *C. zeina*-infected plants, since the pathogen switches to a necrotrophic growth habitat after first growing intercellularly in the leaf mesophyll where it would require plant nutrients.

NAC (NAM/ATAF/CUC) transcription factors have important functions in regulating plant growth, development, and abiotic and biotic stress responses. In *Arabidopsis*, the NAC transcription factor ATAF1 negatively regulates the defense response to necrotrophic fungi and bacterial pathogens (Wang *et al.*, 2009) and ATAF2 acts as a repressor of PR gene expression (Delessert *et al.*, 2005). Conversely, ANAC019 and ANAC055 are involved in the JA-dependent expression of defense genes in *Arabidopsis* (Bu *et al.*, 2008). In maize, ZmNAC41 and ZmNAC100 were identified to be transcriptionally induced both during the initial biotrophic as well as the ensuing necrotrophic colonization of maize leaves by the hemibiotrophic ascomycete fungus *C. graminicola* (Voitsik *et al.*, 2013). ZmNAC41 was present in the WGCNA data set and had a strong correlation of 0.5 to GLS susceptibility. In a promoter element analysis of six pathogen-induced maize NAC transcription factors, Voitsik *et al.* (2013) identified response elements for ERF, WRKY, TGA and NAC transcription factors in five of the six analysed genes, suggesting an involvement of these pathogen-induced NACs in the transcriptional network controlling the plant defense response. Three of the six pathogen-induced maize NACs mentioned above, including ZmNAC41, were present in the WGCNA data set: ZmNAC38

had a significant correlation to GLS susceptibility (0.25) and ZmNAC97 slightly correlated to GLS susceptibility (0.1). None of three mentioned NACs belonged to the greenyellow module. The greenyellow module included 12 transcription factors of which 6 were annotated with a "response to stress" GO-term: 2×WRKY, 2×AP2/ERF, 1×MYB and 1×NAC transcription factors (included in the "regulation of transcription" category in Figure 3.13).

Other potential driver genes in the greenyellow module

Table 3.3 contains a few regulatory genes that potentially modulate processes in the greenyellow module. Patatin/phospholipase (A_92_P040918; Table 3.3) contributed to enriched GO-terms such as "small molecule biosynthetic process", "response to stimulus", "lipid metabolic process" and "monocarboxylic acid metabolic process". Another reporter (A_92_P034498) representing the same maize gene was also part of the greenyellow module, thus confirming the expression profile and significance of this gene. Patatin/phospholipase is not only a storage protein, it also catalyses the cleavage of fatty acids from membrane lipids (Mignery *et al.*, 1988). Plants widely use phospholipid-based signal transduction to transfer the recognition of extracellular signals. Activation of phospholipases also initiates the production of defense signalling molecules, such as oxylipins and jasmonates (Canonne *et al.*, 2011). Camera *et al.* (2005) showed that two members of the patatin-like gene family (PLPs) are strongly induced in leaves challenged with fungal and bacterial pathogens in *Arabidopsis*. The accumulation of PLP2 in response to *Botrytis cinerea* or *Pseudomonas syringae* pv. tomato is dependent on JA and ET signalling, but is not dependent on SA. Their data indicate that PLP2-encoded lipolytic activity can be exploited by pathogens with different lifestyles to facilitate host colonisation.

The AP2 domain containing protein (A_92_P029518; Table 3.3) was part of the "regulation of transcription" functional category in Figure 3.13 and contributed to the "response to stimulus" enriched GO-term in Table 3.2. Pré *et al.* (2008) reported that the AP2/ERF domain transcription factor ORA59 in *Arabidopsis* integrates ET and JA signals in plant defense. Over-expression of ORA59 caused increased resistance against the fungus *Botrytis cinerea*, whereas ORA59-silenced plants were more susceptible. Several genes involved in the ET and JA biosynthesis were present in the greenyellow module. Specifically, ACC synthase (A_92_P039018; Table 3.3) and 3-ketoacyl-thiolase, key en-

zymes involved in ET and JA biosynthesis respectively, were part of the greenyellow module. However, a few other ET and JA biosynthesis and response genes were identified to significantly correlate with susceptibility and a few to resistance. It can be hypothesised that ET and JA signalling-related responses could be effective in conferring resistance to *C. zeina*, but occurred too late after infection in the susceptible response.

The WRKY transcription factor (A_92_P018873) in Table 3.3 contributed to the "response to chitin" enriched GO-term (Table 3.2). WRKY transcription factors are global regulators of host responses in reaction to pathogen challenge, typically via SA and JA signalling (Pandey and Somssich, 2009).

Other high GLS-correlating genes in the greenyellow module

The gene with the second highest correlation to GLS disease severity was a glutathione S-transferase (GST) (Table 3.4). GSTs are required for detoxification of lipid hydroperoxides that are generated during oxidative stress (Bhattacharjee, 2012). Although best known for their ability to detoxify cellular environments, GSTs are also capable of binding non-substrate ligands, with important cell signalling implications. Wisser *et al.* (2011) discovered high positive genetic correlations between GLS, Northern leaf blight and Southern leaf blight in a diverse panel of maize inbred lines and found a GST gene to be associated with modest levels of resistance to all three fungal diseases. They proposed that variability in detoxification pathways underlie natural variation in maize multiple disease resistance. Dean *et al.* (2005) highlighted that plants respond to foliar challenge by GST expression. They demonstrated that different GST genes respond in different ways to fungal infection and were the first to show a plant GST gene that plays a role in susceptibility to fungal infection: NbGSTU1, a *Nicotiana benthamiana* GST gene was in part responsible for susceptibility to fungal infection by *Colletotrichum destructivum* and *Colletotrichum orbiculare*. Two GSTs, highly correlated to GLS susceptibility, were present in the greenyellow module. Susceptible harvested samples had mature lesions (Figure 3.7), so these plants might be responding by detoxifying toxins produced by *C. zeina*.

Tonoplast dicarboxylate transporter (A_92_P012461; Table 3.3 and 3.4) contributed to many of the enriched GO-terms including "carboxylic acid transmembrane transporter activity", "dicarboxylic acid transport", "malate transport" and "C4-dicarboxylate trans-

port", as mentioned earlier. It can be hypothesised that this plant tonoplast dicarboxylate transporter and potentially other genes with similar functionality, could be up-regulated due to fungal manipulation, in order to benefit from manipulated stomatal aperture, hormone signalling and/or to gain nutrients.

Another gene with an exceptional GLS susceptibility correlation (Table 3.4) in the greenyellow module, is an EF hand / calmodulin-related calcium sensor protein (A_92_P029666). The greenyellow module also includes two other calmodulin-related calcium sensor proteins (A_92_P037035 and A_92_P040397), two calmodulin binding proteins (A_92_P021227 and A_92_P015623) and one calmodulin-dependent protein kinase (A_92_P006123). The calmodulin family is a major class of calcium sensor proteins with a key role in the regulation of numerous target proteins via cellular signalling cascades. It has been shown that pathogen infection causes significant ion fluxes across membranes in plants. These ion fluxes and specifically the activation of Ca^{2+} signalling pathways are vital for the activation of defense responses (Ranty *et al.*, 2006). Signal-specific changes in the cellular Ca^{2+} level were also found to function as a messenger in modulating diverse physiological processes that are important for stress adaptation (Reddy *et al.*, 2011). Due to various different defense responses that are activated in response to different stages of *C. zeina* infection, it is proposed that calcium signalling play a significant role in regulation of induced defense-related signalling cascades as well as plant adaptation to fungal attack. Figure 3.13 highlights calcium signalling as a key factor in the greenyellow module.

Other modules correlating with GLS susceptibility

The paleturquoise module correlated with GLS susceptibility, with a correlation coefficient of 0.31. It consisted of 41 reporters and no enriched GO-terms were detected using BiNGO. According to MapMan, a total of five reporters in this small module were annotated as transcriptional regulators and three as part of the "miscellaneous enzyme families" category. Table 3.7 gives the ten genes with the highest MM scores and Table 3.8 gives the ten genes with the highest GS scores. Two of the genes in Table 3.7, without a specific annotation from their rice and *Arabidopsis* best-matched homologous genes, were annotated by MapMan as putative transcriptional regulators (A_92_P026329 and A_92_P028020). One of the two, A_92_P028020, was also in

Table 3.8. Two other genes that were in both Tables 3.7 and 3.8, thus potentially regulating processes relating to GLS susceptibility in the paleturquoise module, were a citrate synthase (A_92_P027135) involved in the tricarboxylic acid (TCA) cycle (a central pathway in the production of energy from carbohydrates, fats, or proteins) and a cell wall hydroxyproline-rich glycoprotein (A_92_P027779). An additional cell wall-related gene, cellulose synthase (A_92_P025632), was also present in the paleturquoise module. In Table 3.8, the reporter with the second highest correlation with GLS susceptibility, was a CAS1 domain-containing protein (A_92_P030287), part of the O-methyltransferase enzyme family. An adenosine triphosphate (ATP) citrate lyase (A_92_P030448), an enzyme that represents an important step in fatty acid biosynthesis was also present in Table 3.8.

The blue module, consisting of 1,521 reporters, correlated with GLS susceptibility with a correlation coefficient of 0.22. Table 3.5 gives the enriched GO-terms of the blue module, according to BiNGO. Most of these GO-terms belonged to the cellular component ontology. The GO-term "intracellular membrane-bounded organelle" was one of the most significant GO-terms and more specific terms included "chloroplast" and "mitochondrial membrane". In the molecular function and biological process ontologies "aminopeptidase activity" and "cellular response to phosphate starvation" were enriched, respectively. For this module, there was no overlap between the top ten genes that best correlated with the blue module eigengene (Table 3.7) and the top ten genes that best correlated with GLS susceptibility (Table 3.8). However, due to the size of the module, when considering the top 100 genes in the two gene lists, seven genes occurred in both and are the best candidates to regulate GLS-related processes in the blue module: a citrate synthase (A_92_P041513; also mentioned in the paragraph above) and an aconitate hydratase (A_92_P039324) that both play a role in the tricarboxylic acid cycle; a mitogen-activated protein kinase kinase (A_92_P037761) involved in postranslational modification; a transport protein particle component (A_92_P035766) involved in vesicular trafficking; an ATP-binding cassette (ABC) transporter / ATP-binding protein (A_92_P037362) involved in the transport of diverse substrates across cell membranes; an ubiquinol-cytochrome C reductase iron-sulfur subunit (A_92_P037947) involved in mitochondrial electron transport and ATP synthesis; and a VHS domain containing protein (A_92_P041610) involved in vesicle transport in the cell. Energy metabolism, vesi-

cle transport and plasma membrane transport thus appear to be linked to the blue co-expression module and GLS susceptibility, where the above-mentioned genes seem to play key roles in the regulation of these procesesses.

The yellowgreen module correlated with GLS susceptibility with a correlation coefficient of 0.21 and consisted of 35 reporters. "Catalytic activity" was the only enriched GO-term for this module. According to MapMan, four genes belonged to the "miscellaneous enzyme families" category, two was involved in protein degradation and two in nucleotide metabolism. Three out of the five reporters that were shared between the top ten genes with the highest MM scores in the yellowgreen module (Table 3.7) and the top ten with the highest GS scores (Table 3.8), represented a single maize gene: a NUDIX hydrolase (A_92_P039569) that is involved in nucleotide metabolism, a protein transport protein SEC24-like (A_92_P035027) in the secretory pathway and a GST (A_92_P038512) involved in the cellular detoxification of both xenobiotic and endobiotic compounds.

The magenta module correlated with GLS susceptibility with a correlation coefficient of 0.2 and consisted of 266 reporters. Table 3.6 gives the enriched GO-terms of the magenta module. Most of these GO-terms belonged to the cellular component ontology, where "chloroplast thylakoid membrane" was one of the most significant GO-terms. Enriched GO-terms in the molecular function ontology included "binding", "unfolded protein binding" and "structural constituent of ribosome" and enriched GO-terms in the biological process ontology included "response to salt stress", "response to metal ion", "cellular protein metabolic process" and "photosynthesis". According to MapMan, thirteen genes in the magenta module were involved in regulation of transcription, of which four encoded C2H2 zinc finger family proteins and two auxin response factor family proteins. MapMan contributed to the BiNGO results by further highlighting and confirming processes of importance: nine genes encoded proteins that were involved in photosynthesis, seven in transport, seven in signalling, six in redox (reduction-oxidation) reactions, six in development, five in mitochondrial electron transport / ATP synthesis, five in amino acid metabolism, five in abiotic stress (heat stress response), five were miscellaneous enzyme families, four were involved in DNA synthesis / chromatin structure and four in cellular organisation. There was no overlap between the top ten genes with highest MM scores (Table 3.7) and the top ten genes with the highest GS scores (Table 3.8) in the magenta

module. However, when the top 30 genes in both lists were considered, four reporters overlapped of which two were annotated. These were the best candidates for regulating processes relating to GLS susceptibility in the magenta module: a mitogen-activated protein kinase 3 (A_92_P019536; in Table 3.8) involved in the MAP-kinase signalling pathway and a cysteine proteinase inhibitor (A_92_P001644). Apart from being involved in many aspects of plant physiology and development, as well as in the control of the programmed cell death or HR (Solomon *et al.*, 1999), cysteine proteases can also act as inhibitors of fungal plant pathogens (López-García *et al.*, 2012). Out of the ten genes with highest MM scores in the magenta module (potential drivers in Table 3.7), three were part of the mentioned MapMan categories: a plastid developmental DAG protein (A_92_P011228) acting in chloroplast development, a plastid-lipid-associated protein (A_92_P028770) involved in cellular organisation and a dehydroascorbate reductase (A_92_P012559) involved in redox control. The highest GLS-correlating gene in the magenta module (with a GS score of 0.52), was a flavanone 3-hydroxylase (A_92_P029634; Table 3.8), an enzyme involved in flavonoid biosynthesis. Two other flavonoid biosynthesis-related enzymes that significantly correlated with GLS susceptibility were also present in the magenta module. Out of the ten genes with highest GS scores (Table 3.8), two genes not yet referred to, were part of the above-mentioned MapMan categories: a chorismate mutase (A_92_P029669) involved in amino acid metabolism and a chloroplast post-illumination chlorophyll fluorescence increase protein (A_92_P042004) involved in photosynthesis and chlororespiration.

GO enrichment and overview of the turquoise module

The turquoise module (1,564 genes) was the co-expression module with the strongest negative correlation with GLS severity (correlation coefficient of -0.31). Several enriched GO-terms were detected for the turquoise module when the analysis was based on BiNGO's plant GO slim. These terms included nucleotide binding, protein binding, transport / transporter activity, cell growth, protein modification process and abscission (Table 3.9).

According to MapMan (Figure 3.15), 19% of the reporters in the turquoise module (271 reporters) were involved in protein degradation (8.1%), post-translational modification (5.4%), protein synthesis (2.5%), protein targeting (1.5%), and amino acid activation (0.9%). Fourteen percent of the reporters in the turquoise module (199 reporters)

were involved in RNA regulation of transcription (9.6%), RNA processing (2.6%), RNA binding (1.4%) and RNA transcription (0.4%). The third largest main category was transport (77 reporters; 5.4% of the turquoise module), which consisted of a variety of different transport-related genes. The largest sub-categories were ABC transporters and multidrug resistance systems (0.7%), metabolite transporters at the mitochondrial membrane (0.7%) and sugar transport (0.6%). G-protein signalling (1.8%), calcium signalling (0.8%) and receptor kinases (0.8%) were the largest sub-categories for signalling (61 reporters; 4.3% of the reporters in the turquoise module). Four percent of the reporters in the turquoise module (60 reporters) were involved in cellular organisation (1.8%), vesicle transport (0.9%) and other cell-related activities (1.4%).

Due to the large number of reporters in this module, GO enrichment was also permormed on subsets of reporters. The 400 reporters (26% of the reporters in the turquoise module) with the strongest positive correlation to the module eigengene (i.e. potential drivers) were enriched for localization, establishment of localization and transport. The 200 (13% of the reporters in the turquoise module) reporters with the strongest negative correlation to the GLS disease profile (i.e. correlation to GLS resistance) were enriched for chloride transport and chloride channel activity. Interestingly, GO-terms related to "transport" was enriched in both cases.

There was no overlap between the top twenty genes that best correlated with the turquoise module eigengene (Table 3.10) and the top twenty genes that best correlated with GLS resistance (Table 3.11). However, due to the size of the module, when considering the top 100 genes in the two gene lists, three genes occurred in both: a DNA binding protein (A_92_P006660) involved in regulation of transcription, a MIF4G domain containing protein (A_92_P005288) involved in protein synthesis and a DNA mismatch repair protein MutS (A_92_P011271) involved in DNA repair. These genes could be potential drivers of processes in the turquoise co-expression module that play a role in GLS resistance.

Potential driver genes in the turquoise module

The twenty top candidate "central nodes" that may serve as module drivers, are given in Table 3.10. The top candidate (A_92_P001413) encoded a coronatine-insensitive 1 (COI1), an F-box protein essential for all jasmonate responses. It interacts with multiple

proteins to form the SCF(COI1) E3 ubiquitin ligase complex and recruits jasmonate ZIM-domain (JAZ) proteins, negative regulators of JA signalling, for degradation by the 26S proteasome (Yan *et al.*, 2009). He *et al.* (2012) showed that COI1 regulate NB–LRR accumulation and function in *Arabidopsis*. They further demonstrated that apart from being involved in JA signalling-dependent disease resistance, COI1 also has a role in disease resistance independent of JA signalling. Melotto *et al.* (2006) showed that coronatine (COR), a compound made by *Pseudomonas syringae*, promotes stomatal reopening through the E3 ligase subunit COI1, which allows bacteria to enter. Since *C. zeina* hyphae also penetrate the leaf mesophyll via stomata (Beckman and Payne, 1983; Lyimo *et al.*, 2013), regulation of stomatal aperture play a key role. However, since high expression of this COI1 relates to GLS resistance, it seems unlikely that *C. zeina* produce compounds analogous to coronatine that alter stomatal aperture. Interestingly, a second gene encoding COI1 (with a GS score of -0.26) is also present in the turquoise module. It can be hypothesised that COI1 plays a key regulatory role in enhanced disease resistance against *C. zeina*.

COI1, together with three other genes in Table 3.10 were annotated as protein degradation-related genes: a zinc finger C3HC4 type domain containing protein (A_92_P007274), a signal peptide peptidase-like 2B (A_92_P006383) and an F-box/RNI-like superfamily protein (A_92_P006712). As mentioned above, 115 of the reporters in the turquoise module (8.1%) were involved in protein degradation. Of these, 74 were associated specifically with ubiquitin-dependent protein degradation. Ubiquitin, the ubiquitination system and the 26S proteasome play a key role in the regulation of plant immune response processes such as the oxidative burst, hormone signalling, gene induction, and programmed cell death (Trujillo and Shirasu, 2010). Without ubiquitin functioning properly, toxins from invading pathogens and other harmful molecules would increase dramatically due to weakened immune defenses.

Two phosphatases (A_92_P006813 and A_92_P005988) and one kinase (A_92_P002512) were also in the top 20 candidate drivers (Table 3.10). In signal transduction pathways, protein kinases modify other proteins by phosphorylation and phosphatases dephosphorylate proteins. Protein kinases and phosphatases play a key role in signalling mechanisms critical for responses to environmental stresses and attack by pathogens (Sessa and Martin, 2000). For example, mitogen-activated protein kinase (MAPK) pathways transfer

information from sensors to cellular responses. Mészáros *et al.* (2006) provided evidence that the MAP kinase kinase (MKK1) signalling pathway modulates the expression of genes responding to the bacterial elicitor flagelli and plays an important role in pathogen defense in *Arabidopsis*. Francia *et al.* (2011) showed that arbuscular-mycorrhizal (AM) fungal exudates activate MAP kinase cascades in plant cells. Here MAPK activation was dependent on the cytosolic Ca^{2+} increase. Yamaguchi *et al.* (2013) showed that fungal chitin recognition by CERK1, a chitin receptor, triggered rapid engagement of a rice MAP kinase cascade, which lead to defense response activation in rice. Specifically, in response to chitin, OsRLCK185 (a rice receptor-like cytoplasmic kinase) is directly phosphorylated by OsCERK1 at the plasma membrane. One can speculate that the phosphatases/kinases in the turquoise module are involved in post-translational modifications associated with MAPK cascades, which are activated in response to fungal elicitors. Interestingly, as mention above, 5.4% of the reporters in the turquoise module were involved in post-translational modifications. The phosphatases/kinases in Table 3.10 can be potential drivers in the form of post-transcriptional regulators.

GLS-correlating genes the turquoise module

The twenty genes in each module with the strongest negative correlation to GLS disease severity (i.e. highest correlation to GLS resistance) are given in Table 3.11. The gene with the highest correlation to GLS resistance (-0.53) in the turquoise module encoded a callose synthase (A_92_P010785). It belongs to the "minor CHO metabolism" category in Figure 3.15. Callose deposition in the form of local cell wall thickenings, called papillae, is a typical response of plants to fungal attack. Hinch and Clarke (1982) documented callose formation in maize roots as a response to infection with *Phytophthora cinnamomi*. In resistant *Arabidopsis* transgenic lines during powdery mildew infection, Ellinger *et al.* (2013) showed that haustoria formation was stopped due to expression of POWDERY MILDEW RESISTANT4 (PMR4), which encodes a stress-induced callose synthase under the control of the constitutive 35S promoter. They concluded that elevated early callose deposition resulted in complete penetration resistance to powdery mildew in *Arabidopsis*, so that activation of subsequent defense mechanisms was not needed. Our results showed that callose synthases generally correlated well with GLS resistance. Of the 10 genes encoding callose synthase proteins (from the input set of 19,281 reporters), five had a

significant negative correlation to GLS severity and one a significant positive correlation (Table 3.12). The top gene belonged to the turquoise module (the gene mentioned above) and three others to the yellow co-expression module.

Interestingly, phenylalanine ammonia-lyase (PAL) (A_92_P031017) and hydroxycinnamoyl-coenzyme A shikimate/quinate hydroxycinnamoyltransferase protein (A_92_P025879), which are involved in lignin biosynthesis, both had very strong positive correlations with GLS severity (0.51 and 0.5, respectively) in the turquoise module. It is thus apparent that lignification does not play a vital role in resistance against *C. zeina.* Both mentioned lignin biosynthesis-related genes also belonged to the turquoise module, due to their strong association with the turquoise module expression profile. The bulk of genes in the turquoise module (83%) had a negative correlation to GLS severity, but genes with a similar expression profile in the opposite direction were also included. There are cases where this can be biologically meaningful, for example when the same regulator repress or activate two different processes. As an example, Majello *et al.* (1997) reported that Sp3 is a dual-function transcription regulator with modular independent activation and repression domains. This gene's activity is dependent upon both the promoter and the cellular context.

The gene with the second best negative correlation to GLS severity in the turquoise module (-052), encoded an auxin response factor (A_92_P008134; Table 3.11). Auxin is an important plant hormone that regulates growth and development. It also regulates the plant's defense response by stimulating the degradation of transcriptional repressors (Bari and Jones, 2009). The turquoise module also contained two other auxin response factors (A_92_P008427 and A_92_P004482) with GLS severity correlations of -0.36 and -0.30, respectively, as well as a "suppressor of auxin resistance" protein (A_92_P026229) with an opposite expression profile (with a GLS severity correlation of 0.22). Interestingly, Navarro *et al.* (2006) found that induced auxin signalling promotes *Arabidopsis* susceptibility to the biotrophic bacterium *Pseudomonas syringae*, whereas Paponov *et al.* (2008) reported that repression of auxin signalling promotes susceptibility of *Arabidopsis* plants to the necrotrophic fungi *Plectosphaerella cucumerina* and *Botrytis cinerea.* Comparable to the latter, a high expression of these auxin response factors in our maize population in all probability promotes resistance to the hemibiotroph *C. zeina.* Furthermore, three auxin-related hormone metabolism genes (O-fucosyltransferase proteins) were also

present in the turquoise module: A_92_P006771, A_92_P004860 (in Table 3.10) and A_92_P008697.

The auxin response factor mentioned above (A_92_P008134), together with three other reporters in Table 3.11 are involved in regulation of transcription: a transcription termination factor nusG family protein (A_92_P012261), a ternary complex factor MIP1-like protein (A_92_P011441) and a basic helix-loop-helix DNA-binding protein (A_92_P015327). Also, two reporters are involved in RNA binding (A_92_P012893 and A_92_P004813; Table 3.11).

Other modules correlating with GLS resistance

The darkred module (63 genes) was the co-expression module with the second strongest negative correlation with GLS severity (correlation coefficient of -0.24). No GO-terms were enriched for the darkred module using BINGO. Protein degradation was the largest category according to MapMan, consisting of 11% of the reporters in the darkred module. There was no overlap between the top ten genes that best correlated with the darkred module eigengene (Table 3.10) and the top ten genes that best correlated with GLS resistance (Table 3.11). A good candidate driver gene could be "enhanced disease resistance 2 protein" (Table 3.10). According to Hiruma *et al.* (2011), *Arabidopsis* enhanced disease resistance 1 protein exerts an important positive role in resistance responses to hemibiotrophic/necrotrophic fungi, in part by inducing antifungal protein expression through depression of MYC2 function. MYC2, a basic helix-loop-helix leucine zipper transcription factor, represses a set of JA-responsive genes, while activating others. Interestingly, the two MYC2 transcription factors in the full data set (not in the darkred module) positively correlated with the GLS severity profile and the majority of the enhanced disease resistance proteins negatively correlated with GLS severity. Thus similar to what Hiruma *et al.* (2011) observed, this enhanced disease resistance protein was likely required for the induced expression of antifungal proteins. Furthermore, two protein degradation-related reporters, two protein targeting reporters, one reporter involved in post-translational modification and one in cell wall degradation, were present in Table 3.10: a ubiquitin-conjugating enzyme (A_92_P005017), a 26S proteasome non-ATPase regulatory subunit (A_92_P004214), a transportin (A_92_P004852), a vacuolar protein sorting-associated protein (A_92_P005333), a polyprenyltransferase

protein (A_92_P003665) and a pectin lyase-like protein (A_92_P004287). A serine carboxypeptidase-like gene was in the top ten best correlating genes to GLS resistance (-0.31) in the darkred co-expression module (Table 3.11). Serine carboxypeptidase is a wound-inducible gene product that functions in signal transduction (Li *et al.*, 2001). Liu *et al.* (2008) isolated a serine carboxypeptidase-like gene from rice, that was significantly up-regulated after treatments with benzothiadiazole, SA, JA and 1-amino cyclopropane-1-carboxylic acid (ACC). It was also up-regulated in incompatible interactions between rice and the blast fungus, *Magnaporthe grisea*. Liu *et al.* (2008) overexpressed the rice serine carboxypeptidase-like gene in transgenic plants to show enhanced resistance to *Pseudomonas syringae* pv. tomato and *Alternaria brassicicola*, as well as increased resistance to oxidative stress. Crampton *et al.* (2009) found that serine carboxypeptidase was one of the candidate genes that were significantly induced by SA treatment, but not up-regulated to the same extent by MeJA in pearl millet (*Pennisetum glaucu*m). It was also induced in response to a biotrophic rust pathogen, *Puccinia substriata*, in pearl millet. Apart from the serine carboxypeptidase-like gene mentioned above, a second gene in Table 3.11 was also related to protein degradation: a peptidase M50 family protein (A_92_P013508).

The yellow module (out of 1,170) had a correlation coefficient to GLS severity of -0.23. According to BINGO, no terms were enriched for the full module, but "nucleic acid binding" was enriched for the top 500 genes in the yellow module that best correlated with GLS resistance. According to MapMan, 108 reporters (10% of the genes in the yellow module) are involved in regulation of transcription, 77 reporters (7%) in protein degradation and 52 reporters (5%) in signalling (of which 22 are specifically involved in G-protein signalling). G-proteins were identified to be involved in signal transduction, induction of stomatal closure and defense responses (Zhang *et al.*, 2012*a*). Even though it is not part of the top ten candidate drivers (Table 3.10), an ethylene response factor (ERF), subfamily of AP2 transcription factor genes, with a MM score of 0.91 and GS score of -0.2, is a good candidate driver gene in the yellow module. According to Gutterson and Reuber (2004), the expression of several ERF genes is regulated by JA, SA and ET, as well as by pathogen challenge. The yellow module includes at least 30 genes involved in JA, ET and SA signalling according to the Blast2GO derived GO-terms of single genes. Allene oxide cyclase (A_92_P004273), with a strong correlation to GLS

resistance (-0.39) but also not in the top ten (Table 3.11), is an important enzyme involved in the JA biosynthesis pathway. Two WRKY transcription factors (with GLS severity correlations of 0.32 and -0.24, respectively) were also present in the yellow module. Two abiotic stress-related reporters that could be drivers of defense mechanisms, were present in Table 3.10: a universal stress protein domain containing protein (A_92_P001412) and a wound-responsive protein (A_92_P004097). Interestingly, 2% of the genes in the yellow module (21 reporters) were stress-related, of which 17 played a role in abiotic stress. Three reporters in Table 3.11 (best correlating with GLS resistance in the yellow module), were involved in regulation of transcription and contributed to the enriched GO-term "nucleic acid binding" mentioned above: a DNA binding domain containing protein (A_92_P002149), a DNA-binding storekeeper-related transcriptional regulator (A_92_P014539) and a homeobox-leucine zipper protein (A_92_P005598).

3.5 Conclusion

The expression patterns in genes and groups of genes across individuals in a *C. zeina*-infected maize RIL population were studied and coordinated responses to *C. zeina* infection were observed. Listed below is evidence in support of the observed coordinated responses. A genome-wide maize co-expression network identified 42 co-expression modules, eight of which were significantly associated with GLS susceptibility or resistance. Functional enrichment analyses of the resulting GLS-linked gene co-expression networks confirmed that the modules were biologically meaningful. In addition, specific genes and processes potentially contributing to GLS susceptibility and resistance were revealed.

The RIL population was sampled during flowering when GLS lesions were evident. Thus sampling was more suited to measure susceptible responses. A rapid response by the plant, typically conferring resistance, would more likely be detected at an earlier growth stage. Thus constitutively expressed genes that were differently expressed between the parental lines, rather than induced genes, were expected to be detected with the microarray analyses, since induced genes might have been activated earlier when the fungus started to invade and was likely switched off by the time of sampling (or turned on equally in susceptible and resistant plants at the later time point). This could explain the exceptionally high positive correlation to GLS severity (susceptibility) of the greenyellow

co-expression module, which was associated with a variety of enriched GO-terms. In order to standardise RNA sampling comparable areas on maize ear leaves of the same age, from plants of the same age, were sampled in the same field. A randomised block design with three replicate blocks was used to eliminate unwanted variation. Furthermore material was pooled, firstly from two plants per row and later from the three biological repeat blocks, allowing the measurement of the average expression for a RIL. A few limitations of this study were that (i) inoculum might be uneven in space and time, since the data came from a field trial; (ii) the observed gene expression levels were the average expression in and around lesions; and (iii) other diseases could influence the results (if genes involved in the GLS disease response were affected), however the Baynesfield site was dominated by GLS symptoms in this particular season (data not shown).

As mentioned above, one of the five co-expression modules that were identified to be related with GLS susceptibility had a remarkably high correlation coefficient of 0.71 (greenyellow co-expression module; p-value=$1e^{-16}$). Enriched processes/mechanisms in this co-expression module, included GA-signalling, response to stimulus and specifically chitin, transmembrane transporter activity, calcium signalling and protein degradation. Prominent processes in two other co-expression modules relating to GLS susceptibility, included (i) energy metabolism, vesicle transport and plasma membrane transport, and (ii) photosynthesis, flavonoid biosynthesis and abiotic stress. From this study it is apparent that up-regulation of GA and consequent down-regulation of DELLAs is associated with GLS susceptibility to *C. zeina* infection. Due to various different defense responses that are activated in response to *C. zeina* infection, one can speculate that calcium signalling play a significant role in regulation of induced defense-related signalling cascades as well as plant adaptation to fungal attack. Since susceptible harvested samples had mature lesions, these plants might be responding by detoxifying and degrading toxins produced by *C. zeina* and by up-regulating chitin-degrading enzymes. It can further be hypothesised that plant trans-membrane transporters could be up-regulated due to fungal manipulation, in order to benefit from altered stomatal aperture, hormone signalling or to gain nutrients.

Out of the three co-expression modules with negative correlations to GLS disease severity, i.e. positive correlation to resistance, the module with the strongest association to GLS resistance was linked with processes including post-translational modifications,

ubiquitin-mediated protein degradation and callose depositions. Another module relating with GLS resistance, was associated with protein degradation and G-protein signalling. From the results of this study, one can speculate that the phosphatases and kinases are involved in post-translational modifications associated with MAPK cascades, which are activated in response to fungal elicitors. It can be hypothesised that COI1 plays a key regulatory role in enhanced disease resistance against *C. zeina*. Ubiquitin-mediated protein degradation seem to act on toxins produced by *C. zeina* to strengthen immune defenses. One can speculate that callose depositions in the form of local cell wall thickenings is an effective response of maize to *C. zeina* infection. Lastly, G-protein signalling seems to play a role in improved GLS disease resistance.

The findings in this chapter were mostly hypotheses, which need to be validated with further studies. However, this chapter confirms that coordinated responses to *C. zeina* infection under field conditions in maize were observed. A major hypothesis that follows from this result is that there is a genetic basis for the observed coordinated responses. This will be focus of the next two chapters.

3.6 Acknowledgement of data contributions

This chapter is part of a collaborative project on the genomics of quantitative disease resistance in African maize varieties to *Cercospora zeina.* I would like to acknowledge the following people for contributing work or data to make the analysis of this chapter possible:

- Prof. Pangirayi Tongoona and his team from the University of KwaZulu-Natal for the field trials and GLS scoring.
- Prof. Dave Berger and members of the Molecular Plant-Pathogen Interactions laboratory for collecting maize leaf samples at Baynesfield Estate, KwaZulu-Natal.
- Dr. Bridget Crampton, Dr. Shane Murray and Ms. Jeanne Korsman for RNA extractions, microarray experiments and related lab work.
- A customised computer script for RIL selection for two colour microarray hybridization based on the method of Fu and Jansen (Fu and Jansen, 2006), the microarray data analysis including normalisation and back-conversion of the normalised data

was developed by a bioinformaticist (who preferred to stay anonymous) employed in the Maize eQTL project.

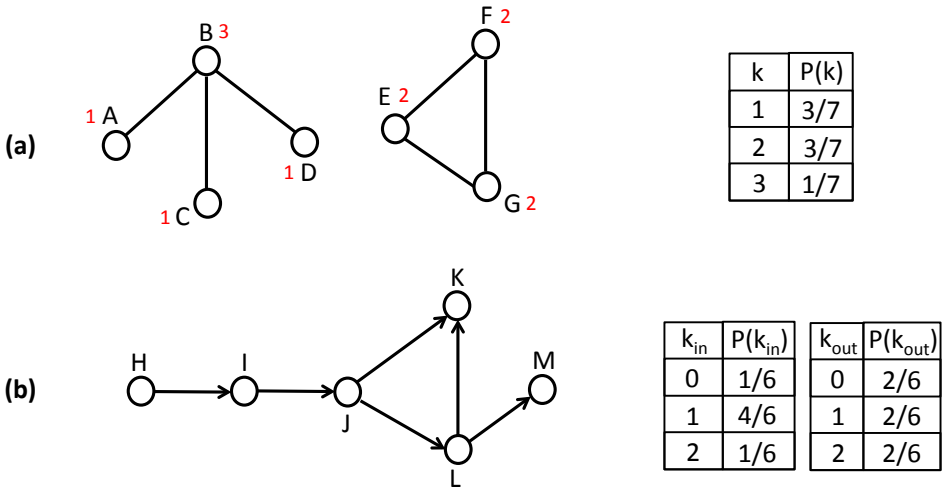


Figure 3.1: An illustration of how degree distributions are calculated. Adapted from Albert (2005). The number of interactions a node participates in is quantified by its degree, k (k_{in} and k_{out} in directed networks). The degree distribution $P(k)$ quantifies the fraction of nodes with degree k ($P(k_{in})$ and $P(k_{out})$ in directed networks). **(a)** This undirected and disconnected graph is composed of two connected components ABCD and EFG. The graph has degrees (k) ranging from 1 to 3, indicated in red, which was used to calculate the degree distribution $P(k)$. As an example, the degree of node B is 3, since it has links with 3 other nodes: A, C and D. The fraction of nodes with a degree of 3 is 1/7, since only node B (out of the 7 nodes in this graph) has a degree of 3. **(b)** This directed graph contains a source node H that can reach every other node in the network. Its out-component consists of the sink nodes M and K. The graph has in- and out-degrees ranging from 0 to 2 (not shown). As an example, node L has in-degree 1 and out-degree 2. Degree distributions $P(k_{in})$ and $P(k_{out})$ are given separately.

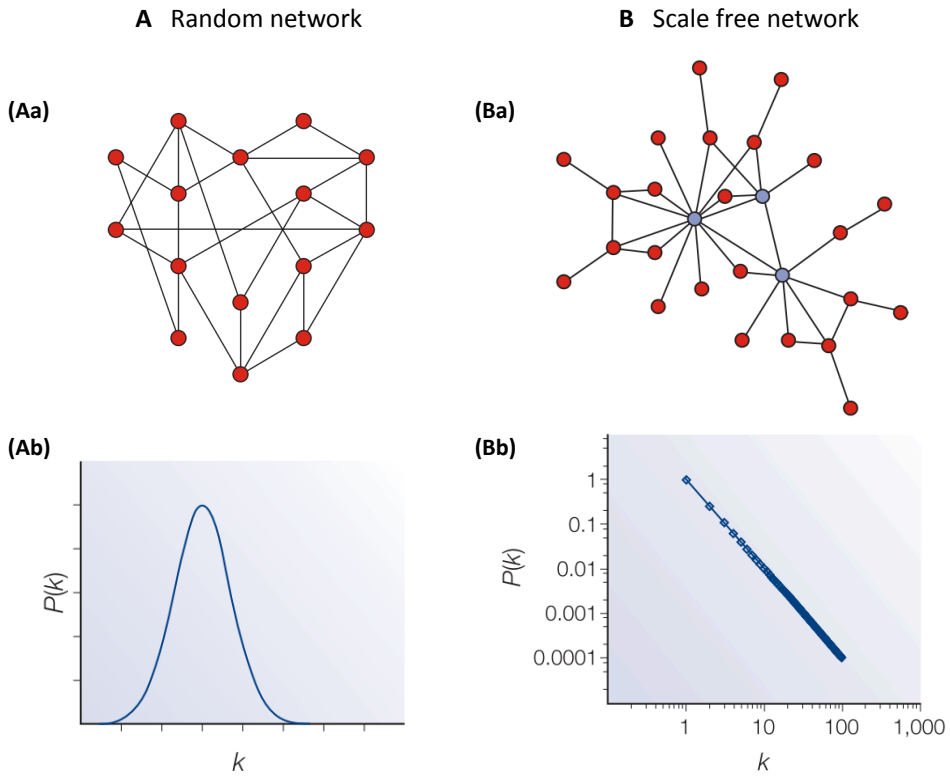


Figure 3.2: The difference between a random and a scale-free network. Adapted from Barabási and Oltvai (2004). **(Aa)** The random network is homogeneous: most nodes have approximately the same number of links. **(Ab)** The node degrees of a random network follow a Poisson distribution, which indicates that most nodes have approximately the same number of links (close to the average degree k). The x-axis represents the degree k and the y-axis the degree distribution $P(k)$. The tail (high k region) of the degree distribution $P(k)$ decreases exponentially, which indicates that nodes that significantly deviate from the average are extremely rare. **(Ba)** The scale-free network is inhomogeneous: the majority of the nodes have one or two links, but a few nodes have a large number of links. In a scale-free network, the probability that a node is highly connected is statistically more significant than in a random graph and the network's properties are often being determined by a relatively small number of highly connected nodes that are known as hubs (the blue nodes). **(Bb)** Scale-free networks are characterised by a power-law degree distribution; the probability that a node has k links follows $P(k) \sim k$–γ, where γ is the degree exponent. The degree exponent γ is usually in the range $2 < \gamma < 3$. Such distributions are seen as a straight line on a log-log plot, where the x-axis represents $log(k)$ and the y-axis $log(P(k))$.

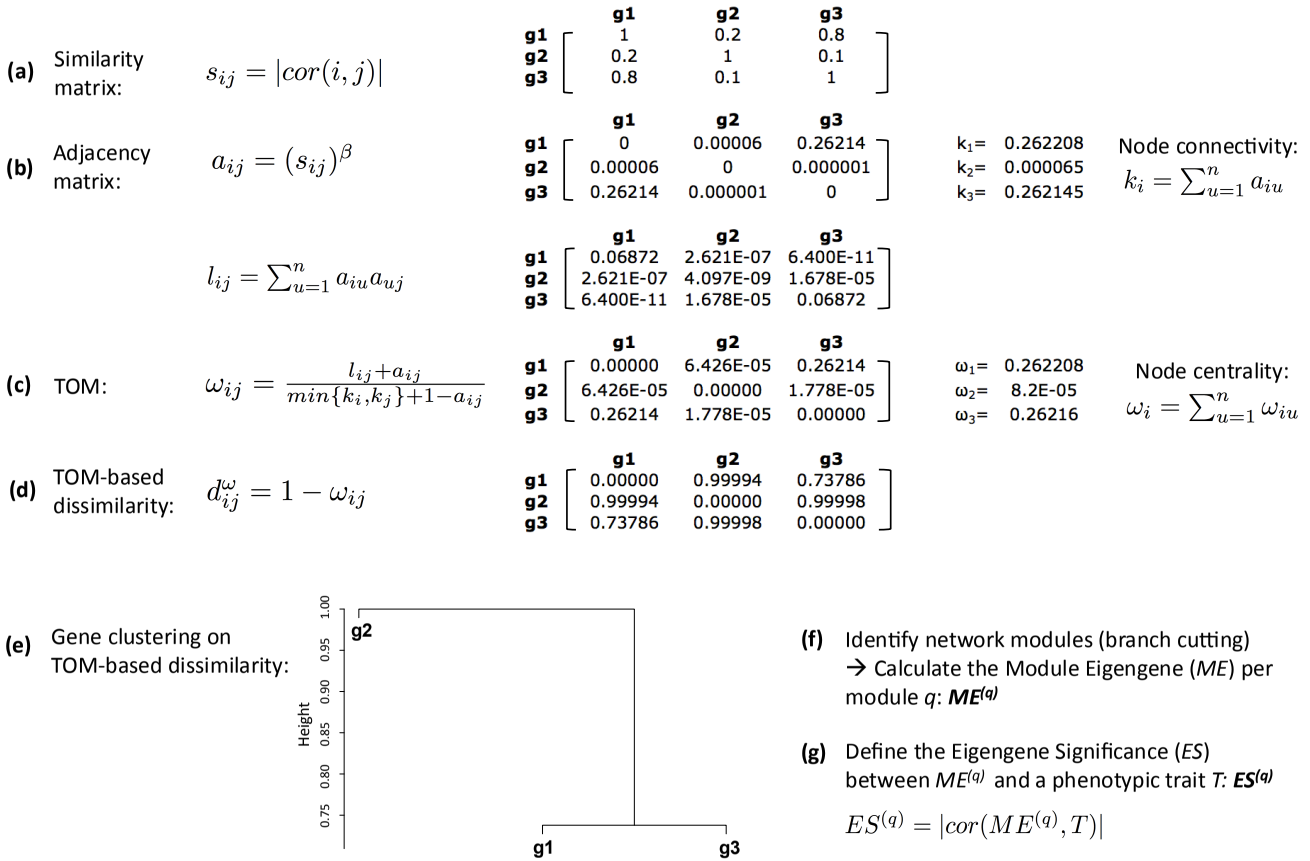


Figure 3.3: An example outlining the basic steps and calculations in WGCNA. Equations were extracted from Zhang and Horvath (2005). For simplicity, only 3 genes are used in steps (a) to (e) to demonstrate the calculations. **(a)** A similarity matrix is typically based on the Pearson correlation coefficient between gene expression profile pairs. **(b)** An adjacency matrix encodes the connection strengths between pairs of nodes. The *power* adjacency function is the default in WGCNA for transforming a similarity matrix, and the parameter β can be chosen so that it satisfies scale-free topology approximately. In this example $\beta = 6$ was used. Node connectivity, k_i, is defined as the sum of row i of the adjacency matrix. **(c)** The topological overlap matrix (TOM) provides a similarity measure, which reflects the relative interconnectedness of pairs of nodes. Node centrality, ω_i, is defined as the sum of row i of TOM. **(d)** TOM is transformed into a dissimilarity measure, by subtracting it from one. **(e)** Hierarchical clustering in WGCNA is based on the TOM-based dissimilarity distance measure. **(f)** Modules correspond to branches of the dendogram, and a module eigengene (ME) representing all the gene expression profiles in a module can be calculated for each module. **(g)** The correlation coefficient between the ME and an external trait (T) can be calculated as an eigengene significance score (ES) of each module.

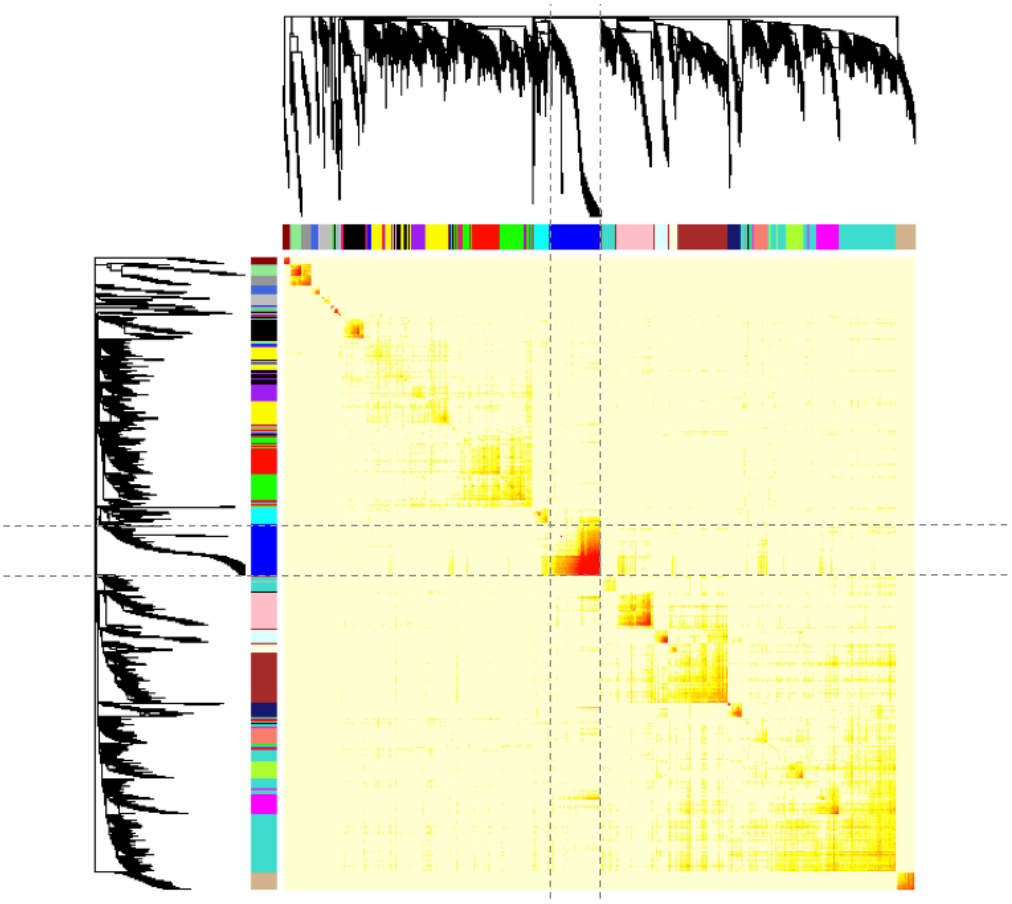


Figure 3.4: Topological Overlap Matrix (TOM) plot generated by Langfelder and Horvath (2012). A TOM plot is a color-coded heatmap of the values of the TOM-based dissimilarity matrix, where genes (in rows and columns) are sorted by the clustering tree and clusters correspond to dark squares along the diagonal. Light colour represents low overlap and increasingly darker red higher overlap. Dark coloured blocks along the diagonal correspond to gene co-expression modules. The "blue" gene co-expression module is highlighted with grey dashed lines. The gene dendrogram and module assignment are shown along the left side and the top.

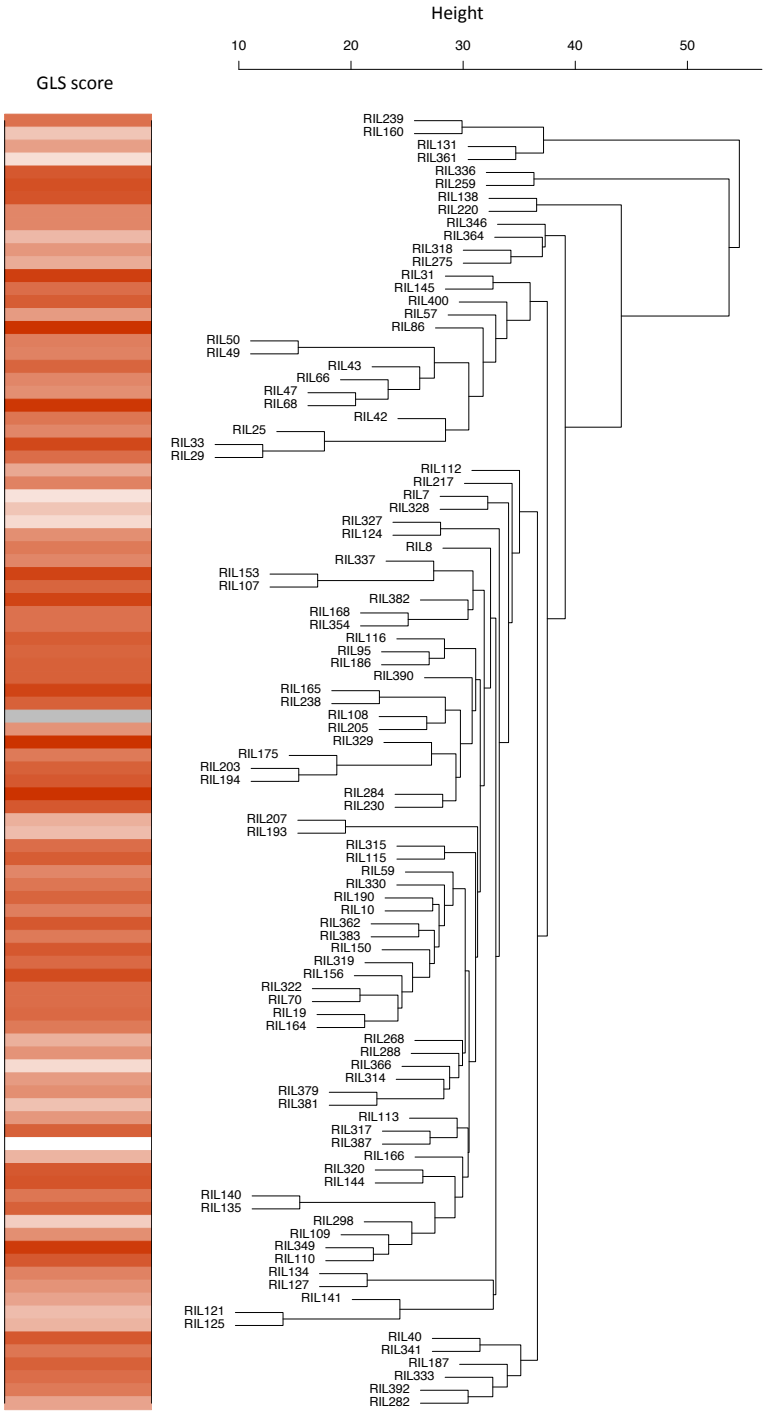


Figure 3.5: Sample clustering to detect outliers. A clustering dendrogram of the individuals in the RIL population was based on Euclidean distances across the reporter expression values on the microarray. The heatmap gives an indication of the GLS disease severity score per RIL: White indicates resistance (minimum score is 1); Red indicates susceptibility (maximum score is 9); Grey indicates a missing entry.

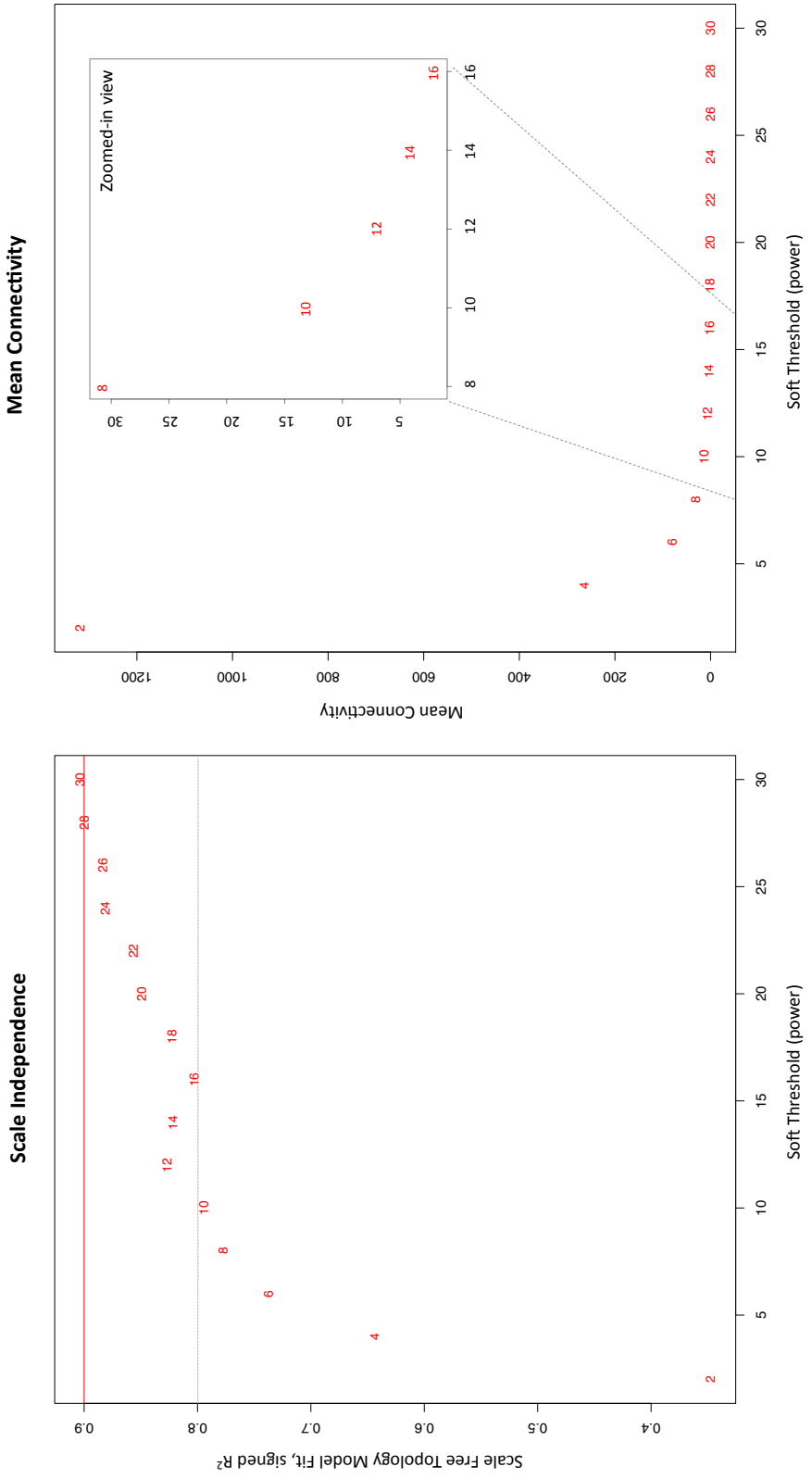


Figure 3.6: Choosing a soft-thresholding power. Analysis of network topology for various soft-thresholding powers. The left hand side graph shows the scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis). The red line indicates a very good fit of 0.9. The right hand side graph displays the mean connectivity (degree, y-axis) as a function of the soft-thresholding power (x-axis). The best estimate soft-thresholding power, is the power with the highest mean connectivity, but for which the scale-free topology fit index is higher than 0.8. For this analysis, a soft-thresholding power of 12 was chosen.

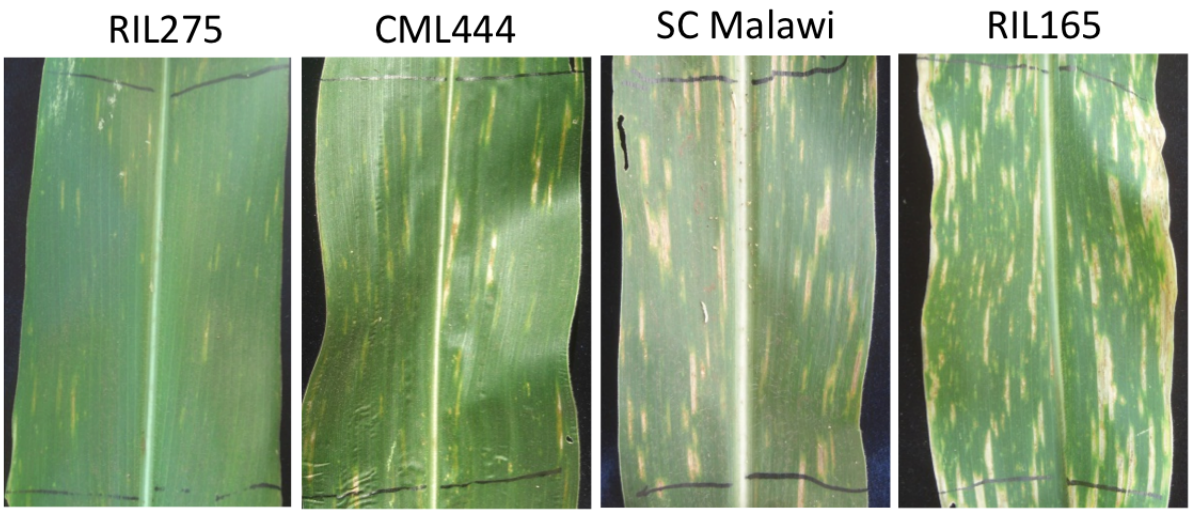


Figure 3.7: Grey leaf spot on maize in the field, caused by *Cercospora zeina*. Symptoms are necrotic lesions on the leaf surface. The range of differences in levels of lesions indicates that there is large variation in disease resistance/susceptibility for *C. zeina* in this population. CML444 was the more resistant parent and SC Malawi the more susceptible parent of the RIL population. Transgressive segregation was observed in this population, since some RILs showed less and other more GLS disease symptoms compared to the parental lines.

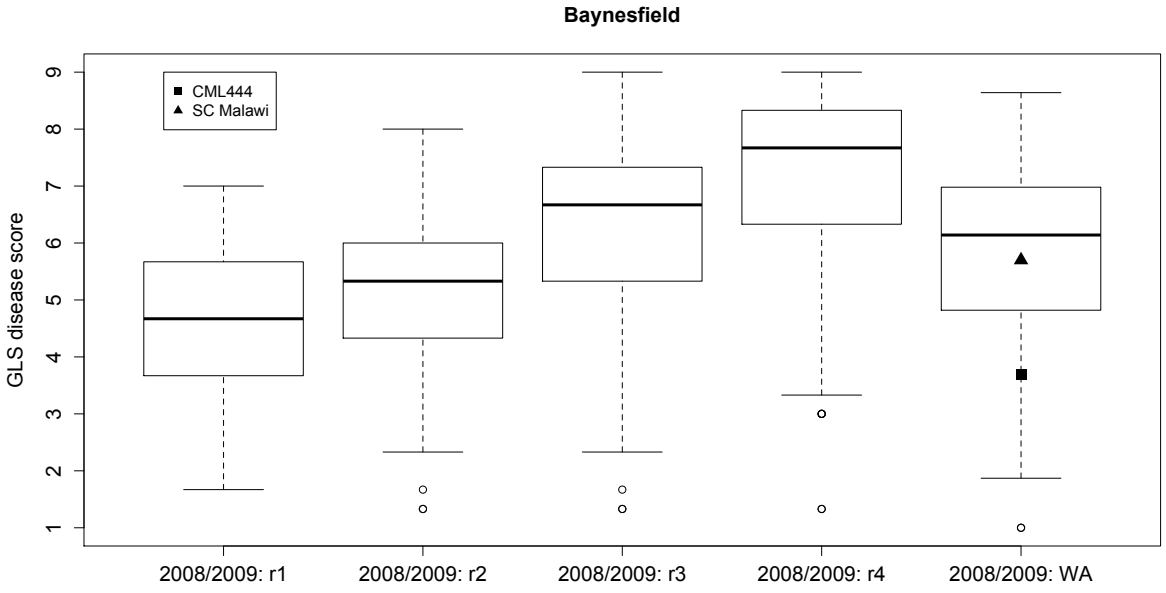


Figure 3.8: Boxplots of GLS disease severity data (y-axis), collected at 92, 99, 109 and 116 days after planting ($r1 - r4$) at Baynesfield Estate in KwaZulu-Natal, South Africa. The last boxplot represents a weighted average (WA) of the GLS severity scores across the four ratings for the 2008/2009 season, depending on the number of days between the rating and the day of sampling.

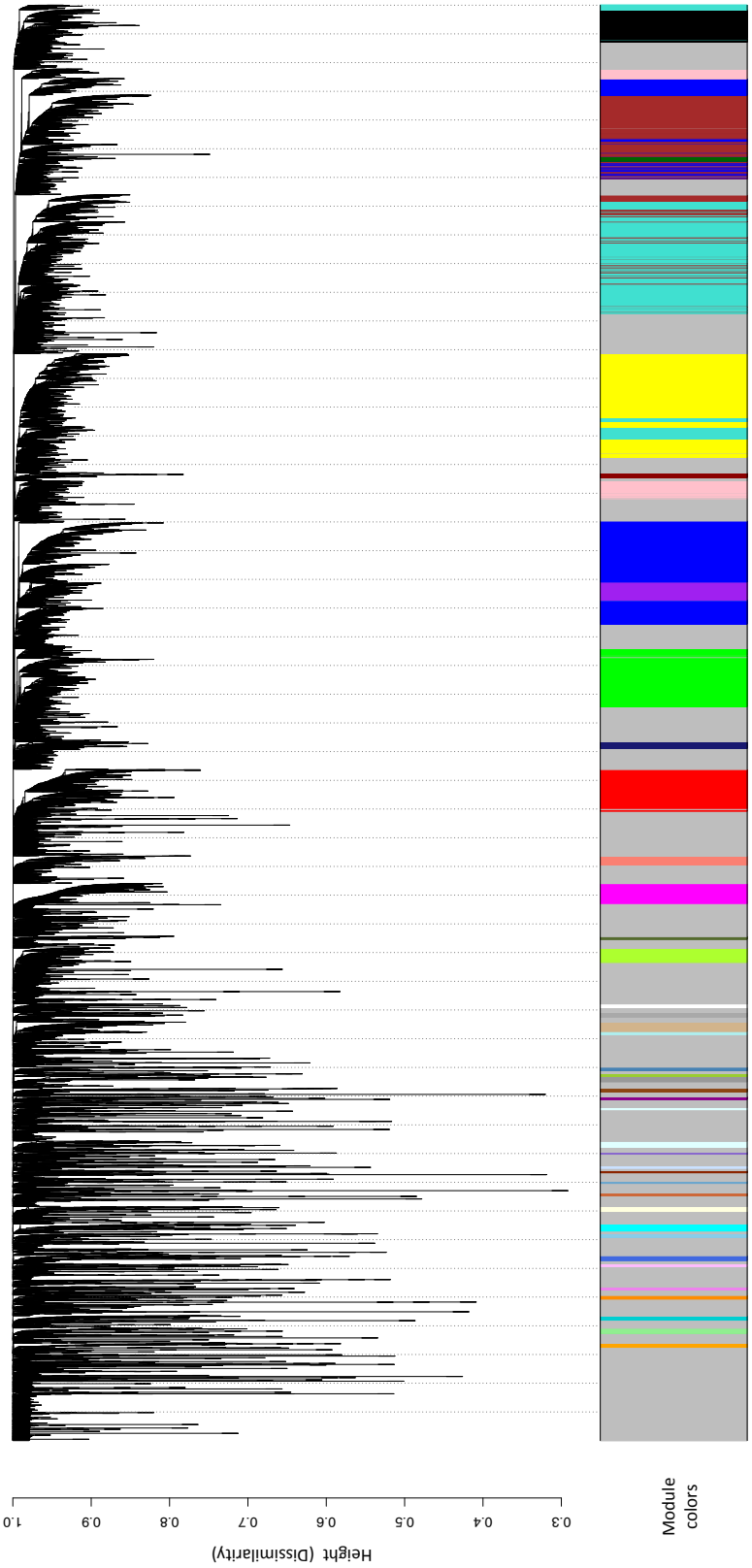


Figure 3.9: Gene dendogram and module colours. Clustering dendrogram of genes, with dissimilarity (y-axis) based on topological overlap, together with assigned module colours (x-axis). Modules were identified with a branch cutting method and very similar expression profiles were merged. Genes not belonging to any module are grey.

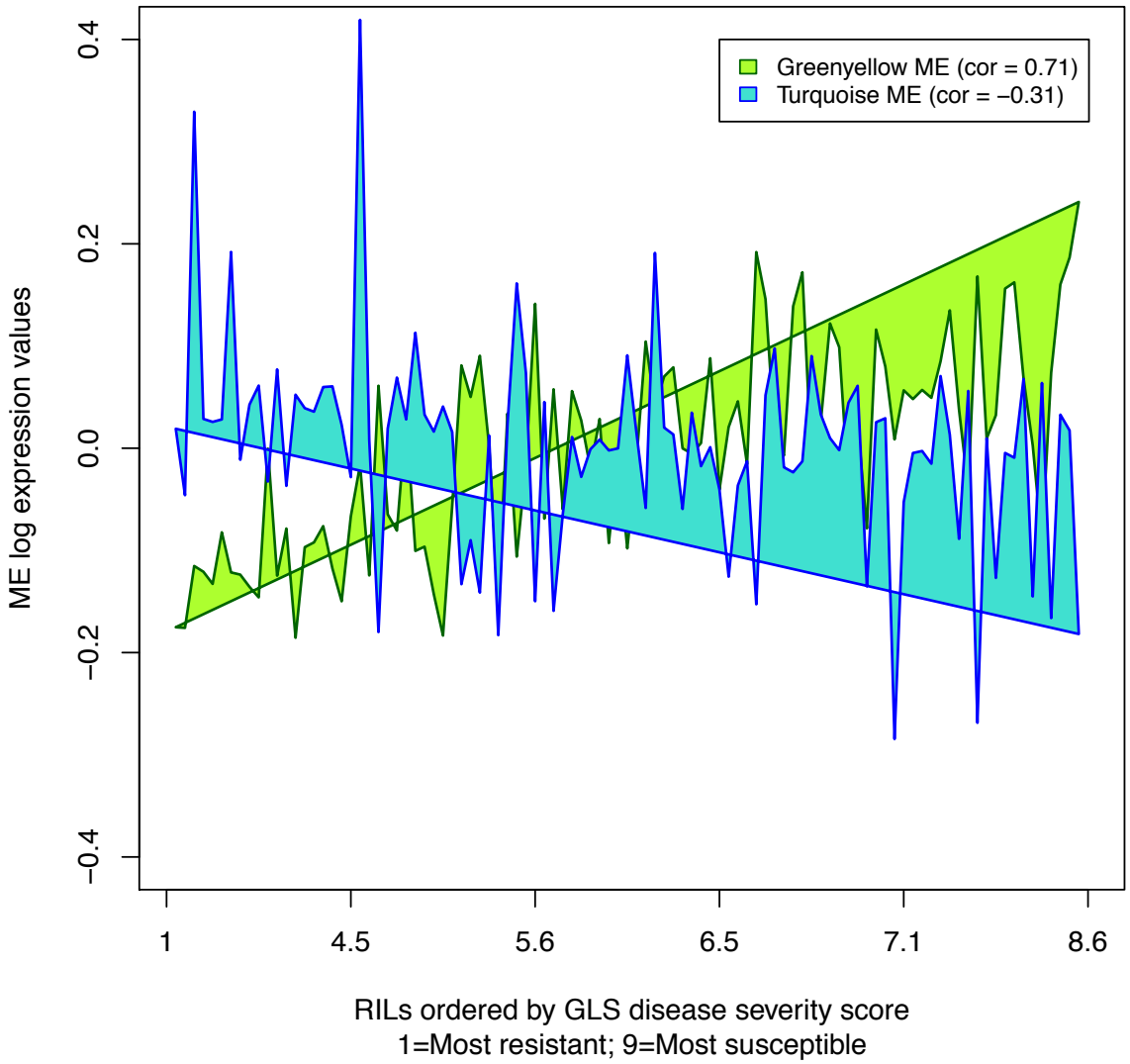


Figure 3.10: Module eigengene (ME) expression values (y-axis) across the RILs (x-axis; sorted by GLS severity scores) for the greenyellow and turquoise modules. A positive correlation indicates that RILs with high ME expression values, also have high disease severity scores (susceptible RILs), and RILs with low ME expression values also have a low disease severity scores (resistant RILs). Genes in the greenyellow module are positively correlated with GLS severity across the RILs. A negative correlation indicates that RILs with high ME expression values have low disease severity scores (resistant RILs), whereas RILs with low ME expression values have high disease severity scores (susceptible RILs). Genes in the turquoise module are negatively correlated with GLS severity across the RILs.

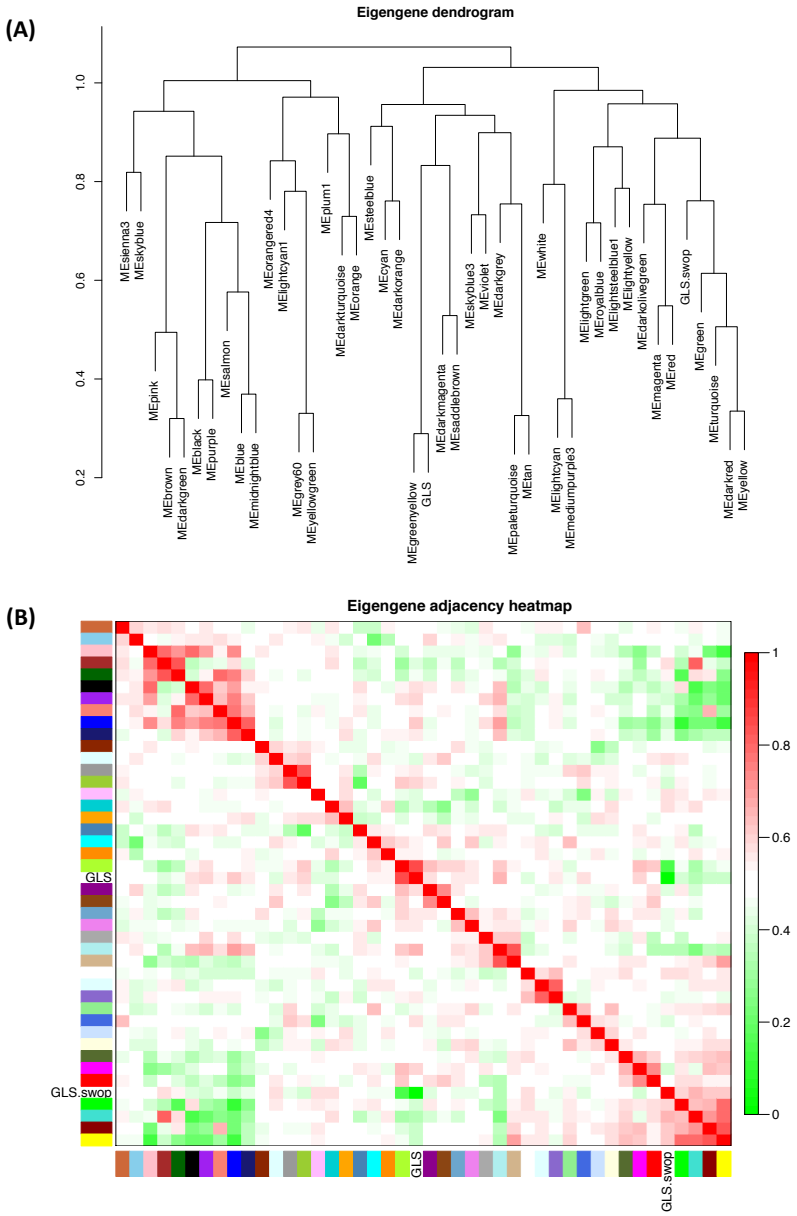


Figure 3.11: The co-expression module eigengene dendogram and adjacency heatmap represent the relationships among the co-expression modules. "GLS" refers to the normal GLS severity profile, where a positive correlation indicates that higher expression correlates with susceptible RILs. For convenience "GLS swop" was also calculated, so that a positive correlation indicates that higher expression correlates with resistant RILs. **(A)** A hierarchical clustering dendrogram of the eigengenes in which the dissimilarity of eigengenes E_I and E_J, is given by $1-cor(E_I, E_J)$. **(B)** The heatmap shows the eigengene adjacency, $A_{IJ} = (1+cor(E_I, E_J))/2$, which preserves the sign of the correlation. Within the heatmap, red indicates high adjacency (i.e. 1 represents a strong positive correlation) and green low adjacency (i.e. 0 represents a strong negative correlation), as shown by the color legend.

Figure 3.12: Co-expression module-GLS disease relationships across 17 field trials. Each row corresponds to a co-expression module eigengene (y-axis on the left-hand side) and each column to a GLS disease trait (x-axis). The rows were sorted according to the correlation coefficients of the first column (B_09_GLS), which was the focus of this study (the weighted average of ratings $1-4$ at Baynesfield season 2008/2009). The other columns represent GLS disease traits from different seasons scored at different locations. The columns were sorted according to location (indicated by the first letter of the name): Baynesfield, Greytown, Cedara and Ukalinga. Each cell contains the corresponding correlation and p-value. The table is colour-coded by correlation coefficient significance, according to the colour legend on the right-hand side y-axis.

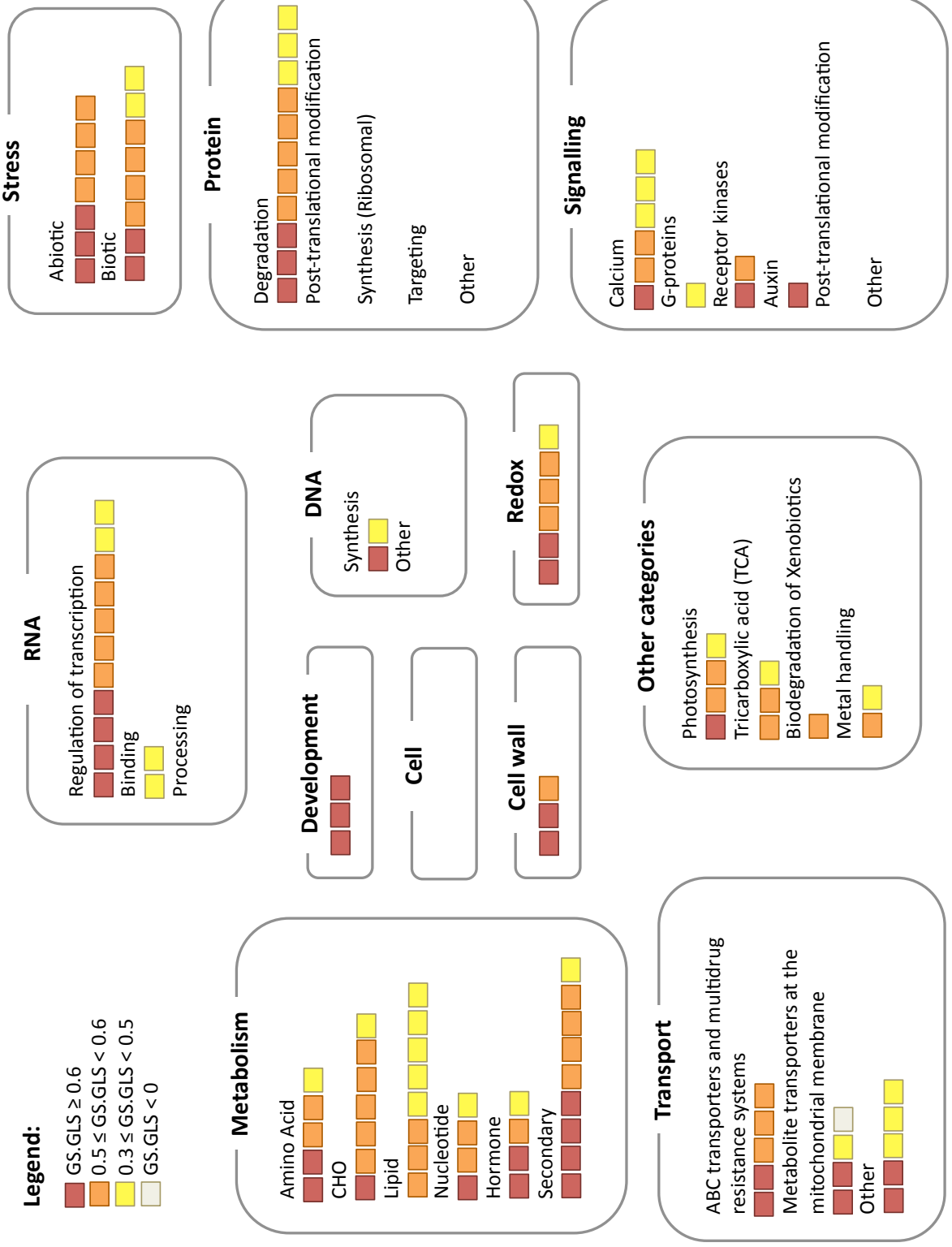


Figure 3.13: An overview of the functional categories that are associated with the 185 reporters in the greenyellow module. MapMan BINs were used as a basis to group the genes into categories. Each block represents a maize gene model. The colour of each block represents the gene significance with the trait (GS.GLS). Red indicates a very high positive correlation with GLS severity: reporters with correlation coefficients above 0.6. Orange indicates reporters with correlation coefficients between 0.5 and 0.6. Yellow indicates reporters with correlation coefficients between 0.3 and 0.5. White indicates reporters with a negative correlation to GLS severity.

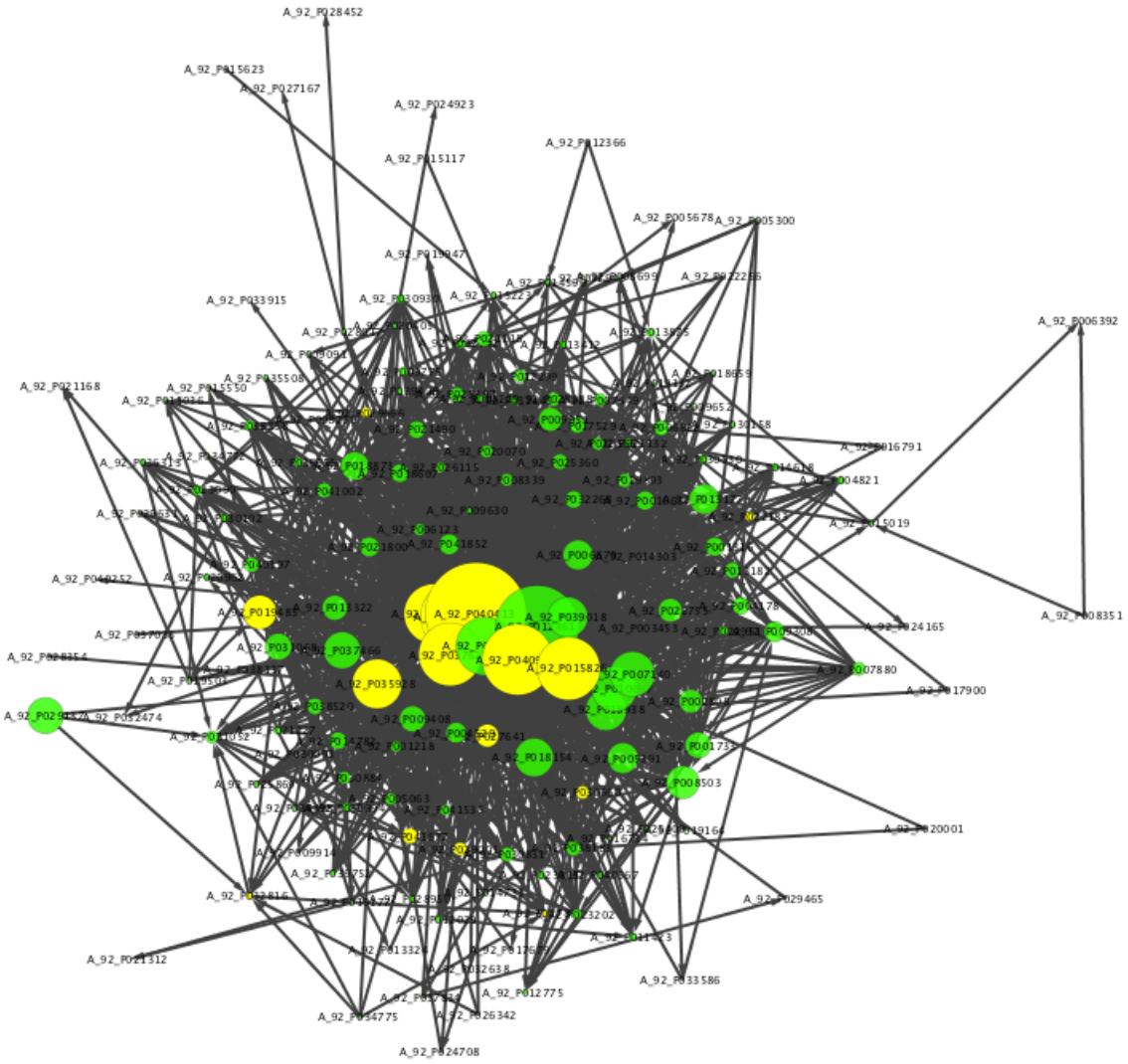


Figure 3.14: Network representation of the greenyellow module. Nodes represent genes and edges represent co-expression. The network layout is force-directed and node size corresponds to node degree. Yellow indicate nodes/genes that are highly correlated to the GLS severity profile (genes with a GS.GLS score > 0.65 in Table 3.4) and green indicate genes with correlation coefficients < 0.65 in the greenyellow module. This figure shows that the bulk of potential driver genes in the greenyellow module (the largest nodes), were also the genes that best correlated with GLS severity (yellow nodes). Table 3.3 lists the top 35 driver genes in the greenyellow module and Table 3.4 the top 35 GLS severity-correlating genes.

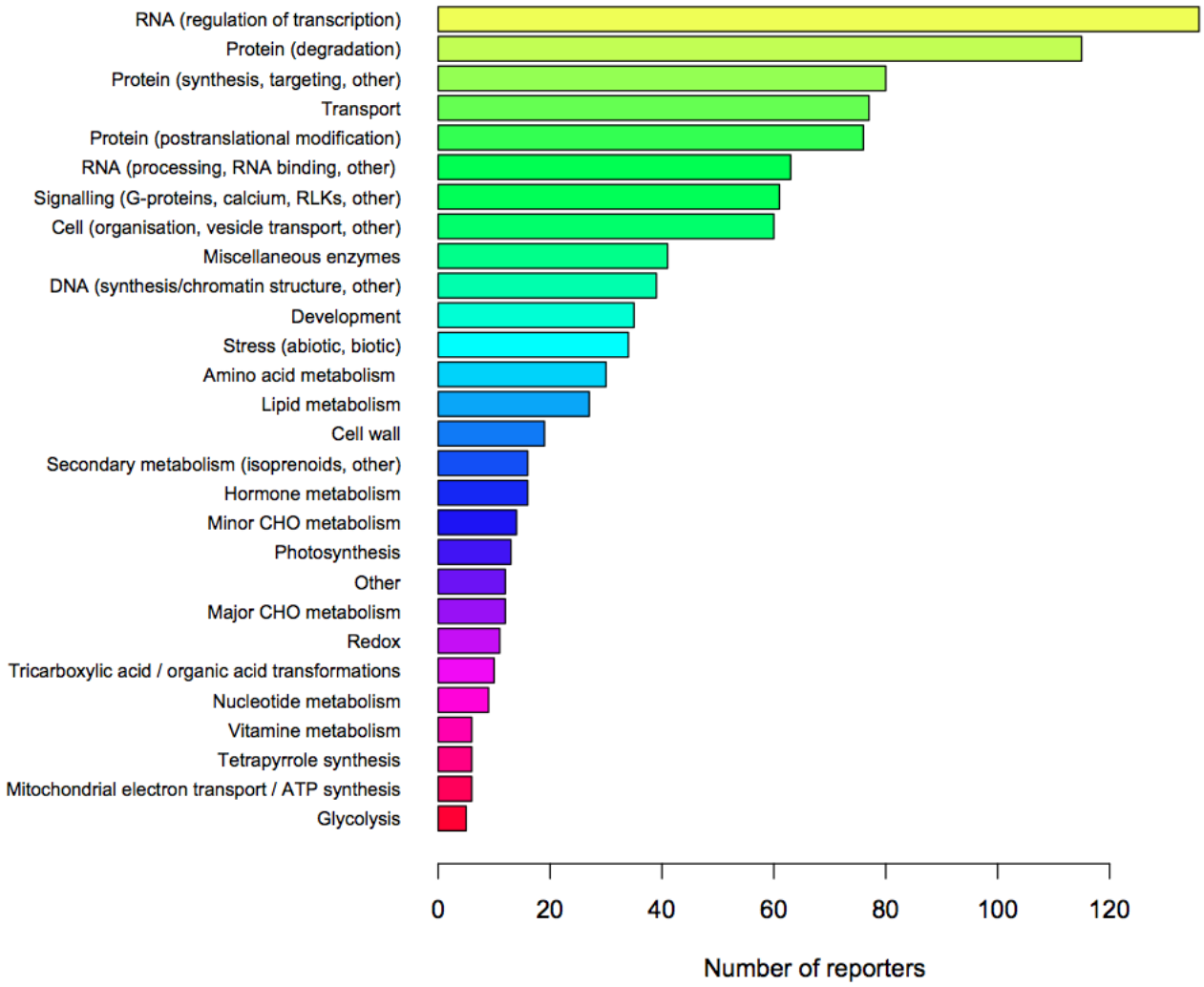


Figure 3.15: Summary of the MapMan categories in the Turquoise module. A total of 1,029 reporters (out of the 1,564 reporters in the turquoise) were assigned to specific MapMan categories.

Table 3.1: Module eigengenes were correlated with the GLS severity scores, using Pearson correlation. Eight co-expression modules were identified to be significantly associated with GLS severity (p-value < 0.05).

Module	Correlation between ME and GLS severity score	p-value[a]	Number of reporters in module	High expression of reporters in module correlates to higher GLS severity scores (H) or lower GLS severity scores (L)
Greenyellow	0.71	1.00E-16	185	H
Turquoise	-0.31	0.002	1564	L
Paleturquoise	0.31	0.002	41	H
Darkred	-0.24	0.02	63	L
Yellow	-0.23	0.02	1170	L
Blue	0.22	0.03	1521	H
Yellowgreen	0.21	0.04	35	H
Magenta	0.2	0.05	266	H

[a] The null hypothesis is that there is no correlation between the ME expression values and GLS severity scores across the RILs

Table 3.2: Enriched GO-terms for the Greenyellow module based on the full GO ontology using BiNGO.

GO-ID	Description	p-val	Corr p-val	Cluster freq	Total freq
16101	diterpenoid metabolic process	2.44E-07	1.07E-04	4/132 3.0%	5/8758 0.0%
9685	gibberellin metabolic process	2.44E-07	1.07E-04	4/132 3.0%	5/8758 0.0%
3824	catalytic activity	1.77E-06	5.18E-04	73/132 55.3%	3091/8758 35.2%
16102	diterpenoid biosynthetic process	1.32E-05	2.33E-03	3/132 2.2%	4/8758 0.0%
9686	gibberellin biosynthetic process	1.32E-05	2.33E-03	3/132 2.2%	4/8758 0.0%
19748	secondary metabolic process	1.98E-05	2.90E-03	9/132 6.8%	101/8758 1.1%
46943	carboxylic acid transmembrane transporter activity	4.46E-05	4.90E-03	5/132 3.7%	27/8758 0.3%
5342	organic acid transmembrane transporter activity	4.46E-05	4.90E-03	5/132 3.7%	27/8758 0.3%
44248	cellular catabolic process	6.33E-05	6.05E-03	13/132 9.8%	240/8758 2.7%
44281	small molecule metabolic process	7.65E-05	6.05E-03	23/132 17.4%	641/8758 7.3%
50896	response to stimulus	8.10E-05	6.05E-03	37/132 28.0%	1316/8758 15.0%
46942	carboxylic acid transport	8.94E-05	6.05E-03	5/132 3.7%	31/8758 0.3%
15849	organic acid transport	8.94E-05	6.05E-03	5/132 3.7%	31/8758 0.3%
44270	cellular nitrogen compound catabolic process	1.28E-04	8.04E-03	4/132 3.0%	18/8758 0.2%
6629	lipid metabolic process	1.66E-04	9.72E-03	13/132 9.8%	264/8758 3.0%
46700	heterocycle catabolic process	1.98E-04	1.09E-02	4/132 3.0%	20/8758 0.2%
32787	monocarboxylic acid metabolic process	2.40E-04	1.24E-02	9/132 6.8%	139/8758 1.5%
9740	gibberellic acid mediated signaling pathway	5.05E-04	2.34E-02	3/132 2.2%	11/8758 0.1%
71370	cellular response to gibberellin stimulus	5.05E-04	2.34E-02	3/132 2.2%	11/8758 0.1%
10476	gibberellin mediated signaling pathway	6.67E-04	2.56E-02	3/132 2.2%	12/8758 0.1%
15140	malate transmembrane transporter activity	6.70E-04	2.56E-02	2/132 1.5%	3/8758 0.0%
5310	dicarboxylic acid transmembrane transporter activity	6.70E-04	2.56E-02	2/132 1.5%	3/8758 0.0%
15556	C4-dicarboxylate transmembrane transporter activity	6.70E-04	2.56E-02	2/132 1.5%	3/8758 0.0%
9056	catabolic process	8.25E-04	2.99E-02	13/132 9.8%	312/8758 3.5%
44283	small molecule biosynthetic process	8.50E-04	2.99E-02	13/132 9.8%	313/8758 3.5%
16298	lipase activity	1.13E-03	3.53E-02	4/132 3.0%	31/8758 0.3%
19752	carboxylic acid metabolic process	1.17E-03	3.53E-02	13/132 9.8%	324/8758 3.6%
43436	oxoacid metabolic process	1.17E-03	3.53E-02	13/132 9.8%	324/8758 3.6%
6082	organic acid metabolic process	1.20E-03	3.53E-02	13/132 9.8%	325/8758 3.7%
6835	dicarboxylic acid transport	1.33E-03	3.53E-02	2/132 1.5%	4/8758 0.0%
43090	amino acid import	1.33E-03	3.53E-02	2/132 1.5%	4/8758 0.0%
15743	malate transport	1.33E-03	3.53E-02	2/132 1.5%	4/8758 0.0%
15740	C4-dicarboxylate transport	1.33E-03	3.53E-02	2/132 1.5%	4/8758 0.0%
42180	cellular ketone metabolic process	1.42E-03	3.66E-02	13/132 9.8%	331/8758 3.7%
6721	terpenoid metabolic process	1.80E-03	4.39E-02	4/132 3.0%	35/8758 0.3%
10200	response to chitin	1.80E-03	4.39E-02	4/132 3.0%	35/8758 0.3%
4190	aspartic-type endopeptidase activity	1.95E-03	4.51E-02	3/132 2.2%	17/8758 0.1%
70001	aspartic-type peptidase activity	1.95E-03	4.51E-02	3/132 2.2%	17/8758 0.1%
10033	response to organic substance	2.07E-03	4.67E-02	14/132 10.6%	388/8758 4.4%

Table 3.3: The 35 best potential drivers in the greenyellow module. Reporters were sorted by MM scores (decreasing).

Agilent.ID	Annotation (AT, Rice, B2G)[a]	Degree[b]	MM.[c] green-yellow	p.MM. green-yellow	GS.[d] GLS	p.GS. GLS	GS and MM scores in top 35[e]
A_92_P040413	Carnitine acylcarnitine translocase	129	0.94	8.0E-49	0.67	1.6E-14	*
A_92_P012461	Tonoplast dicarboxylate transporter	107	0.94	2.6E-48	0.62	4.2E-12	*
A_92_P015826	Nodulin MtN3 family protein	81	0.92	3.4E-41	0.66	4.6E-14	*
A_92_P040918	Patatin group A-3-like / phospholipase A 2A	91	0.92	9.6E-41	0.65	2.2E-13	*
A_92_P035171	Heptahelical transmembrane protein receptor	77	0.91	9.8E-39	0.67	4.2E-14	*
A_92_P032766	NAC transcription factor	90	0.91	2.6E-38	0.66	6.4E-14	*
A_92_P037621	F-box kelch-repeat protein skip11-like	81	0.90	1.5E-37	0.68	5.5E-15	*
A_92_P029518	AP2 domain containing protein	75	0.90	1.6E-37	0.63	1.5E-12	*
A_92_P019483	Protein of unknown function	44	0.89	3.2E-35	0.65	2.1E-13	*
A_92_P018873	WRKY transcription factor	37	0.89	7.6E-35	0.64	1.0E-12	*
A_92_P018154	Cytokinin-o-glucosyltransferase 2	50	0.89	1.0E-34	0.57	7.9E-10	
A_92_P018938	CHY zinc finger family	53	0.88	4.7E-34	0.63	2.2E-12	*
A_92_P013177	Protein of unknown function	39	0.88	4.4E-33	0.62	4.7E-12	*
A_92_P006679	Bowman-birk type trypsin inhibitor	39	0.88	5.3E-33	0.62	4.7E-12	*
A_92_P022755	Chitinase	31	0.88	6.2E-33	0.62	4.7E-12	*
A_92_P010667	Branched-chain-amino-acid aminotransferase	58	0.88	6.5E-33	0.52	2.4E-08	
A_92_P007140	Laccase 7	58	0.88	7.0E-33	0.59	6.8E-11	
A_92_P013322	Glycosyltransferase family 61 protein	33	0.86	4.9E-31	0.56	1.2E-09	
A_92_P025360	.	21	0.86	5.6E-31	0.58	3.3E-10	
A_92_P018607	PfkB-like carbohydrate kinase family protein	24	0.86	1.2E-30	0.52	3.1E-08	
A_92_P021118	Serine-type endopeptidase inhibitors	20	0.86	2.5E-30	0.62	6.5E-12	
A_92_P009408	COP1-interacting protein 7-like protein	34	0.86	8.9E-30	0.52	2.5E-08	
A_92_P039018	ACC synthase	52	0.85	2.2E-29	0.61	2.0E-11	
A_92_P017529	Secondary cell wall-related glycosyltransferase	21	0.85	5.7E-29	0.60	5.6E-11	
A_92_P008503	Chaperone DnaJ-domain superfamily protein	43	0.85	8.8E-29	0.62	5.5E-12	*
A_92_P025397	.	20	0.85	1.3E-28	0.56	1.2E-09	
A_92_P027641	.	30	0.85	1.5E-28	0.68	8.4E-15	*
A_92_P029001	Polyphenol oxidase	18	0.84	3.2E-28	0.66	6.1E-14	*
A_92_P022183	Lactate dehydrogenase	13	0.84	6.7E-28	0.66	1.2E-13	*
A_92_P023090	Cytochrome P450	14	0.84	8.0E-28	0.63	3.4E-12	*
A_92_P021800	.	26	0.84	9.0E-28	0.56	1.4E-09	
A_92_P022166	Auxin regulated gene	17	0.84	1.0E-27	0.61	2.5E-11	
A_92_P001733	Protein of unknown function	34	0.83	4.2E-27	0.55	4.1E-09	
A_92_P009951	C2 calcium/lipid-binding and GRAM domain containing	31	0.83	5.7E-27	0.44	3.5E-06	
A_92_P030884	ABC transporter, ATP-binding protein	17	0.83	6.1E-27	0.58	2.3E-10	

[a] Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description. A dot indicates that the reporter annotation was inconclusive.
[b] Cytoscape was used to calculate the node degree index (the number of edges connected to a specific node).
[c] MM.greenyellow is the module membership of a specific gene with the greenyellow module. It is the correlation of the gene expression profile with the module eigengene expression profile, across the RILs.
[d] GS.GLS is the gene significance value of a specific gene with the GLS disease severity profile. It is the correlation of the gene expression profile with the GLS disease severity scores, across the RILs.
[e] Genes that were also part of the top 35 potential driver genes in the greenyellow module (Table 3.4) are marked with a star.

Table 3.4: The 35 strongest GLS disease-correlating reporters in the greenyellow module. Reporters were sorted by GS scores (decreasing).

Agilent.ID	Annotation (AT, Rice, B2G)[a]	Degree[b]	MM.[c] green-yellow	p.MM. green-yellow	GS.[d] GLS	p.GS. GLS	GS and MM scores in top 35[e]
A_92_P041977	Protein of unknown function	21	0.82	5.5E-25	0.70	2.8E-16	
A_92_P032816	Glutathione S-transferase GSTU6	9	0.71	1.4E-16	0.68	4.9E-15	
A_92_P037621	F-box kelch-repeat protein skip11-like	81	0.90	1.5E-37	0.68	5.5E-15	*
A_92_P027641	.	30	0.85	1.5E-28	0.68	8.4E-15	*
A_92_P040413	Carnitine acylcarnitine translocase	129	0.94	8.0E-49	0.67	1.6E-14	*
A_92_P035171	Heptahelical transmembrane protein receptor	77	0.91	9.8E-39	0.67	4.2E-14	*
A_92_P015826	Nodulin MtN3 family protein	81	0.92	3.4E-41	0.66	4.6E-14	*
A_92_P029001	Polyphenol oxidase	18	0.84	3.2E-28	0.66	6.1E-14	*
A_92_P032766	NAC transcription factor	90	0.91	2.6E-38	0.66	6.4E-14	*
A_92_P030660	Nuclease I	18	0.81	5.3E-24	0.66	6.5E-14	
A_92_P029666	EF hand / Calmodulin-related calcium sensor protein	14	0.77	1.6E-20	0.66	7.2E-14	
A_92_P027836	Shikimate kinase	9	0.74	1.3E-18	0.66	1.1E-13	
A_92_P035928	Dihydrolipoyllysine-residue acetyltransferase	64	0.82	1.7E-25	0.66	1.1E-13	
A_92_P026342	Chitinase 2-like	3	0.72	3.8E-17	0.66	1.1E-13	
A_92_P022183	Lactate dehydrogenase	13	0.84	6.7E-28	0.66	1.2E-13	*
A_92_P019483	Protein of unknown function	44	0.89	3.2E-35	0.65	2.1E-13	*
A_92_P040918	Patatin group A-3-like / phospholipase A 2A	91	0.92	9.6E-41	0.65	2.2E-13	*
A_92_P024708	ABC transporter C family member 8-like	3	0.74	2.1E-18	0.64	6.6E-13	
A_92_P018873	WRKY transcription factor	37	0.89	7.6E-35	0.64	1.0E-12	*
A_92_P038555	Protein of unknown function	13	0.77	7.8E-21	0.64	1.1E-12	
A_92_P036313	Mitochondrial phosphate carrier	8	0.71	7.8E-17	0.63	1.3E-12	
A_92_P029518	AP2 domain containing protein	75	0.90	1.6E-37	0.63	1.5E-12	*
A_92_P031052	Bifunctional 3-dehydroquinate dehydratase/shikimate	16	0.73	1.3E-17	0.63	1.5E-12	
A_92_P030158	Aspartic proteinase nepenthesin I	7	0.79	4.0E-22	0.63	1.7E-12	
A_92_P009775	Shikimate/quinate hydroxycinnamoyl transferase	12	0.79	1.7E-22	0.63	2.2E-12	
A_92_P018938	RING finger and CHY zinc finger domain-containing	53	0.88	4.7E-34	0.63	2.2E-12	*
A_92_P038520	Cytidine/deoxycytidylate deaminase family protein	22	0.78	1.8E-21	0.63	2.5E-12	
A_92_P040567	Polyphenol oxidase	9	0.74	2.9E-18	0.63	3.3E-12	
A_92_P023090	Cytochrome P450	14	0.84	8.0E-28	0.63	3.4E-12	*
A_92_P012461	Tonoplast dicarboxylate transporter	107	0.94	2.6E-48	0.62	4.2E-12	*
A_92_P006679	Bowman-birk type trypsin inhibitor	39	0.88	5.3E-33	0.62	4.7E-12	*
A_92_P013177	Protein of unknown function	39	0.88	4.4E-33	0.62	4.7E-12	*
A_92_P022755	Chitinase	31	0.88	6.2E-33	0.62	4.7E-12	*
A_92_P021168	High-affinity potassium transporter	2	0.72	5.0E-17	0.62	5.2E-12	
A_92_P008503	Chaperone DnaJ-domain superfamily protein	43	0.85	8.8E-29	0.62	5.5E-12	*

[a] Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description. A dot indicates that the reporter annotation was inconclusive.
[b] Cytoscape was used to calculate the node degree index (the number of edges connected to a specific node).
[c] MM.greenyellow is the module membership of a specific gene with the greenyellow module. It is the correlation of the gene expression profile with the module eigengene expression profile, across the RILs.
[d] GS.GLS is the gene significance value of a specific gene with the GLS disease severity profile. It is the correlation of the gene expression profile with the GLS disease severity scores, across the RILs.
[e] Genes that were also part of the top 35 potential driver genes in the greenyellow module (Table 3.3) are marked with a star.

Table 3.5: Enriched GO-terms for the Blue module based on the full GO using BiNGO.

GO-ID	Description	p-val	Corr p-val	Cluster freq	Total freq
5622	intracellular	2.04E-08	4.39E-05	580/1220 47.5%	3530/8758 40.3%
43231	intracellular membrane-bounded organelle	5.00E-08	4.39E-05	488/1220 40.0%	2909/8758 33.2%
43227	membrane-bounded organelle	5.00E-08	4.39E-05	488/1220 40.0%	2909/8758 33.2%
44424	intracellular part	1.92E-07	1.26E-04	555/1220 45.4%	3401/8758 38.8%
43229	intracellular organelle	4.02E-07	1.76E-04	501/1220 41.0%	3041/8758 34.7%
43226	organelle	4.02E-07	1.76E-04	501/1220 41.0%	3041/8758 34.7%
5737	cytoplasm	9.37E-06	3.53E-03	422/1220 34.5%	2567/8758 29.3%
44444	cytoplasmic part	2.74E-05	8.92E-03	392/1220 32.1%	2387/8758 27.2%
19866	organelle inner membrane	3.14E-05	8.92E-03	34/1220 2.7%	121/8758 1.3%
44446	intracellular organelle part	3.96E-05	8.92E-03	202/1220 16.5%	1130/8758 12.9%
44422	organelle part	4.19E-05	8.92E-03	202/1220 16.5%	1131/8758 12.9%
4177	aminopeptidase activity	4.35E-05	8.92E-03	9/1220 0.7%	15/8758 0.1%
31090	organelle membrane	4.40E-05	8.92E-03	61/1220 5.0%	267/8758 3.0%
5739	mitochondrion	4.83E-05	9.09E-03	94/1220 7.7%	459/8758 5.2%
44429	mitochondrial part	9.40E-05	1.64E-02	32/1220 2.6%	117/8758 1.3%
31975	envelope	1.06E-04	1.64E-02	79/1220 6.4%	379/8758 4.3%
31967	organelle envelope	1.06E-04	1.64E-02	79/1220 6.4%	379/8758 4.3%
9536	plastid	1.12E-04	1.64E-02	219/1220 17.9%	1260/8758 14.3%
9507	chloroplast	1.50E-04	2.08E-02	212/1220 17.3%	1220/8758 13.9%
16036	cellular response to phosphate starvation	1.60E-04	2.11E-02	11/1220 0.9%	24/8758 0.2%
5740	mitochondrial envelope	1.96E-04	2.46E-02	27/1220 2.2%	96/8758 1.0%
31966	mitochondrial membrane	2.35E-04	2.82E-02	26/1220 2.1%	92/8758 1.0%

Table 3.6: Enriched GO-terms for the Magenta module based on the full GO using BiNGO.

GO-ID	Description	p-val	Corr p-val	Cluster freq	Total freq
16020	membrane	5.63E-07	5.88E-04	61/182 33.5%	1607/8758 18.3%
5886	plasma membrane	2.35E-06	1.10E-03	36/182 19.7%	767/8758 8.7%
5622	intracellular	5.42E-06	1.10E-03	103/182 56.5%	3530/8758 40.3%
5623	cell	6.00E-06	1.10E-03	124/182 68.1%	4555/8758 52.0%
44464	cell part	6.00E-06	1.10E-03	124/182 68.1%	4555/8758 52.0%
44424	intracellular part	6.32E-06	1.10E-03	100/182 54.9%	3401/8758 38.8%
43229	intracellular organelle	1.33E-05	1.48E-03	91/182 50.0%	3041/8758 34.7%
43226	organelle	1.33E-05	1.48E-03	91/182 50.0%	3041/8758 34.7%
55035	plastid thylakoid membrane	1.42E-05	1.48E-03	12/182 6.5%	129/8758 1.4%
9535	chloroplast thylakoid membrane	1.42E-05	1.48E-03	12/182 6.5%	129/8758 1.4%
9579	thylakoid	1.99E-05	1.89E-03	15/182 8.2%	203/8758 2.3%
42651	thylakoid membrane	2.81E-05	2.43E-03	12/182 6.5%	138/8758 1.5%
34357	photosynthetic membrane	3.02E-05	2.43E-03	12/182 6.5%	139/8758 1.5%
44434	chloroplast part	5.55E-05	4.13E-03	25/182 13.7%	513/8758 5.8%
44435	plastid part	9.12E-05	5.31E-03	25/182 13.7%	529/8758 6.0%
43231	intracellular membrane-bounded organelle	9.32E-05	5.31E-03	85/182 46.7%	2909/8758 33.2%
43227	membrane-bounded organelle	9.32E-05	5.31E-03	85/182 46.7%	2909/8758 33.2%
44446	intracellular organelle part	9.81E-05	5.31E-03	42/182 23.0%	1130/8758 12.9%
44422	organelle part	1.00E-04	5.31E-03	42/182 23.0%	1131/8758 12.9%
32991	macromolecular complex	1.02E-04	5.31E-03	31/182 17.0%	734/8758 8.3%
5488	binding	1.11E-04	5.50E-03	95/182 52.1%	3377/8758 38.5%
31984	organelle subcompartment	1.43E-04	6.20E-03	12/182 6.5%	163/8758 1.8%
31976	plastid thylakoid	1.43E-04	6.20E-03	12/182 6.5%	163/8758 1.8%
9534	chloroplast thylakoid	1.43E-04	6.20E-03	12/182 6.5%	163/8758 1.8%
6970	response to osmotic stress	1.66E-04	6.95E-03	13/182 7.1%	191/8758 2.1%
44436	thylakoid part	2.01E-04	8.05E-03	12/182 6.5%	169/8758 1.9%
9628	response to abiotic stimulus	2.23E-04	8.62E-03	25/182 13.7%	560/8758 6.3%
44267	cellular protein metabolic process	2.62E-04	9.78E-03	35/182 19.2%	918/8758 10.4%
44444	cytoplasmic part	3.32E-04	1.14E-02	71/182 39.0%	2387/8758 27.2%
5198	structural molecule activity	3.50E-04	1.14E-02	13/182 7.1%	206/8758 2.3%
3735	structural constituent of ribosome	3.52E-04	1.14E-02	11/182 6.0%	154/8758 1.7%
5737	cytoplasm	3.56E-04	1.14E-02	75/182 41.2%	2567/8758 29.3%
9651	response to salt stress	3.60E-04	1.14E-02	12/182 6.5%	180/8758 2.0%
9987	cellular process	4.77E-04	1.46E-02	84/182 46.1%	2992/8758 34.1%
9570	chloroplast stroma	5.47E-04	1.63E-02	14/182 7.6%	244/8758 2.7%
5634	nucleus	5.64E-04	1.63E-02	35/182 19.2%	956/8758 10.9%
15979	photosynthesis	5.93E-04	1.67E-02	7/182 3.8%	70/8758 0.7%
51082	unfolded protein binding	7.18E-04	1.95E-02	5/182 2.7%	35/8758 0.3%
44237	cellular metabolic process	7.30E-04	1.95E-02	66/182 36.2%	2231/8758 25.4%
9532	plastid stroma	8.78E-04	2.29E-02	14/182 7.6%	256/8758 2.9%
10287	plastoglobule	1.19E-03	3.03E-02	5/182 2.7%	39/8758 0.4%
10038	response to metal ion	1.30E-03	3.22E-02	11/182 6.0%	180/8758 2.0%
19538	protein metabolic process	1.80E-03	4.36E-02	36/182 19.7%	1059/8758 12.0%
9536	plastid	1.90E-03	4.51E-02	41/182 22.5%	1260/8758 14.3%

Table 3.7: The ten genes that best correlated with the module eigengenes of each of the remaining modules significantly associated with GLS susceptibility. These lists include potential driver genes of the respective modules.

Module colour	Agilent.ID	Annotation (AT, Rice, B2G)[a]	Degree[b]	MM.[c] module	p.MM. module	GS.[d] GLS	p.GS. GLS
paleturquoise	A_92_P027935	ATP synthase / copper, cobalt, zinc ion binding	25	0.94	4.2E-49	0.30	2.3E-03
paleturquoise	A_92_P027135	Citrate synthase	26	0.94	6.9E-48	0.40	4.2E-05
paleturquoise	A_92_P026402	Protein translation factor SUI1	19	0.93	2.7E-44	0.28	4.9E-03
paleturquoise	A_92_P027779	Proline-rich protein	21	0.92	1.5E-42	0.35	3.3E-04
paleturquoise	A_92_P025610	SCM-like with four MBT domains 1	18	0.90	1.2E-37	0.25	1.2E-02
paleturquoise	A_92_P026329	Protein with unknown function	12	0.89	1.5E-34	0.29	3.6E-03
paleturquoise	A_92_P022171	Vacuolar ATP synthase subunit G	18	0.88	1.8E-33	0.21	3.8E-02
paleturquoise	A_92_P026244	Protein with unknown function	17	0.88	3.1E-33	0.22	2.9E-02
paleturquoise	A_92_P028020	Protein with unknown function	13	0.87	1.6E-31	0.40	3.6E-05
paleturquoise	A_92_P027288	HSP70-interacting protein 1	11	0.85	2.4E-29	0.23	2.3E-02
blue	A_92_P035544	Magnesium transporter NIPA2-like	812	0.95	3.1E-52	0.26	1.0E-02
blue	A_92_P038748	Vacuolar-sorting receptor 3	739	0.94	6.5E-49	0.24	1.7E-02
blue	A_92_P035048	Ser/Thr protein phosphatase family protein	627	0.94	1.9E-47	0.27	6.9E-03
blue	A_92_P038049	Aspartyl aminopeptidase	842	0.94	2.6E-47	0.29	3.7E-03
blue	A_92_P035165	Hydroxyacylglutathione hydrolase	663	0.94	5.4E-47	0.25	1.3E-02
blue	A_92_P039811	Mitochondrial-processing peptidase subunit beta-like	898	0.94	9.6E-47	0.28	5.0E-03
blue	A_92_P041201	Aldehyde dehydrogenase	674	0.93	1.3E-45	0.21	3.6E-02
blue	A_92_P038663	NADH dehydrogenase	503	0.93	5.8E-44	0.16	1.2E-01
blue	A_92_P039354	Chloroplast protease	561	0.93	5.0E-43	0.21	4.0E-02
blue	A_92_P038572	Protein of unknown function	627	0.92	1.7E-42	0.16	1.2E-01
yellowgreen	A_92_P034676	Serine/threonine-protein kinase HT1	534	0.92	1.7E-42	0.18	7.3E-02
yellowgreen	A_92_P037706	Serine carboxypeptidase-like 48	678	0.92	1.8E-42	0.16	1.2E-01
yellowgreen	A_92_P039569	NUDIX hydrolase 13	14	0.88	6.1E-34	0.44	5.1E-06
yellowgreen	A_92_P005381	.	15	0.88	4.5E-33	0.06	5.3E-01
yellowgreen	A_92_P018033	Amidase family protein	11	0.87	1.4E-31	0.23	2.2E-02
yellowgreen	A_92_P039815	.	14	0.85	1.4E-29	0.27	5.8E-03
yellowgreen	A_92_P007213	NAD(P)-binding Rossmann-fold superfamily protein	12	0.85	4.2E-29	0.19	6.2E-02
yellowgreen	A_92_P035027	Sec23/Sec24 protein transport family protein	10	0.83	2.3E-26	0.32	1.4E-03
yellowgreen	A_92_P038037	.	5	0.79	6.8E-23	0.27	7.4E-03
yellowgreen	A_92_P038446	.	2	0.79	1.6E-22	0.24	1.5E-02
magenta	A_92_P010731	60S ribosomal protein L28	229	0.95	4.2E-53	0.25	1.1E-02
magenta	A_92_P029339	N-terminal nucleophile aminohydrolase	220	0.94	6.8E-48	0.28	5.1E-03
magenta	A_92_P024192	.	223	0.94	2.0E-47	0.35	4.4E-04
magenta	A_92_P011228	DAG protein, chloroplast precursor	217	0.93	5.7E-45	0.24	1.8E-02
magenta	A_92_P028770	Plastid-lipid associated protein PAP / fibrillin protein	217	0.93	9.5E-45	0.29	3.0E-03
magenta	A_92_P015841	.	215	0.93	2.9E-44	0.32	9.7E-04
magenta	A_92_P018198	Allyl alcohol dehydrogenase-like protein	220	0.93	6.2E-44	0.21	3.8E-02
magenta	A_92_P016596	Programmed cell death protein 5	211	0.93	6.3E-44	0.19	5.7E-02
magenta	A_92_P013494	Bifunctional aminoacyl-tRNA synthetase	222	0.93	8.6E-44	0.17	9.4E-02
magenta	A_92_P012559	Dehydroascorbate reductase	212	0.93	2.3E-43	0.15	1.4E-01

[a] Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description. A dot indicates that the reporter annotation was inconclusive.
[b] Cytoscape was used to calculate the node degree index (the number of edges connected to a specific node).
[c] MM.module is the module membership of a gene with a specific module. It is the correlation of the gene expression profile with the module eigengene expression profile, across the RILs. The "module" corresponds to the specific module colour in the first column.
[d] GS.GLS is the gene significance value with the GLS disease trait. It is the correlation of the gene expression profile with the GLS disease severity scores, across the RILs.

Table 3.8: The ten genes that best correlated with the GLS disease scores, in each of the remaining modules significantly associated with GLS susceptibility.

Module colour	Agilent.ID	Annotation (AT, Rice, B2G)[a]	Degree[b]	MM.[c] module	p.MM. module	GS.[d] GLS	p.GS. GLS
paleturquoise	A_92_P027772	.	5	0.82	1.8E-25	0.45	2.6E-06
paleturquoise	A_92_P030287	CAS1 domain-containing protein 1-like	3	0.74	6.7E-19	0.44	3.7E-06
paleturquoise	A_92_P028020	Protein with unknown function	13	0.87	1.6E-31	0.40	3.6E-05
paleturquoise	A_92_P028331	Rickettsia surface antigen family protein	6	0.80	1.1E-23	0.40	4.0E-05
paleturquoise	A_92_P027135	Citrate synthase	26	0.94	6.9E-48	0.40	4.2E-05
paleturquoise	A_92_P030448	ATP citrate lyase	5	0.80	8.0E-24	0.36	2.4E-04
paleturquoise	A_92_P027779	Proline-rich protein	21	0.92	1.5E-42	0.35	3.3E-04
paleturquoise	A_92_P022106	RAS-related protein RIC1	6	0.79	7.5E-23	0.35	3.8E-04
paleturquoise	A_92_P027052	GDP-mannose transporter	8	0.81	4.3E-24	0.34	6.2E-04
paleturquoise	A_92_P023878	Ribosomal protein L34E-like protein	4	0.76	2.1E-20	0.33	6.6E-04
blue	A_92_P038236	High-affinity potassium transporter	33	0.62	4.9E-12	0.54	6.2E-09
blue	A_92_P029774	6-phosphogluconate dehydrogenase	91	0.63	3.1E-12	0.50	1.2E-07
blue	A_92_P032712	Heme binding protein	113	0.77	8.7E-21	0.50	1.4E-07
blue	A_92_P035922	Ferrochelatase-2, chloroplast precursor	205	0.74	1.9E-18	0.48	4.6E-07
blue	A_92_P038771	.	232	0.74	1.3E-18	0.47	9.7E-07
blue	A_92_P025853	Cyclin-dependent kinase D1	125	0.75	5.3E-19	0.46	1.6E-06
blue	A_92_P021967	NAD(P)H-dependent oxidoreductase	132	0.68	4.1E-15	0.46	1.8E-06
blue	A_92_P030141	Ribonuclease T2 family domain containing protein	36	0.58	3.1E-10	0.46	1.8E-06
blue	A_92_P039583	Undecaprenyl pyrophosphate synthetase	180	0.74	2.6E-18	0.45	3.1E-06
blue	A_92_P039524	CAS1 domain-containing protein 1	186	0.78	1.4E-21	0.44	3.7E-06
yellowgreen	A_92_P013271	.	.	0.69	2.8E-15	0.46	1.8E-06
yellowgreen	A_92_P039569	NUDIX hydrolase 13	14	0.88	6.1E-34	0.44	5.1E-06
yellowgreen	A_92_P035027	Sec23/Sec24 protein transport family protein	10	0.83	2.3E-26	0.32	1.4E-03
yellowgreen	A_92_P026055	Aspartic proteinase A1	6	0.77	8.7E-21	0.28	5.0E-03
yellowgreen	A_92_P009397	Protein of unknown function	19	0.93	2.1E-44	0.28	5.5E-03
yellowgreen	A_92_P039815	.	14	0.85	1.4E-29	0.27	5.8E-03
yellowgreen	A_92_P025850	Salt response protein	8	0.77	1.1E-20	0.27	5.9E-03
yellowgreen	A_92_P038037	.	5	0.79	6.8E-23	0.27	7.4E-03
yellowgreen	A_92_P034995	.	7	0.74	1.1E-18	0.26	9.3E-03
yellowgreen	A_92_P038512	Glutathione S-transferase	1	0.72	4.6E-17	0.25	1.3E-02
magenta	A_92_P029634	Flavanone 3-hydroxylase	96	0.80	9.7E-24	0.52	3.3E-08
magenta	A_92_P035711	Probable polyamine oxidase 2-like	1	0.64	7.7E-13	0.45	2.8E-06
magenta	A_92_P029669	Chorismate mutase	74	0.75	5.2E-19	0.41	2.5E-05
magenta	A_92_P025158	Chalcone synthase	69	0.77	1.8E-20	0.40	3.2E-05
magenta	A_92_P027160	.	50	0.74	6.4E-19	0.37	1.6E-04
magenta	A_92_P038681	Probable E3 ubiquitin-protein ligase ARI7-like	109	0.82	1.1E-25	0.36	2.3E-04
magenta	A_92_P019536	Mitogen-activated protein kinase 3	203	0.92	1.6E-40	0.36	2.5E-04
magenta	A_92_P031672	Selenium-binding protein 2	156	0.86	3.5E-30	0.35	3.7E-04
magenta	A_92_P024192	.	223	0.94	2.0E-47	0.35	4.4E-04
magenta	A_92_P042004	Chlorophyll fluorescence increase protein	NA	0.63	3.2E-12	0.34	4.5E-04

[a] Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description. A dot indicates that the reporter annotation was inconclusive.
[b] Cytoscape was used to calculate the node degree index (the number of edges connected to a specific node).
[c] MM.module is the module membership of a gene with a specific module. It is the correlation of the gene expression profile with the module eigengene expression profile, across the RILs. The "module" corresponds to the specific module colour in the first column.
[d] GS.GLS is the gene significance value with the GLS disease trait. It is the correlation of the gene expression profile with the GLS disease severity scores, across the RILs.

Table 3.9: Enriched GO-terms for the Turquoise module based on plant GO slim using BiNGO.

GO-ID	Description	p-val	Corr p-val	Cluster freq	Total freq
5622	intracellular	1.56E-04	9.50E-03	573/1275 44.9%	3530/8758 40.3%
166	nucleotide binding	2.87E-04	9.50E-03	160/1275 12.5%	857/8758 9.7%
40007	growth	3.00E-04	9.50E-03	30/1275 2.3%	109/8758 1.2%
5886	plasma membrane	6.55E-04	1.28E-02	143/1275 11.2%	767/8758 8.7%
5515	protein binding	7.52E-04	1.28E-02	168/1275 13.1%	924/8758 10.5%
5737	cytoplasm	8.06E-04	1.28E-02	422/1275 33.0%	2567/8758 29.3%
5623	cell	1.66E-03	2.09E-02	712/1275 55.8%	4555/8758 52.0%
6464	protein modification process	1.91E-03	2.09E-02	91/1275 7.1%	469/8758 5.3%
9838	abscission	1.98E-03	2.09E-02	4/1275 0.3%	5/8758 0.0%
6810	transport	2.26E-03	2.15E-02	114/1275 8.9%	611/8758 6.9%
9987	cellular process	3.19E-03	2.76E-02	479/1275 37.5%	2992/8758 34.1%
5215	transporter activity	4.99E-03	3.95E-02	87/1275 6.8%	460/8758 5.2%
16049	cell growth	6.57E-03	4.80E-02	22/1275 1.7%	88/8758 1.0%

Table 3.10: The genes that best correlated with the module eigengenes of the modules significantly associated with GLS resistance. These lists include potential driver genes of the respective modules.

Module colour	Agilent.ID	Annotation (AT, Rice, B2G)[a]	Degree[b]	MM.[c] module	p.MM. module	GS.[d] GLS	p.GS. GLS
turquoise	A_92_P001413	Coronatine-insensitive protein 1	951	0.94	2.9E-47	-0.21	3.6E-02
turquoise	A_92_P005691	Digalactosyldiacylglycerol synthase	988	0.94	4.6E-46	-0.24	1.7E-02
turquoise	A_92_P006813	Protein phosphatase 2C family protein	960	0.93	7.4E-45	-0.23	2.4E-02
turquoise	A_92_P005522	LEM3 (ligand-effect modulator 3) family protein	846	0.93	1.0E-44	-0.30	2.4E-03
turquoise	A_92_P001487	Protein of unknown function	928	0.93	1.8E-44	-0.23	2.1E-02
turquoise	A_92_P002512	Serine/threonine-protein kinase HT1	760	0.93	4.0E-44	-0.34	5.0E-04
turquoise	A_92_P007274	Zinc finger, C3HC4 type domain containing protein	722	0.93	3.8E-43	-0.30	2.2E-03
turquoise	A_92_P006383	Signal peptide peptidase-like 2B	784	0.92	7.6E-43	-0.20	4.4E-02
turquoise	A_92_P002825	Transmembrane protein 184C-like	612	0.92	3.6E-42	-0.23	2.0E-02
turquoise	A_92_P004638	.	800	0.92	1.2E-41	-0.28	4.8E-03
turquoise	A_92_P007102	Protein kinase PKN/PRK1, effector	870	0.92	1.6E-41	-0.26	8.2E-03
turquoise	A_92_P004301	aspartyl-tRNA synthetase	642	0.92	1.5E-40	-0.29	3.6E-03
turquoise	A_92_P004860	O-fucosyltransferase family protein	512	0.91	3.1E-40	-0.23	2.0E-02
turquoise	A_92_P002786	Inositol-pentakisphosphate 2-kinase family protein	575	0.91	2.2E-39	-0.32	1.2E-03
turquoise	A_92_P004887	Heat shock factor 3	401	0.91	2.3E-39	-0.22	2.6E-02
turquoise	A_92_P007135	KH domain-containing protein	671	0.91	2.4E-39	-0.35	3.3E-04
turquoise	A_92_P004641	Magnesium transporter NIPA2	528	0.91	2.9E-39	-0.17	9.9E-02
turquoise	A_92_P006712	F-box/RNI-like superfamily protein	833	0.91	3.6E-39	-0.17	8.2E-02
turquoise	A_92_P006063	Nucleoporin autopeptidase	716	0.91	5.8E-39	-0.28	5.1E-03
turquoise	A_92_P005988	Serine threonine protein phosphatase 2A regulatory subunit	569	0.91	1.2E-38	-0.23	2.3E-02
darkred	A_92_P004852	Transportin 1	40	0.95	5.9E-51	-0.20	4.6E-02
darkred	A_92_P005017	Ubiquitin-conjugating enzyme E2 7	36	0.94	4.7E-48	-0.21	3.4E-02
darkred	A_92_P005333	Vacuolar protein sorting-associated protein 18	35	0.93	8.7E-46	-0.24	1.4E-02
darkred	A_92_P005091	.	35	0.93	1.3E-43	-0.27	6.6E-03
darkred	A_92_P004287	Pectin lyase-like superfamily protein	34	0.91	2.2E-38	-0.06	5.4E-01
darkred	A_92_P014451	No exine formation 1	27	0.89	1.3E-35	-0.07	4.8E-01
darkred	A_92_P003909	.	26	0.89	3.5E-35	-0.18	7.4E-02
darkred	A_92_P004214	26S proteasome non-ATPase regulatory subunit 3	31	0.89	1.5E-34	-0.15	1.5E-01
darkred	A_92_P003665	Polyprenyltransferase 1	24	0.89	1.6E-34	-0.29	3.1E-03
darkred	A_92_P013670	Enhanced disease resistance 2 protein	26	0.88	8.7E-34	-0.20	5.1E-02
yellow	A_92_P003122	Shugoshin-1	945	0.95	2.8E-53	-0.24	1.5E-02
yellow	A_92_P003532	Actin-related protein 2/3 complex subunit 2	923	0.95	8.5E-50	-0.30	2.4E-03
yellow	A_92_P001412	Universal stress protein domain containing protein	883	0.94	2.5E-48	-0.27	6.9E-03
yellow	A_92_P001640	Acyl carrier protein phosphodiesterase	858	0.94	7.5E-47	-0.21	3.2E-02
yellow	A_92_P006321	ATP-dependent protease La (LON) domain protein	821	0.94	3.8E-46	-0.18	6.9E-02
yellow	A_92_P004097	Wound-responsive family protein	825	0.93	2.9E-45	-0.26	9.2E-03
yellow	A_92_P005135	Protein of unknown function	807	0.93	3.3E-45	-0.26	1.0E-02
yellow	A_92_P003784	5'-3' exoribonuclease 4	827	0.93	3.5E-45	-0.24	1.7E-02
yellow	A_92_P002244	Hydrolase, NUDIX family, domain containing protein	868	0.93	9.5E-45	-0.07	4.7E-01
yellow	A_92_P003208	Arginine-tRNA protein transferase 2	837	0.93	2.1E-44	-0.24	1.4E-02

[a] Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description. A dot indicates that the reporter annotation was inconclusive.
[b] Cytoscape was used to calculate the node degree index (the number of edges connected to a specific node).
[c] MM.module is the module membership of a gene with a specific module. It is the correlation of the gene expression profile with the module eigengene expression profile, across the RILs. The "module" corresponds to the specific module colour in the first column.
[d] GS.GLS is the gene significance value with the GLS disease trait. It is the correlation of the gene expression profile with the GLS disease severity scores, across the RILs.

Table 3.11: The genes that best correlated negatively with the GLS disease scores, in the modules significantly associated with GLS resistance.

Module colour	Agilent.ID	Annotation (AT, Rice, B2G)[a]	Degree[b]	MM.[c] module	p.MM. module	GS.[d] GLS	p.GS. GLS
turquoise	A_92_P010785	Callose synthase 3-like	20	0.74	1.1E-18	-0.53	2.0E-08
turquoise	A_92_P008134	Auxin response factor 2	97	0.79	2.0E-22	-0.52	3.2E-08
turquoise	A_92_P002148	Queuine tRNA-ribosyltransferase	218	0.84	1.4E-27	-0.49	2.7E-07
turquoise	A_92_P012261	Transcription termination factor nusG family protein	56	0.74	1.4E-18	-0.48	5.3E-07
turquoise	A_92_P012893	RNA-binding (RRM/RBD/RNP motifs) family protein	5	0.60	6.6E-11	-0.48	5.6E-07
turquoise	A_92_P010694	UDP-Glycosyltransferase superfamily protein	.	0.67	1.9E-14	-0.47	7.3E-07
turquoise	A_92_P005274	ATP-dependent protease La (LON) domain protein	18	0.76	2.7E-20	-0.47	8.3E-07
turquoise	A_92_P005020	Protein of unknown function	75	0.79	3.7E-22	-0.47	1.0E-06
turquoise	A_92_P010296	Phospholipid-translocating ATPase	38	0.72	5.6E-17	-0.46	1.3E-06
turquoise	A_92_P011441	Ternary complex factor MIP1-like	28	0.69	1.9E-15	-0.46	1.4E-06
turquoise	A_92_P010621	Ubiquinone biosynthesis protein UbiB	2	0.63	3.4E-12	-0.46	1.6E-06
turquoise	A_92_P011123	Protein of unknown function	22	0.76	6.6E-20	-0.45	2.2E-06
turquoise	A_92_P006601	Protein of unknown function	35	0.77	1.7E-20	-0.45	2.4E-06
turquoise	A_92_P001822	Ferrochelatase-2, chloroplast precursor	1	0.70	9.8E-16	-0.45	2.5E-06
turquoise	A_92_P004813	double-stranded RNA binding motif containing protein	319	0.83	3.4E-26	-0.45	2.6E-06
turquoise	A_92_P010388	Thioesterase superfamily member 2	67	0.82	3.2E-25	-0.45	2.8E-06
turquoise	A_92_P005508	Methyltransferase domain containing protein	12	0.73	5.3E-18	-0.44	6.1E-06
turquoise	A_92_P012822	Pectin lyase-like superfamily protein	9	0.72	2.4E-17	-0.43	6.7E-06
turquoise	A_92_P015327	basic helix-loop-helix (bHLH) DNA-binding protein	3	0.57	5.7E-10	-0.43	7.5E-06
turquoise	A_92_P002874	.	68	0.82	3.6E-25	-0.43	9.6E-06
darkred	A_92_P014935	F-box protein, phloem protein 2-A13	8	0.75	2.4E-19	-0.39	5.4E-05
darkred	A_92_P005174	TCP-domain protein	17	0.86	5.1E-31	-0.39	7.5E-05
darkred	A_92_P015248	Pod-specific dehydrogenase SAC25	4	0.77	7.2E-21	-0.33	8.7E-04
darkred	A_92_P013981	Protein of unknown function	6	0.78	1.5E-21	-0.32	1.2E-03
darkred	A_92_P014534	Histone-lysine N-methyltransferases	7	0.73	1.2E-17	-0.31	1.6E-03
darkred	A_92_P002788	Tetratricopeptide repeat (TPR)-like superfamily protein	.	0.67	2.3E-14	-0.31	1.7E-03
darkred	A_92_P004066	Serine carboxypeptidase-like 34	18	0.86	1.7E-30	-0.31	1.7E-03
darkred	A_92_P013508	Peptidase M50 family protein	19	0.83	8.1E-27	-0.30	2.1E-03
darkred	A_92_P002784	60s ribosomal protein L7Ae	17	0.87	1.3E-31	-0.30	2.5E-03
darkred	A_92_P014528	Protein of unknown function	5	0.69	1.6E-15	-0.30	2.9E-03
yellow	A_92_P006568	Protein of unknown function	82	0.71	2.1E-16	-0.46	1.6E-06
yellow	A_92_P004943	Protein of unknown function	175	0.84	9.5E-28	-0.44	4.5E-06
yellow	A_92_P006889	Transducin/WD40 repeat-like superfamily protein	104	0.76	8.9E-20	-0.42	1.6E-05
yellow	A_92_P002149	B3 DNA binding domain containing protein	372	0.88	2.8E-34	-0.41	2.2E-05
yellow	A_92_P013130	Retrotransposon protein	.	0.56	1.0E-09	-0.41	2.2E-05
yellow	A_92_P014539	DNA-binding storekeeper-related transcriptional regulator	15	0.68	5.9E-15	-0.41	2.4E-05
yellow	A_92_P009321	Translation initiation factor 2, small GTP-binding protein	2	0.64	5.7E-13	-0.41	2.6E-05
yellow	A_92_P007774	Protein of unknown function	21	0.79	2.9E-22	-0.41	2.7E-05
yellow	A_92_P005598	Homeobox-leucine zipper protein HAT14	153	0.78	1.0E-21	-0.41	2.8E-05
yellow	A_92_P004061	Circadian clock coupling factor ZGT	113	0.78	1.4E-21	-0.40	3.1E-05

[a] Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description. A dot indicates that the reporter annotation was inconclusive.
[b] Cytoscape was used to calculate the node degree index (the number of edges connected to a specific node).
[c] MM.module is the module membership of a gene with a specific module. It is the correlation of the gene expression profile with the module eigengene expression profile, across the RILs. The "module" corresponds to the specific module colour in the first column.
[d] GS.GLS is the gene significance value with the GLS disease trait. It is the correlation of the gene expression profile with the GLS disease severity scores, across the RILs.

Table 3.12: A list of all the reporters encoding callose synthases, in the WGCNA input set of 19,281 input reporters. The correlation of their expression profiles in relation to the GLS scores are also given, together with p-values indicating significance.

Agilent.ID	ZM.ID	B2G.description	moduleColor	GS.GLS	p.GS.GLS	p-val < 0.05
A_92_P010785	GRMZM2G059212_T01	callose synthase 3-like	turquoise	-0.53	2.0E-08	*
A_92_P001619	GRMZM2G111529_T01	callose synthase	yellow	-0.27	0.007	*
A_92_P000970	GRMZM2G022856_T01	callose synthase 11-like	yellow	-0.23	0.024	*
A_92_P016646	GRMZM2G465764_T01	callose synthase	grey	-0.22	0.031	*
A_92_P005832	GRMZM2G430680_T01	callose synthase	yellow	-0.21	0.032	*
A_92_P033767	GRMZM2G180951_T01	callose synthase 3-like	grey	-0.19	0.063	
A_92_P010595	GRMZM2G453794_T01	callose synthase	turquoise	-0.15	0.128	
A_92_P010167	GRMZM2G326643_T03	callose synthase 10	green	-0.07	0.514	
A_92_P025009	GRMZM5G840560_T01	callose synthase 3-like	grey	0.07	0.509	
A_92_P034724	GRMZM2G084802_T01	callose synthase	grey	0.22	0.025	*

Chapter 4

Global expression QTL analysis towards identifying the molecular basis of grey leaf spot disease in maize

4.1 Introduction

Grey leaf spot (GLS) is a foliar disease of maize of great economic importance in many countries. GLS is caused by the fungal pathogens *Cercospora zeae-maydis* Tehon and E. Y. Daniels and *Cercospora zeina* Crous & U. Braun and can significantly reduce grain yield depending on the level of susceptibility of the hybrid (Latterell and Rossi, 1983). Resistant commercial hybrids are not readily available. In order to breed new resistant hybrids it is essential to understand the mechanisms, determined by the expression of certain genes and pathways, underlying disease resistance.

As reviewed in Chapter 1, quantitative trait locus (QTL) mapping is used to analyse natural variation by identifying genomic regions affecting a quantitative trait such as GLS resistance in maize. Determining the causal genes underlying phenotypic QTLs is currently a major challenge, for which the combination of large scale expression profiling and genetic analyses to identify expression QTLs (eQTLs) can play a significant role. Figure 4.1 gives an overview of the steps in a full eQTL experiment, where population-wide genotyping and gene expression profiling precede eQTL mapping and interpretation. eQTLs are thus not linked to specific phenotypic traits, but can be used to search for associations between gene expression polymorphisms and phenotypic QTLs.

In Chapter 3, expression traits (e-traits) across a CML444×SC Malawi maize recombinant inbred line (RIL) population infected with *C. zeina* were used in a correlation analysis to show that coordinated responses to GLS disease in subtropical white dent maize lines were apparent. Clusters of genes that significantly correlated with GLS resistance or susceptibility were determined and central nodes were identified as potential network drivers. However, a limitation of the analysis so far was that there was no reference to what the genetic basis for the response to *C. zeina* infection in the maize RIL population could be. The natural variation captured in this RIL population provided a valuable resource for answering this question.

According to previous studies, complex genetic interactions underlie GLS resistance in maize (Bubeck *et al.*, 1993; Saghai Maroof *et al.*, 1996; Lehmensiek *et al.*, 2001; Gordon *et al.*, 2004; Menkir and Ayodele, 2005; Zhang *et al.*, 2012*b*). However, in most of the published GLS quantitative trait locus (QTL) studies, *C. zeae-maydis* instead of *C. zeina* was the causal agent of GLS disease. Therefore, adding the genetic dimension will reveal valuable information about the genomic locations responsible for *C. zeina* resistance or susceptibility in the CML444×SC Malawi maize cross. Once the genetic basis (i.e. the causal genes or polymorphisms) is determined, this could be utilised in crop improvement through marker-assisted breeding or via genetic modification.

The aim of the work reported in this chapter was to combine QTL mapping for GLS severity with eQTL analyses to investigate the molecular basis of the quantitative disease response to *C. zeina* infection. The ultimate goal was to identify hypotheses of genes and mechanisms that could explain the GLS severity phenotypic QTLs.

4.1.1 Using expression QTLs for the identification of genes and pathways affecting phenotypic traits

QTL analysis of gene expression profiles (i.e. eQTL analysis) identifies genomic regions, which are likely to contain a causal gene with regulatory effect on the gene whose expression profile is affected by the eQTL. eQTLs are classified as *cis* or *trans*, where a *cis*-eQTL represent a polymorphism physically located near the gene itself and a *trans*-eQTL represent a polymorphism at a different location than the position of the gene affected by the eQTL.

Identifying *cis*-eQTLs that co-localise with a phenotypic QTL can be a valuable ap-

proach for causal gene discovery, since these show differences in gene expression that are under the control of DNA sequence variants in or close to the gene itself, potentially contributing to the phenotypic QTL. Furthermore, the identification of a set of genes with *trans*-eQTLs at a common locus can be used to dissect genetic variation that influences an entire pathway or biological process, which can lead to the identification of initiating polymorphisms upstream in a transcriptional network. Additionally, one could also perform a correlation analysis between the phenotype scores and the gene expression values (e-traits) across the individuals in a population, to filter the lists of candidate genes underlying phenotypic QTLs. Since a causal relationship cannot be inferred based on correlation alone, a gene with a good correlation and with an eQTL that coincides with the phenotypic QTL is a stronger candidate for causing the trait. Furthermore, the combination of e-trait correlations (correlation of gene expression patterns) and co-localisation of *trans*-eQTL positions with phenotypic QTLs, could provide a powerful strategy to infer potential regulatory gene networks affecting the phenotypic trait. Ultimately, phenotypic QTLs can be searched for regulatory genes with *cis*-eQTLs that potentially affect the phenotype through its effect on many downstream genes.

The first two examples given below, illustrate that the combination of e-trait correlations (the correlation between pairs of gene expression profiles) with eQTL mapping approaches is a powerful strategy to infer regulatory gene networks. The third example illustrates that a correlation analysis between phenotype values and gene expression values of the genes that underly a phenotypic QTL can give considerable insight, when the aim of the investigation is to identify candidate genes responsible for a phenotypic trait.

Lan *et al.* (2006) used gene expression data across the individuals of a segregating population to identify co-regulated genes as well as genomic locations of putative regulatory loci (i.e. eQTLs). Their study was based on 45,000 e-traits derived from 60 mice in an F_2 population segregating for obesity and diabetes. They first used eQTL mapping to identify 6,016 "seed" transcripts (with LOD scores of 3.4 or greater at one or more locations in the genome). Subsequently, they identified transcripts whose expression profiles were highly correlated to those of the seed transcripts (using Pearson correlation coefficient cut-off of 0.7). They further tested for enrichment of common biological functions among the lists of correlated transcripts. Out of the 6,016 seed transcripts, 1,341 produced lists of e-traits that were enriched for at least one gene ontology (GO) term.

Thirty-eight of the identified seeds belonged to the G protein-coupled receptor (GPCR) protein signalling pathway and were correlated with 174 transcripts that were also part of the GPCR protein signalling pathway according to GO-term annotations. Remarkably, 131 of the identified transcripts shared a regulatory locus on chromosome 2, even though the linkage was not always significant across the 60 F_2 mice. Thus, Lan *et al.* (2006) used e-trait correlation combined with eQTL mapping to reveal regulatory networks that would otherwise be missed. Several QTLs for obesity and related traits have been mapped to this region on chromosome 2, making it a region of interest for further study.

Bing and Hoeschele (2005) reanalysed the e-trait data from a yeast study (Brem *et al.*, 2002). The data set contained 6,215 gene expression values as well as genotypes at 3,312 markers for each of 40 haploid segregants from a cross between a laboratory strain and a wild strain of *Saccharomyces cerevisiae*. They incorporated a correlation analysis for transcription network inference and their strategy was to: (i) identify eQTLs; (ii) determine a set of regulatory candidate genes physically located within each eQTL confidence region (using the sequenced yeast genome map); (iii) reduce the number of candidate causal genes in each eQTL interval by correlation analysis of expression; (iv) draw directional links from the remaining candidate causal gene(s) in each eQTL region to the gene affected by the eQTL; and (v) join the putative regulatory links to form networks. This method was used for reducing the set of candidate genes for both *cis*- and *trans*-eQTL regions. For 65% of the identified eQTL regions, a single candidate regulatory gene was retained in the e-trait correlation analysis (step iii); for 7% of the eQTL regions, no gene was retained; and for 28% of the eQTL regions, more than one gene were retained. A few biological processes were identified as significantly overrepresented in either independent network structures or in highly interconnected sub-networks. Interestingly, most of the transcription factors that were found in the inferred networks had a putative regulatory link to only one other gene or exhibited *cis*-regulation (Bing and Hoeschele, 2005). Since the aim of this study was to illustrate a method for transcriptional regulatory network inference, no phenotypic QTLs were investigated.

Druka *et al.* (2008) exploited regulatory variation to identify genes underlying quantitative resistance to the wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici* in barley. They used used 1,536 SNP markers and 139 DH lines of the Steptoe×Morex reference barley mapping population to explore the potential of using e-traits as surrogates for

the identification of candidate genes underlying the interaction between barley and the wheat stem rust fungus. Six phenotypic QTLs associated with barley's reaction to stem rust were identified and one of these coincided with the major stem rust resistance locus, *Rpg1* on chromosome 7H, that were previously positionally cloned using this population. Correlation analysis between phenotype values for rust infection and the gene expression values of the genes underlying the major QTL on chromosome 7H, placed *Rpg1* in the top five candidate genes.

4.1.2 Tools for eQTL discovery and interpretation

eQTL analysis recently moved from a cutting-edge genomics concept to a more mature area of investigation (Wright *et al.*, 2012). This led to a rapid expansion of available datasets and numerous attempts to functionally relate eQTLs to phenotypic traits. To this end, tools for detecting eQTLs (Basten *et al.*, 2001; van Ooijen, 2009), ways to display or browse pre-analysed eQTL data (Mueller *et al.*, 2006; Zou *et al.*, 2007) and databases to store available datasets (Wang *et al.*, 2003; Mueller *et al.*, 2006) became essential to numerous researchers. However, only a few computational tools for the analysis and post-processing of eQTL data are available and these are generally not widely distributed or not accompanied by user-friendly software packages. As a result, investigators interested in eQTL analyses often need to develop their own code or use tools not specifically suited for eQTL analysis. Wright *et al.* (2012) mentioned that the ever-increasing complexity and the added dimensions of the resulting datasets not only create opportunities for further development and refinement of the computational tools, but also produce challenges for the visualisation of the results.

QTL Cartographer is a powerful, flexible and widely used program for mapping QTLs onto a genetic linkage map (Basten *et al.*, 2001). It can handle data from a variety of experimental designs and from any organism. Most of the genome-wide eQTL mapping studies in crops that were described in Section 1.5 on page 19 used QTL Cartographer for eQTL mapping. However, investigators had to either use additional post-processing tools for eQTL analysis or develop their own code in order to interpret the resulting eQTL data. The biological interpretation of detected eQTLs may include the classification of *cis*- and *trans*-eQTLs, the evaluation of functional enrichment and the modeling of causal interactions among eQTL (to identify candidate genes influencing biologic pathways)

or between eQTL and other genetic associations (phenotypic traits for example disease associated loci) (Michaelson *et al.*, 2009) (Figure 4.1).

WebQTL (`http://www.webqtl.org/`) is a web-based tool that combines databases of complex traits with software for mapping QTLs (Wang *et al.*, 2003). It also calculates correlations among traits. WebQTL includes well-curated genotype data for various supported populations and users need to provide only the quantitative trait data to identify QTLs. QTL reaper is a batch-oriented version of WebQTL, which establishes genome-wide significance via a permutation approach. Hubner *et al.* (2005) used it for an eQTL study carried out in the BXH×HXB panel of recombinant inbred rat strains using the Affymetrix microarray platform. The algorithm in QTL reaper was used to build eQTL Explorer (Mueller *et al.*, 2006), an application that enables integrated visualisation and mining of results from genome-wide linkage analyses and expression profiling. eQTL Explorer allows eQTL results across the whole genome from multiple array experiments to be displayed alongside phenotypic QTLs mapped to the genome. It consists of a Java interface as well as a relational database, to store and manage expression, linkage and external data. Upon import into the database, the software also determines and indicates whether the eQTLs are *cis*- or *trans*-acting. eQTL Explorer is useful when studying a model organism with genome, genotype and phenotypic QTL data already stored in the database.

eQTL Viewer is a web-based tool for visualising the relationships between the e-trait genes and the candidate genes in the eQTL regions using Scalable Vector Graphics. It does not include functionality for eQTL mapping. The output display is a scalable and annotated two-dimensional plot of the e-trait gene positions versus the linked eQTL positions across the genome. As a result, the eQTLs on the diagonal indicate potential *cis*-eQTLs. It provides biologists with an efficient and intuitive way to explore transcriptional regulation patterns and to generate hypotheses on the genetic basis of transcriptional regulations (Zou *et al.*, 2007).

The different types of data that need to be incorporated as well as different study objectives, especially when working with non-model organisms, make the development of flexible eQTL data analysis tools complex and challenging. There is a need for user-friendly eQTL post-processing tools and browsers that can provide assistance to those interested in mining eQTL data, also from less-studied species, for their own research.

4.2 Aims and objectives

The main aim of this chapter was to investigate the genetic basis for the response to *C. zeina* infection in the CML444×SC Malawi maize RIL population. Specific objectives were to (i) identify phenotypic QTLs for GLS severity; (ii) identify eQTLs for infected leaf tissue, i.e. identify genomic regions where genetic variation could explain gene expression differences; (iii) classify eQTLs as *cis* or *trans*; (iv) identify eQTLs that coincide with the GLS severity QTLs; (v) filter the lists of candidate genes, for genes with a significant correlation between their gene expression profiles and the GLS severity scores; and (vi) identify *trans*-eQTL hotspots linked to disease resistance or susceptibility responses; (vii) determine enriched functional categories within the candidate *trans*-eQTL hotspots.

Ultimately, the overlap between GLS severity QTLs and *cis*-eQTLs could reveal individual candidate genes that are responsible for the respective GLS QTLs; and the overlap between GLS severity QTLs and *trans*-eQTLs could reveal the mechanisms conferred by the respective GLS QTLs. A refined analysis, including only the genes with *cis*- and *trans*-eQTL peaks in the *trans*-eQTL hotspot intervals that coincide with the GLS severity QTLs, might reveal potential regulatory network models contributing to the various GLS severity QTLs.

4.3 Materials and methods

The data used in this chapter is described in Section 3.3.1 (Germplasm and field trials) and section 3.3.2 (RNA extraction and microarray analysis) from the previous chapter.

4.3.1 Construction of linkage map and QTL identification

The CIMMYT maize RIL population was derived from a CML444×SC Malawi cross (see section 3.3.1). A linkage map was previously constructed by Messmer *et al.* (2009) with 160 publicly available restriction fragment length polymorphism (RFLP) and simple sequence repeat (SSR) markers, using MapMaker v3.0 software (Lander *et al.*, 1987). The markers were tested primarily for polymorphism between the parental lines. Regions where there were gaps of approximately 20 centimorgan (cM) or more were identified and SSR markers in these regions were selected from the Maize Genetics and Genomics

database (`http://www.maizegdb.org/`). An improved version of this linkage map, named QMap 2.0, was constructed using 148 of the markers that were used by Messmer *et al.* (2009) together with 19 additional SSR markers (thus a total of 167 markers) across 145 RILs. Markers that were closer than 5cM from each other were removed to reduce possible distortion of the map. The following SSR markers were added: bnlg1811, bnlg615, umc1111, phi073, bnlg1449, bnlg1108, umc1720, bnlg105, dupssr10, umc1155, umc1572, bnlg2191, umc1413, umc1424, umc1562, umc1170, bnlg1375, umc1137, umc1337. MapManager QTX software (Manly *et al.*, 2001) was used to construct QMap 2.0 (available in the electronic Appendix) and the final genetic map was displayed using MapChart (Voorrips, 2001). The order of the markers corresponded to that from Messmer *et al.* (2009). The total map length was 1,862 cM Kosambi. It had good coverage over the 10 chromosomes, with approximately one marker every 11 cM. QMap 2.0 was used for QTL and eQTL mapping.

GLS disease severity data was recorded at 92, 99, 109 and 116 days after planting (DAP). QTL Cartographer (Basten *et al.*, 1994; Wang *et al.*, 2012*a*) was employed to map QTLs for GLS disease severity, at each of the four ratings, in the CML444×SC Malawi RIL population of 145 individuals. A walking speed of 2 cM was used in composite interval mapping (CIM) with forward regression and backward elimination (p-value = 0.1). The LR threshold was set to 11.5 after permutation testing showed that 11.5 corresponded to approximate $\alpha = 0.05$ experiment-wise (Doerge and Churchill, 1996). The eQTL data analysis pipeline was utilised for eQTL mapping (see section 4.3.2).

4.3.2 Development of a Galaxy workflow for global eQTL analysis

An eQTL data analysis pipeline was developed as part of this study. It was implemented in Python and R, and developed as a workflow in the online data analysis platform Galaxy (`http://galaxyproject.org`). The eQTL data analysis pipeline is currently an in-house tool, which employs a computer cluster at the Bioinformatics and Computational Biology Unit at the University of Pretoria for eQTL mapping. Figure 4.2 gives an overview of the six modules in the pipeline as well as the required input files. The six modules are described below in general terms and the application of the pipeline to this study are given in section 4.3.3.

Mapping eQTLs

The eQTL pipeline employs QTL Cartographer to map eQTLs. Apart from an input file that contains a matrix of the expression values (where rows correspond to genes and columns to individuals in a mapping population) called the "e-traits file", two additional files need to be generated using (Windows) QTL Cartographer beforehand: the "map file" containing information on the genetic linkage map (marker order, chromosome assignment and recombination fractions) and the "cross file" containing information on the population (the marker names, marker genotypes, trait values and other explanatory variables). This first module in the pipeline consists of three parts: the first component splits the e-traits file into 48 sub-files; the second component identifies eQTLs by running QTL Cartographer as 48 parallel tasks using different nodes on a computer cluster; and the third component concatenates the eQTL result files after the parallel runs. The module permits the setting of several parameters, including the model for stepwise regression, the model for interval mapping, the walking speed and the likelihood ratio (LR) threshold. The module also allows the inclusion of "other traits", i.e. factors that will be "regressed out" in the regression analysis.

Linking the genetic and physical maps

Before a distinction between *cis*- and *trans*-eQTLs can be made, it is necessary to link the genetic and physical maps (since eQTLs have cM positions, while genes and reporters have base pair (bp) positions). It is assumed that a draft genome sequence of the species under study is available and that the physical marker positions are known or at least estimated. The second module in the pipeline requires a "chromosome length file", containing the bp length of each chromosome, as well as a "markers file", containing the marker names, cM positions and bp positions of each marker on the genetic map. When mapping QTLs, QTL Cartographer scans the entire genome at a constant walking speed (typically 2 cM). The resulting interval positions, from here on referred to as "bins", are used in CIM and are fixed for every QTL mapping run based on the same genetic map. The marker cM and bp positions in the "markers file", are used as anchor points to proportionally estimate the bp positions at each corresponding bin. As a result, a "lookup table" is generated, containing the cM and bp positions of each bin (an example of a lookup table is provided in the electronic Appendix). It is important to note that the last interval before the next

marker is smaller than the selected walking speed if the marker spacing is not a multiple of the bin size. Every eQTL starts and ends at a specific cM position, and the lookup table makes it possible to find the corresponding eQTL coordinates on the physical map in order to extract the gene models or single-nucleotide polymorphisms (SNPs) within specified regions.

Classifying eQTLs as *cis* or *trans*

To classify eQTLs as *cis* or *trans*, a "gene positions file" that provides the bp positions of all the genes in the genome is required. The lookup table is used to proportionally estimate a cM-based position for each gene to compare to the corresponding eQTL start and end positions. This pipeline defines an eQTL that is located within a distance of 5 cM from the location of its linked gene as *cis* and an eQTL that is located further than 5 cM from the location of its linked gene, often on a different chromosome, as *trans*.

Identifying eQTL hotspots

To identify eQTL hotspots, the frequency of eQTLs and genes throughout the genome are calculated per bin. A sliding-window approach is implemented within the pipeline, where users can choose to include two or three bins per window. Whenever a bin smaller than 2 cM (e.g. the last interval before the next marker) is part of a "sliding window", an additional bin is added so that the window size for two or three bins per sliding window is always $4-6$ cM or $6-8$ cM, respectively. eQTL peaks and gene models are then counted per sliding window. With this information available, two tests are conducted: (i) whether the eQTL frequency is higher than that expected by chance, and (ii) whether gene density is an explanatory factor for eQTL hotspots.

Firstly, the expected maximum number of eQTL peaks per cM is calculated with a permutation approach. Each of the identified eQTLs are randomly assigned to 1 cM of the total number of cM on the map, and the resulting maximal number of eQTLs per bin is stored. The procedure is repeated 1000 times and the threshold corresponding to 95% of the obtained distribution established (Potokina *et al.*, 2008). This serves as a threshold to test if the eQTL frequency per sliding window is significantly high. Sliding windows for which the number of eQTL peaks per cM are above the permutation threshold are marked as potential eQTL hotspots.

Secondly, the proportion of genes to eQTL peaks per cM is calculated for each sliding window. Here the null hypothesis which is tested for each sliding window is that the proportion of genes to eQTL peaks in a specific sliding window is the same as the proportion of genes to eQTL peaks across the whole genome, i.e. that the number of eQTLs can be explained simply by local gene density. Sliding windows for which the null hypothesis is rejected, thus with a significant eQTL excess or deficiency compared to gene number, are identified (chi-squared test p-value < 0.0001).

Finally, sliding windows with (i) the number of eQTL peaks per cM are above the permutation threshold and (ii) significant eQTL excess compared to gene number, are called "unbiased" eQTL hotspots. Adjacent sliding windows that meet the specified hotspot criteria are merged to form larger hotspot regions. Three different sets of hotspots are identified, namely hotspots based on "all eQTLs", "*cis*-eQTLs only" and "*trans*-eQTLs only".

Determining enriched GO-terms per eQTL hotspot

The last module performs a GO over-representation analysis on each identified hotspot using the TopGO R package (Alexa and Rahnenfuhrer, 2010). TopGO was ideal for incorporation within the Galaxy workflow, since (i) it is it is a script-based R package, instead of a web-based tool, and (ii) it is not species-specific, i.e. one can provide an independent "gene2GO map file" including genes from any species, which can be generated from external information. A "gene2GO map file" is required as input for this analysis, listing all the genes on the array or in the genome (generally it will be the genes in the "gene positions file" mentioned above) together with their associated GO-terms. A Fisher's exact test is applied and significant GO-terms (`http://www.geneontology.org`) are listed in an output file, per hotspot. In an additional optional step, each hotspot is split according to the parental allele associated with higher expression and further GO over-representation analyses are performed on the resulting subsets.

4.3.3 Input files and use of the eQTL data analysis pipeline

The input files that were used for the eQTL analysis in this study are available in the electronic Appendix. The e-traits file consisted of microarray-based gene expression profiles for 30,280 reporters in leaf samples across 100 RILs. These were the back-converted

intensity expression profiles after removal of flagged reporters from the original set of 42,034 reporters (see section 3.3.2). Functions from Windows QTL Cartographer was used to convert the linkage map and cross information (see section 4.3.1) into a format suitable for QTL Cartographer (i.e. a map file and a cross file were generated). eQTL mapping was performed using the parameters mentioned above for QTL mapping, i.e. forward and backward stepwise regression (p-value = 0.1), a 2 cM walking speed and CIM was implemented. The LR threshold for eQTL mapping was set to 11.5.

Since the genome sequences of the two parental maize lines, CML444 and SC Malawi, were not known, the maize B73 reference genome (RefGen) v2 (Schnable *et al.*, 2009) was used as a physical map. Sequences of all 10 chromosomes (in FASTA format) were downloaded from the maizesequence.org FTP site (`http://ftp.maizesequence.org/current/`). The chromosome length file was constructed with this information; it consisted of the base pair (bp) length of each chromosome in order to assist in linking the genetic and physical maps. To generate the markers file (for which cM and bp positions of each marker was needed), the MaizeGDB locus lookup tool (Andorf *et al.*, 2010) was used to extract the physical positions of most markers. In cases where the physical coordinates on the B73 genome sequence were not available, primer or marker sequences were downloaded from MaizeGDB (`http://www.maizegdb.org`) and located on the maize B73 reference genome v2.0 using the basic local alignment search tool (BLAST) (Altschul *et al.*, 1990).

The gene positions file consisted of the start and end bp positions, on the B73 RefGen v2 genome sequence, of the gene models representing each e-trait reporter. The reporter positions were necessary for classification of eQTLs as *cis* or *trans*. BLASTN results of the 42,034 60-mer microarray reporter sequences were used to assign the reporters into six genomic annotation groups (Chapter 2; Coetzer *et al.*, 2011). Information on the reporters in the following annotation groups were included (32,937 reporters) in the file: "annotated by sense gene model"; "annotated by antisense gene model"; "annotated by gDNA" (reporters without a working gene set (WGS) transcript hit); and "annotated by EST" (in which case the EST from which the reporter was designed, but not the reporter itself, has a WGS transcript hit). Reporters in the "ambiguous annotation" and "inconclusive annotation" groups were excluded (9,097 reporters), since no single gene position could be assigned for these reporters. Therefore, the reporter positions of 32,937

reporters were included in the gene positions file.

For eQTL hotspot identification, two 2 cM bins per sliding window were selected. This resulted in 4-6 cM sized sliding windows. The average cM size of an eQTL region was 10.9 cM. Thus a full eQTL region was mostly not included in a 4-6 cM sized sliding window, however a significant part of each eQTL region surrounding its peak was included.

In order to generate a suitable "gene2GO mapping" input file for GO enrichment analysis, the "Zea mays V5a" GO annotation file was downloaded from the AgriGO (a web-based GO analysis toolkit for the agricultural community) website (`http://bioinfo.cau.edu.cn/agriGO/`). This file consisted of maize gene IDs annotated with GO-terms. According to the reporter−gene model annotations from Coetzer *et al.* (2011), GO-terms were assigned to the matching Agilent reporter IDs. All the genes that were represented on the array (the genes in the "gene positions file" mentioned above) together with their associated GO-terms were included in the gene2GO mapping file.

4.3.4 Overlap analysis between QTLs and eQTLs

An eQTL was said to overlap a GLS severity QTL if it spanned at least one common 2 cM bin. A customised python script was developed to extract the QTL-overlapping eQTLs and the gene expression profiles of the genes to which these eQTLs belonged.

Genes with c*is*-eQTLs that overlapped the GLS severity QTLs were identified as candidates that could be responsible for the respective GLS severity QTLs. Overlapping *cis*-eQTLs (per GLS severity QTL) were divided into two groups, based on the parental allele associated with higher expression. The Pearson correlation coefficient of each *cis*-eQTL gene's expression profile with the GLS severity scores was calculated and genes with an associated p-value < 0.01 (experiment-wise) were identified as the best candidates (Tables S4.1 and S4.2 in the electronic Appendix). The null hypothesis in each case stated that there was no linear relationship between the *cis*-eQTL gene's expression profile and the GLS severity scores, i.e. that the value of the Pearson correlation coefficient was zero. In cases where the p-value was significantly small, this null hypothesis was rejected and a significant correlation was inferred.

Genes with *trans*-eQTLs that overlapped the GLS severity QTLs were identified as candidates involved in the mechanisms that might explain the respective GLS severity QTLs. Overlapping *trans*-eQTLs (per GLS severity QTL) were also divided into two

groups, based on the parental allele associated with higher expression. The Pearson correlation coefficient of each *trans*-eQTL gene's expression profile with the GLS severity scores were calculated and genes with an associated p-value < 0.01 (experiment-wise) were identified. The null hypothesis in each case stated that there was no linear relationship between the *trans*-eQTL gene's expression profile and the GLS severity scores. GO enrichment analysis were performed on different sets of overlapping *trans*-eQTL genes (Table S4.3 in the electronic Appendix).

4.3.5 Overlap analysis between QTLs and *trans*-eQTL hotspots; and gene regulatory network reconstruction

A *trans*-eQTL hotspot was said to overlap a GLS severity QTL if it spanned at least one common 2 cM bin. Note that a gene belonged to a *trans*-eQTL hotspot if it had a *trans*-eQTL with a peak in the identified hotspot interval. After the QTL-overlapping *trans*-eQTL hotspots were identified, the eQTLs within each of the identified hotspots were extracted (output from the eQTL data analysis pipeline) and a customised python script was used to extract the gene expression profiles of the genes affected by these eQTLs. For each identified hotspot, the *trans*-eQTLs were split into two groups based on the parental allele associated with higher expression and the *trans*-eQTL genes whose gene expression profiles significantly correlated to the GLS severity scores were extracted for further analysis (p-value < 0.01). In order to identify genes that potentially regulated the identified groups of genes with *trans*-eQTLs, genes with c*is*-eQTL peaks within the same hotspot intervals were identified. These *cis*-eQTL genes were split into two groups based on the parental allele associated with higher expression and *cis*-eQTL genes whose gene expression profiles significantly correlated to the GLS severity scores were extracted for further analysis (p-value < 0.01). This overlap analysis was considered a "refined" analysis, since the bulk of genes included in the analysis was a subset of the previously mentioned overlap analysis (see section 4.3.4).

Subsequently, for each hotspot, the candidate genes with *cis*-eQTLs for which a specific allele (e.g. CML444) was associated with increased expression and the candidate genes with *trans*-eQTLs for which the same allele was associated with increased expression, were grouped together as a basis for regulatory network reconstruction. For each group, the Pearson correlation coefficients (and associated p-values) between the gene

expression profiles of each identified *cis*-eQTL gene and all the *trans*-eQTL genes in its group were calculated. The null hypothesis that was tested for each pair of gene expression profiles stated that there was no linear relationship between the gene expression values of the *cis*- and *trans*-eQTL genes. In cases where the p-value was < 0.00001, this null hypothesis was rejected and a significant correlation between the gene expression profiles was inferred.

Therefore, for each identified hotspot per parental allele associated with higher expression, *cis*- and *trans*-eQTL genes with significantly correlated gene expression profiles were included in "regulatory network" models per GLS severity QTL, where *cis*-eQTL genes were potential regulators of *trans*-eQTL genes (Figures 4.9, 4.10, 4.11, 4.12, 4.13, 4.14, 4.15). The R package igraph (Csardi and Nepusz, 2006) was used to create directed graphs for visualisation of the putative regulatory networks. Nodes represented genes and directed edges connected genes with *cis*-eQTLs to highly co-expressed (correlated) genes with *trans*-eQTLs. The phenotypic trait "GLS severity" was included as an additional node in each network. Black dotted lines were used to indicate a remarkably strong correlation (p-value <0.00001) between GLS resistance or susceptibility and an eQTL gene's expression profile (the previous filter for correlation between GLS severity and the eQTL gene expression profiles was: p-value <0.01). No annotation filters were applied before the regulatory networks were constructed. To assist with interpretation, predefined functional categories from MapMan, distinguished by colour, were used to group eQTL genes per network. Table 4.7 gives the functional categories as well as the colours that were assigned to each category.

4.3.6 Functional annotation and GO over-representation analysis

Out of the 30,280 e-traits (that remained after removal of flagged reporters from the original set of 42,034 reporters), 23,848 (78.8%) were assigned a single maize gene ID according to the Maize Microarray Annotation Database (reporters in annotation groups "annotated by sense gene model", "annotated by antisense gene model" and "annotated by EST" (Chapter 2; Coetzer *et al.*, 2011)). A *Z. mays* annotation file, which was released as part of Phytozome version 7.0 (`http://www.phytozome.net`), was downloaded from the FTP site. The file included the best *Arabidopsis* TAIR10 and rice BLAST hits for each maize gene. The resulting *Arabidopsis* and rice hit descriptions together with the

BLAST2GO description for each gene (which was extracted from the Maize Microarray Annotation Database), were used to formulate a final functional annotation per reporter.

TopGO was used to identify enriched GO-terms per eQTL hotspot as part of the eQTL data analsis pipeline (see section 4.3.2). However, due to more specific functional annotations that are available for *Arabidopsis* (compared to maize), BiNGO (Maere *et al.*, 2005) was also used in additional analyses to identify enriched GO-terms in order to determine whether genes in the same *trans*-eQTL hotspots (or sub-groups) were involved in the same biological processes. "Best BLAST hit" IDs from Phytozome corresponding to the relevant maize genes were used as input to the BiNGO analyses. Default BiNGO parameters were used and a reference set of 11,291 *Arabidopsis* IDs corresponded to the 30,280 e-trait reporters. As an alternative to using the full GO hierarchy, BiNGO provides several GOSlim ontologies that are organism-specific slimmed-down versions of the full GO hierarchy. GOSlim ontologies generally give a broad overview of the ontology content without the detail of the specific fine-grained terms. In cases where the full GO hierarchy did not produce significantly enriched GO-terms, additional BiNGO analyses based on the plant GOSlim ontology were performed. All BiNGO GO enrichment output tables list the enriched GO-terms from the three categories (i.e. biological process, molecular function and cellular component) together in one table, sorted by significance (Table S4.3 in the electronic Appendix and Table 4.6).

In addition, MapMan was used to functionally classify genes into predefined bins (Thimm *et al.*, 2004). MapMan's classification, together with manual revision, were used to group eQTL genes into functional categories. The MapMan ontology comprises a set of 34 tree-structured bins, describing a variety of cellular processes.

4.4 Results and discussion

4.4.1 Identification of QTLs for a *C. zeina*-infected maize RIL population

A phenotypic QTL analysis was performed to identify the regions of the genome where genetic variation explained GLS disease severity differences in the CML444×SC Malawi RIL population. From the GLS disease severity data that was collected at 92, 99, 109

and 116 days after planting (DAP) from the field trial at Baynesfield Estate (2008/2009 season), GLS disease severity QTLs were mapped. Note that the samples for gene expression profiling (and subsequent eQTL analysis) were collected at 103 DAP, from the same field trial. Eight consensus QTLs were identified (Table 4.1). For six of the eight QTLs, CML444 (the more resistant parent) was the parent with the resistance associated allele, whereas SC Malawi (the more susceptible parent) was the parent with the resistance associated allele for the other two QTLs (Table 4.1). It was concluded that complex genetic interactions appeared to lie at the foundation of the *C. zeina* disease response in maize.

QTL 9-7 was consistently present in three out of the four ratings (the first three time points: 92, 99 and 109 DAP) and accounted for 12% of the total phenotypic variation. QTL 9-7 also had the highest peak LR of 22.99, thus the strongest marker-trait association. Interestingly, SC Malawi was the parent with the resistance associated allele for this QTL even though it is regarded as the more susceptible parent. Since both parents are neither fully resistant nor fully susceptible, it could be expected that the more resistant parent would be the source of most of the resistance associated alleles, but that the susceptible parent could also contribute to resistance. QTL 6-13 and QTL 10-10 were detected in the latest two ratings (109 and 116 DAP). These QTLs accounted for 18% and 14% of the total phenotypic variation, respectively. QTL 10-10 had the second largest peak LR of 19.67 and CML444 was the parent with the resistance associated allele. QTL 6-13 had a peak LR of 17.86 and SC Malawi was the parent with the resistance associated allele. QTL 3-3, QTL 3-14 and QTL 5-3 were detected in the first and either the second or third ratings. These QTLs accounted for 9%, 10% and 7% of the total phenotypic variation and had below average peak LRs of 12.68, 16.15 and 12.82, respectively (the average for all eight QTL was 16.24). QTL 4-11 and QTL 9-5 were only detected in the earliest rating (92 DAP), accounted for 7% and 9% of the total phenotypic variation, and had peak LRs of 14.10 and 13.67, respectively. For the five QTL lastly mentioned, CML444 was the parent with the resistance associated allele.

The lookup table (available in the electronic Appendix), which is an output from of eQTL data analysis pipeline (see section 4.3.2), was used to proportionally estimate the start, peak and end bp positions of each GLS severity QTL region. QTL 9-5 spanned the largest region of 36 Megabase (Mb), whereas the remaining seven QTLs spanned

regions of between 3.5 and 12 Mb. However, on the cM scale QTL 9-5 spanned a region of only 15.08 cM, which is only slightly more than the average of 14.44 cM across the eight QTL. On this scale, QTL 3-14 spanned the largest region of 20.35 cM (12.27 Mb) and the remaining seven QTLs spanned regions of between 8.54 and 18.38 cM.

4.4.2 Global analysis of eQTLs in *C. zeina*-challenged leaves using the CML444×SC Malawi maize RIL population

Global eQTL analysis was performed to identify regions of the genome where genetic variation explained gene expression differences in the CML444×SC Malawi RIL population. Subsequent analyses were aimed at finding overlap between GLS severity QTLs and either individual eQTLs or eQTL hotspots. The overlap between GLS severity QTLs and *cis*-eQTLs might reveal candidate genes causing the respective GLS severity QTLs, whereas the overlap between GLS severity QTLs and *trans*-eQTLs might reveal mechanisms conferred by the respective GLS severity QTLs. Furthermore, when a *trans*-eQTL hotspot overlaps a GLS severity QTL, one could hypothesise that there may be a polymorphism in a gene that affects the expression of many related genes and thereby underlies a GLS severity QTL.

Before the above-mentioned analyses could be carried out, the following main steps were necessary: (i) identification of eQTLs for the 30, 280 e-traits, (ii) classification of the resulting eQTLs as *cis* or *trans*, and (iii) identification of eQTL hotspots. For this purpose, an eQTL data analysis pipeline was developed as part of this study and developed as a workflow in the online data analysis platform Galaxy (Figure 4.2).

eQTL identification using the eQTL data analysis pipeline

Out of the 30,280 initial reporters, 24,732 (81.7%) were identified to belong to a single genomic location (these were not part of the "ambiguous" or "inconclusive" annotation groups from Coetzer *et al.*, 2011). Furthermore, the 24,732 reporters represented 17,250 unique maize gene models (according to the analysis by Coetzer *et al.*, 2011; see Chapter 2). Thus approximately 30% of the 63,331 protein coding gene models in the working gene set (WGS) and 40% of the 39,656 gene models in the functional gene set (FGS) was included as e-traits in this study (FGS is a subset of the WGS in which transcripts that are "probable pseudogene", "possible transposon", "contamination" or "low confidence"

have been filtered out). Therefore, at least 60% of the genes in the maize genome were not included in this study.

From the 30,280 input reporter expression profiles across the individuals in the RIL population i.e. e-traits, 31,549 eQTLs were identified based on a LR threshold of 11.5 (LOD=2.5). These eQTLs explained the expression of 18,000 reporters (thus no eQTLs were identified for the remaining 12,280). Out of the 18,000 reporters with eQTLs, 15,006 were identified to belong to a single genomic location representing 12,035 unique maize gene models (Coetzer *et al.*, 2011; see Chapter 2). Therefore, eQTLs were identified for less than 30% of the maize gene models. Approximately 80% of the reporters had only one eQTL, whereas 20% of the reporters' expression profiles were affected by up to nine eQTLs.

Cis/*trans* classification using the eQTL data analysis pipeline

Before eQTLs could be classified as *cis* or *trans*, the "lookup table" (available in the electronic Appendix; see section 4.3.2 on page 149) was used to identify the 2 cM bin where each Agilent reporter was located and subsequently to proportionally estimate a cM position per Agilent reporter. An eQTL that was located within 5 cM from the location of its linked gene was classified as *cis* and an eQTL that was located further than 5 cM from the location of its linked gene, often on a different chromosome, was classified as *trans*. Table 4.2 gives a summary per chromosome, of the numbers of markers, 2 cM bins, reporters as well as *cis*- and *trans*-eQTLs. The average number of 2 cM bins (sometimes smaller than 2 cM) per chromosome was 101 and the bins had an average physical size of 2 Mb.

Seventeen percent of the identified eQTL (5,258/31,549) could not be classified due to uncertainty concerning the genomic location of its linked genes, i.e. the reporters to which these eQTL belonged were part of the "ambiguous" or "inconclusive" annotation groups from Coetzer *et al.* (2011). One-fifth of the classifiable eQTLs were identified as *cis*-eQTLs: 19% were *cis*-eQTLs and 81% were *trans*-eQTLs. Of the 15,006 reporters with classifiable eQTLs, 12% had only *cis*-eQTLs, 67% had only *trans*-eQTLs and 20% had *cis*- and *trans*-eQTLs.

Various different ways of determining whether a gene is locally regulated were previously reported. For the studies that were compared in section 1.5 on page 19, most

authors classified eQTLs as *cis* if the gene whose expression profile is affected by the eQTL was less than a fixed cM distance away from the eQTL. Popular intervals were 3.5 cM (West *et al.*, 2007), 5 cM (Potokina *et al.*, 2008; Swanson-Wagner *et al.*, 2009) and 10 cM (Holloway and Li, 2010). West *et al.* (2007) reported that using a 5 cM distance instead of a 3.5 cM distance had minimal effect on the number of *cis*-eQTLs identified. Kloosterman *et al.* (2012) used the same linkage group as a criterion for identifying *cis*-eQTL. Keurentjes *et al.* (2008) and Wang *et al.* (2010*b*) calculated support intervals per eQTL and when the gene's position coincided with the support interval, classified it as *cis*-acting.

For the eQTL dataset in the current study, different rules for eQTL classification were applied to test the effect on the percentage *cis*-eQTLs detected (data not shown). Classifying eQTLs as *cis* if its linked gene's position was less than 1 cM away from the eQTL peak versus if its linked gene's position was less than 5 cM away from the eQTL peak, resulted in a 2% difference in the percentage of *cis*-eQTLs detected. Furthermore, classifying eQTLs as *cis* if its linked gene's position was less than 1 cM away from the eQTL peak versus if its linked gene's position is on the same chromosome than the eQTL, resulted in a 10% difference in the percentage of *cis*-eQTLs detected. Thus different rules for eQTL classification for this dataset resulted in a change in ratio between *cis*- and *trans*-eQTLs. A decision was made to continue with the criterion for classification set to a distance of 5 cM.

Figure 4.3 on page 194 gives a scatter plot of the genomic relationships between the eQTL positions and the corresponding e-trait gene positions. The eQTL and e-trait positions corresponded to the 1009 bins (from the lookup table) across the genome. The *cis*-eQTLs (on the diagonal) appeared to be roughly evenly spread throughout the genome. Horizontal bands indicate gene-rich regions, for example the region in the middle of chromosome 2; and vertical bands indicate eQTL-rich regions, for example at the end of chromosome 5.

It was hypothesised that the *cis*-eQTLs were mainly larger-effect polymorphisms, whereas *trans*-eQTLs were mainly smaller-effect polymorphisms. This hypothesis was based on three reasons: (i) *cis*-eQTL sequence polymorphisms (e.g. in a promoter) have a direct influence on expression of a gene giving rise to a *cis*-eQTL and *trans*-eQTLs are caused by a polymorphism located elsewhere in the genome (e.g. in a regulatory

factor) (Hansen *et al.*, 2008); (ii) gene expression of most genes are regulated by multiple factors and thus a polymorphism in one regulatory factor might only result in a small change in the expression of genes controlled in *trans* by that polymorphism; and (iii) the polymorphism underlying a *trans*-eQTL typically affects numerous other genes and could therefore be pleiotropic. Large-effect mutations in pleiotropic genes are likely to be deleterious and, as such, there might be a constraint on the effect size of *trans*-eQTL loci. The hypothesis was tested by comparing for *cis*- and *trans*-eQTLs: (i) the average peak LR, which gives the likelihood that an eQTL truly exists in the region of the marker; and (ii) the average proportion of variation explained, which gives an indication of the relative importance of an eQTL in influencing gene expression variation for a specific e-trait. In this study, the average peak LR was 36 for *cis*-eQTLs and 15.6 for *trans*-eQTLs; and the average proportion of variation explained was 29% for *cis*-eQTLs and 13% for *trans*-eQTLs. Since the average peak LR and proportion of variation explained were more than double for *cis*-eQTLs than for *trans*-eQTLs, this result confirmed the hypothesis that *cis*-eQTLs were mainly larger-effect polymorphisms, whereas *trans*-eQTLs were mainly smaller-effect polymorphisms.

It could further be hypothesised that most of the large-effect (mainly *cis*) eQTLs were detected in this study, but that the numerous small-effect eQTLs (mainly *trans*) remained undetected. An increase in the number of RILs included in the analysis and the number of replicates performed might result in more statistical power for the identification of even more small-effect eQTLs (Mackay, 2001).

Identification of *trans*-eQTL hotspots using the eQTL data analysis pipeline

Figure 4.4 gives the frequency distribution of genes, *cis*- and *trans*-eQTLs per sliding window bin across the 10 chromosomes. Two 2 cM bins per sliding window were selected and as a result the sliding window sizes varied between 4 and 5.9 cM (since when one of the two bins were smaller that 2 cM a third was added). This figure highlights gene-rich regions as well as potential *trans*-eQTL hotspots across the genome. Also, there appear to be fewer *cis*-eQTL hotspots than *trans*-eQTL hotspots.

The first aim towards identifying significant *trans*-eQTL hotspots was to identify those sliding windows for which the number of *trans*-eQTLs per cM was significantly higher than expected by chance. A permutation approach was used to calculate this

threshold, indicated by the horizontal line in Figure 4.5 (the calculated threshold for *trans*-eQTLs was 27). Since these apparent *trans*-eQTL hotspots may reflect regions of the maize genome with little recombination or higher gene density, and therefore more genes per cM than elsewhere in the genome, the second aim was to determine whether gene density was an explanatory factor for *trans*-eQTL hotspots. The proportion of reporters to *trans*-eQTLs was 0.6:0.4. Thus on average, for every three reporters, two *trans*-eQTLs were expected. The null hypothesis, that the proportion of genes to eQTL peaks in a specific sliding window is the same than the proportion of genes to eQTL peaks across the whole genome, was tested for each sliding window. Finally, sliding windows where the number of eQTLs per cM was above the permutation threshold and with significant eQTL excess compared to reporter number (chi-squared test p-value < 0.0001), were declared "unbiased" eQTL hotspots. Thirty-two significant *trans*-eQTL hotspots were identified and are marked in red in Figure 4.5. Table 4.3 gives a summary of the 32 *trans*-eQTL hotspots. The average size of a *trans*-eQTL hotspot was 6.4 cM or 4.2 Mb.

A significant directional bias was evident for three quarters of the 32 *trans*-eQTL hotspots (Table 4.3), such that the same parental allele was associated with higher expression for most of the transcripts in with a hotspot (Pearson's chi-squared test was used to test for excess of positive alleles from one parent; $p < 0.05$). For 11 *trans*-eQTL hotspots, the CML444 allele was associated with higher expression (significant positive effect); and for 13 other hotspots, the opposite allelic effect was observed (in this case either the SC Malawi allele significantly increased accumulation or the CML444 allele significantly decreased accumulation). The directional bias for the hotspots was in contrast to the global average for the 31,549 eQTLs, where 50.6% of had a positive CML444 allelic effect and 49.4% had a positive SC Malawi allelic effect (data not shown).

The TopGO package in R was employed via the eQTL data analysis pipeline to identify significantly over-represented GO-terms per *trans*-eQTL hotspot. The average number of enriched GO-terms per hotspot was 13.2 with an unadjusted p-value < 0.01, and 0.3 with an adjusted p-value < 0.05 (Table 4.3). GO-terms that are enriched in lists of genes that share a *trans*-eQTL hotspot may reveal the involvement of these genes in a shared biological process. Since gene expression data of *C. zeina*-infected maize plants was used in the eQTL analysis, it was expected that some of the eQTLs and eQTL hotspots would reveal processes associated with the GLS disease response. However, it should be borne

in mind that some *trans*-eQTL hotspots may reflect transcriptional responses that are unrelated to defences against GLS. Also, some aspects of leaf development and structure may in fact indirectly affect resistance/susceptibility and one would not recognise it as resistance associated. The genes with *trans*-eQTLs per hotspot can also be exported from Galaxy and imported to other tools, such as BiNGO, for GO enrichment analysis.

4.4.3 Overlap analysis between QTLs and eQTLs to identify genes and pathways involved in the GLS disease response

QTLs may define relatively large regions of the genome and the identification of genes responsible for the respective QTLs may be difficult. In the current study, there are on average 508 gene models per GLS severity QTL interval. A gene with a *cis*-eQTL is generally a good candidate for also explaining the phenotypic QTL, assuming that the polymorphism responsible for the *cis*-eQTL is also responsible for the phenotypic QTL. Therefore, an overlap analysis of GLS severity QTLs and *cis*-eQTLs was carried out to narrow down and prioritise the list of causal candidate genes explaining the GLS disease response. In addition, the overlap between phenotypic QTLs and *trans*-eQTLs could reveal the mechanisms conferred by the respective phenotypic QTLs. This is easiest understood for the case where a gene causing the phenotypic QTL is a transcriptional regulator that would influence the expression levels of downstream genes with *trans*-eQTLs at the position of the regulator (and hence also of the phenotypic QTL). However, a polymorphism in almost any gene can affect many other genes; it does not have to be a transcriptional regulator. Functional enrichment analyses of the genes with *trans*-eQTLs that overlapped the respective GLS severity QTLs might reveal these mechanisms.

The top section of Table 4.4 gives a summary of the numbers of gene models and eQTLs that overlapped each of the eight GLS severity QTLs. Sixteen percent (640/4,060) of the eQTLs that overlapped the eight GLS severity QTLs could not be classified (due to uncertainty concerning the genomic location of its linked genes). Furthermore, 19% of the classifiable eQTLs that overlapped the GLS severity QTLs were identified as *cis*-eQTLs and 81% were *trans*-eQTLs. These results were similar to the results for the full set of eQTLs across the entire genome (see section 4.4.2). On average, there were 508 gene models per GLS severity QTL and approximately 18% of these had *cis*-eQTLs.

Figure 4.6 gives an overview of the process that was followed to narrow down the

candidate genes responsible for and those likely to play a role in GLS disease. Out of the 63,331 protein coding gene models in the maize B73 RefGen v2 working gene set (WGS), 17,250 gene models were included as e-traits in the eQTL analysis. These gene models were represented by 24,732 reporters on the Agilent microarray that were annotated to belong to a single genomic location (Coetzer *et al.*, 2011). In Figure 4.6, part (a) considered the full overlap between all GLS severity QTLs and eQTLs (the analysis discussed in the current section), whereas part (b) focused on the subset of eQTLs with peaks in the *trans*-eQTL hotspots that overlapped the GLS severity QTLs (see section 4.4.4). The aim with both analyses was the same, i.e. to identify candidate genes and pathways (and ultimately to predict regulatory networks) associated with GLS resistance or susceptibility.

Filtering of *cis*-eQTL candidates underlying GLS severity QTLs

Assuming that there is an underlying DNA polymorphism that gives rise to a change in gene expression which in turn affects GLS severity, all reporters with *cis*-eQTLs that overlapped the GLS severity QTLs were identified as candidates potentially affecting the phenotypic trait. A total of 654 reporters with *cis*-eQTLs, representing 590 gene models (Figure 4.6 (a)), overlapped the eight GLS severity QTLs (see the electronic Appendix for a full list of the QTL-overlapping *cis*-eQTLs). This difference between the number of reporters and gene models was mainly attributed to reporters in the gDNA annotation group (see Coetzer *et al.*, 2011) that were not linked to gene models; these could be classified as *cis*-eQTLs due to their known position on the genome. A second reason was that more than one reporter sometimes represented the same gene model.

Further assuming that the allele associated with the trait of interest could be associated with higher or lower expression depending on the underlying polymorphism, the *cis*-eQTL candidate genes were split into two groups based on whether higher expression was associated with the allele linked to resistance or susceptibility (Figure 4.6 (a)). Importantly, for each GLS severity QTL, Table 4.1 stated whether the allele associated with resistance was inherited from CML444 or SC Malawi; hence in each case the other parental allele was associated with susceptibility. For convenience in the rest of the text "QTL 10-10 R", for example, will be used to refer to the GLS severity QTL (on chromosome 10, starting at marker 10) where the resistance associated allele was associated

with higher expression (compared to the susceptibility associated allele).

It is generally expected that a causal gene's expression profile will correlate with the quantitative trait of interest (Mackay *et al.*, 2009). A significant correlation between a gene's expression profile and the GLS severity scores can either be due to a positive or a negative linear relationship between the gene expression values and the GLS severity scores. Figure 4.7 uses two examples to illustrate that a negative correlation between a gene's expression profile and the GLS severity scores indicates a correlation with GLS resistance and a positive correlation indicates a correlation with GLS susceptibility. This is due to low GLS severity scores of 1 meaning resistant and a high GLS severity score of 9 meaning susceptible. Figure 4.7 (a) gives an example of a gene whose expression values correlate with GLS resistance and Figure 4.7 (b) an example of a gene whose expression values correlate with GLS susceptibility.

Table S4.1 (in the electronic Appendix) gives the annotations of the 36 reporters (representing 35 gene models; Figure 4.6 (a)) where: (i) the susceptibility associated parental allele was linked with higher expression; (ii) and a significant (p-value < 0.01) correlation to GLS severity was evident. The predominantly positive correlation coefficients in these tables confirm that these genes are higher expressed in susceptible plants and lower expressed in resistant plants. Table S4.2 (in the electronic Appendix) gives the annotations of the 59 reporters (representing 53 gene models; Figure 4.6 (a)) where: (i) the resistance associated parental allele was linked with higher expression; and (ii) a significant correlation (p-value < 0.01) to GLS severity was evident. The predominantly negative correlation coefficients in these tables confirm that these genes are higher expressed in resistant plants and lower expressed in susceptible plants. The correlation analysis brought the number of candidate causal gene models with *cis*-eQTLs down from an average of 74 to an average of 11 per GLS severity QTL (Figure 4.6 (a)). The middle section of Table 4.4 gives a breakdown of the numbers of *cis*-eQTLs that overlapped each GLS severity QTL after: (i) it was divided into groups based on the parental allele associated with higher expression; and (ii) genes that significantly correlated to GLS severity were identified (p-value < 0.01).

Filtering of *trans*-eQTL candidates underlying GLS severity QTLs

Genes with *trans*-eQTLs that overlapped the GLS severity QTLs were identified as candidate genes involved in the biological processes, or mechanisms that might explain the respective GLS severity QTLs. A total of 2,766 reporters with *trans*-eQTLs (Table 4.4), representing 2,636 gene models (Figure 4.6 (a)), overlapped the eight GLS severity QTLs (see the electronic Appendix for a full list of the QTL-overlapping *trans*-eQTLs). Again assuming that the allele associated with the trait of interest was also associated with higher expression depending on the underlying polymorphism, the *trans*-eQTL genes were split into two groups based on allele associated with higher expression (Figure 4.6 (a)).

Since a causal gene's expression profile was expected to correlate (positively or negatively) with the quantitative trait of interest, the expression profiles of the response genes being regulated by the causal gene can also be expected to correlate with the trait of interest. *Trans*-eQTL candidate genes with a significant correlation to GLS severity were identified (Figure 4.6 (a)) and a GO over-representation analysis was performed on various sets of *trans*-eQTL genes that overlapped the respective GLS severity QTLs to elucidate the mechanisms that these genes were involved in (Table S4.3 in the electronic Appendix). The bottom section of Table 4.4 gives a breakdown of the numbers of *trans*-eQTLs that overlapped each GLS severity QTL after: (i) it was divided into groups based on the parental allele associated with higher expression; and (ii) genes that significantly correlated to GLS severity were identified (p-value < 0.01).

No direct and obvious relationships were identified between the annotations of the *cis*-eQTL candidate genes (see Tables S4.1 and S4.2) and the biological processes identified in the GO enrichment analysis (see Tables S4.3), for the respective GLS severity QTLs. This could be due to a few reasons regarding the "true regulatory gene": it could be that (i) it was not present on the microarray; (ii) its *cis*-eQTL effect was too small, so that it was not detected as an eQTL; (iii) its gene expression profile did not correlate well with GLS severity profile across the RILs (not a significant correlation); (iv) it was not well-annotated (or it was mis-annotated), so that it was not recognised in Tables S4.1 and S4.2; or (v) it was recognised in Tables S4.1 and S4.2, but it is a transcriptional regulator with multiple functions or one that indirectly regulates specific processes. Finally it is likely that some GLS severity QTLs are caused by polymorphisms that do not give rise

to expression variation, but rather to variation in gene splicing or protein sequence.

4.4.4 Exploiting *trans*-eQTL hotspots to identify candidate genes and pathways that play a role in the GLS disease response

The *trans*-eQTL hotspots that were previously identified using the eQTL data analysis pipeline (Section 4.4.2), were hypothesised to each disclose a significant number of response genes regulated by one or more regulatory genes within the respective hotspot loci. The hotspot regions that overlapped the GLS severity QTLs were found to be considerably smaller than the GLS severity QTL intervals (Figure 4.8); and only genes with *trans*-eQTL peaks in a hotspot region were considered part of the hotspot, since the eQTL peak can be considered the most likely position of an eQTL causing polymorphism. Therefore, a "refined" analysis based on only those genes with *trans*-eQTL peaks in each QTL-overlapping hotspot region was performed to gain a more focused perspective regarding the mechanisms and pathways that might affect GLS severity (Figure 4.6 (b)).

Additionally, it was hypothesised that the polymorphism(s) explaining each overlapping *trans*-eQTL hotspot might be *cis* variation in a gene that affects the expression of many related genes and thereby underlies a GLS severity QTL. Consequently, genes with *cis*-eQTL peaks in the same confined hotspot regions were also investigated. The identified genes might suggest a "model" per QTL-overlapping *trans*-eQTL hotspot, which could suggest hypotheses of which genes with *cis*-eQTLs potentially control processes or mechanisms responsible for GLS resistance or susceptibility, respectively.

The presence of *trans*-eQTL hotspots overlapping phenotypic QTLs, suggest possible epistatic interactions among these QTLs. However, no significant epistatic effects between any pairwise combination of GLS severity QTL was observed (data not shown) following an assessment by using the Multiple Interval Mapping (MIM) utility in Windows QTL Cartographer (Balint-Kurti *et al.*, 2008).

Filtering of genes with eQTLs in QTL-overlapping *trans*-eQTL hotspots

Out of the 32 genome-wide *trans*-eQTL hotspots (Table 4.3), five were identified to overlap GLS severity QTLs. The five hotspots of interest were the last hotspots on chromosomes 3, 4 and 10 and the first two hotspots on chromosome 9 (Figure 4.5). All

five *trans*-eQTL hotspot regions were less than half the size of its overlapping GLS severity QTL region (Figure 4.8).

A similar strategy than the one implemented for the QTL-eQTL overlap analysis (Figure 4.6 (a)), was also implemented to narrow down the genes with eQTL peaks in the QTL overlapping hotspot regions (Figure 4.6 (b)). A total of 1,375 reporters (Table 4.5), representing 1,321 gene models, were identified with *trans*-eQTL peaks in the five *trans*-eQTL hotspots that overlapped the GLS severity QTLs. The difference between the number of reporters and the number of gene models, in this case, was mainly due to genes with *trans*-eQTLs in more than one QTL overlapping hotspot. The *trans*-eQTL candidate genes were split into two groups: for 494 the resistance associated allele was associated with higher expression; and for 827 the susceptibility associated allele was associated with higher expression (Figure 4.6 (b)). For convenience, *trans*-eQTL hotspots were named such that "HS 10-10 R" referred the *trans*-eQTL hotspot (on chromosome 10, starting at marker 10 on QMap 2.0) overlapping a GLS severity QTL where the resistance associated allele (that was determined by the overlapping GLS severity QTL; Table 4.1) was associated with higher expression.

After filtering for *trans*-eQTL genes with a significant correlation to GLS severity (p-value < 0.01), the number of eQTLs where the resistance associated allele was associated with higher expression dropped to 129 reporters (representing 126 gene models) and the number of eQTLs where the susceptibility associated allele was associated with higher expression dropped to 261 reporters (representing 256 gene models). The bottom section of Table 4.5 gives a breakdown per QTL-overlapping *trans*-eQTL hotspot, of the numbers of *trans*-eQTLs with peaks in the hotspot regions, after (i) it was divided into groups based on the parental allele associated with higher expression; and (ii) genes that significantly correlated to GLS severity were identified (p-value < 0.01). Interestingly, for each of the identified overlapping hotspots, more *trans*-eQTL genes had an increased expression positively associated with the allele of the susceptible plant compared to the allele of the resistant plant (bottom section of Table 4.5).

A GO overrepresentation analysis was performed on sets of *trans*-eQTL genes that belonged to the respective QTL-overlapping *trans*-eQTL hotspots, to identify the biological processes that these genes were involved in (Figure 4.6 (b)). Over-represented GO-terms were identified for three out of the five hotspots (Table 4.6). The 121 genes with *trans*-

eQTLs that were regulated from HS 9-6 S whose gene expression profiles significantly correlated to GLS severity (Table 4.5), were enriched for cell wall-related genes ("HS 9-6 S cor" in Table 4.6); similar to the genes with *trans*-eQTLs in "QTL 9-5 S cor" (Table S4.3). Also, the 74 *trans*-eQTL genes that were regulated from HS 10-10 S whose gene expression profiles significantly correlated to GLS severity (Tables 4.5), were enriched for phenylpropanoid biosynthesis-related genes ("HS 10-10 S cor" in Table 4.6); similar to the genes with *trans*-eQTLs in "QTL 10-10 S cor" (Table S4.3). These results were not unexpected, since 84% of the *trans*-eQTLs in "QTL 10-10 S cor" were also in "HS 10-10 S cor"; and the *trans*-eQTLs in "QTL 9-5 S cor" and in "HS 9-6 S cor" overlapped with 95%.

Genes with *cis*-eQTL peaks in the *trans*-eQTL hotspot intervals could explain the expression of the target genes with *trans*-eQTLs at the hotspot loci. A total of 132 reporters (representing 125 gene models) with *cis*-eQTL peaks in the 5 QTL-overlapping *trans*-eQTL hotspots were identified (Table 4.5). Finally, 14 reporters (representing 14 gene models) and 9 reporters (representing 9 gene models) with the resistance and susceptibility associated alleles, respectively, linked with higher expression were identified to significantly correlate with the GLS severity scores (Figure 4.6 (b)). The middle section of Table 4.5 gives a breakdown per QTL-overlapping trans-eQTL hotspot, of the numbers of cis-eQTLs with peaks in the hotspot regions, after (i) it was divided into groups based on the parental allele associated with higher expression; and (ii) genes significantly correlated to GLS severity were identified (p-value < 0.01).

Reconstruction of gene regulatory networks

"Regulatory networks" for GLS resistance and susceptibility were constructed per QTL-overlapping *trans*-eQTL hotspot. Nodes represented reporters (transcripts) with eQTL peaks in the hotspot interval and directed edges connected genes with *cis*-eQTLs to highly co-expressed (correlated) genes with *trans*-eQTLs. Steps in the construction of the regulatory networks per GLS severity QTL (Figures 4.9, 4.10, 4.11, 4.12, 4.13, 4.14 and 4.15) are given below.

Genes with *cis*- and *trans*-eQTL peaks in the QTL-overlapping *trans*-eQTL hotspot regions were identified (see the sub-sections above). For each hotspot, the *cis*-eQTL genes for which a specific allele (e.g. CML444) was associated with increased expression

and the *trans*-eQTL target genes for which the same allele was associated with increased expression, were extracted for regulatory network reconstruction. This step assumed that when a specific allele was present (e.g. CML444), more mRNA transcripts of the regulator resulted in more mRNA transcripts of each of the target genes. Thus, cases where negative regulation have a positive effect, for example, will not be detected in this analysis.

Candidate genes were filtered so that only *cis*- and *trans*-eQTL genes that positively correlated (p-value <0.01) with either GLS resistance or susceptibility were included (the number of filtered eQTL genes per hotspot are given in Table 4.5). An additional filter was applied to ensure that "regulators" and "target genes" were co-expressed; the *cis*-eQTL genes were only linked to *trans*-eQTL genes when significantly correlated gene expression profiles (p-value <0.00001) were observed. As a result of various filtering steps, it is likely that some information or candidate genes will be lost. Predefined functional categories, distinguished by colour, were used to group eQTL genes involved in similar gene activities per regulatory network (Table 4.7).

QTL 4-11 regulatory network for genes associated with GLS susceptibility

Figure 4.9 and Table 4.8 give the QTL 4-11 regulatory network for genes positively associated with GLS susceptibility. Fifty-nine percent of the *trans*-eQTLs that overlapped QTL 4-11 S (201 eQTLs) had *trans*-eQTL peaks within HS 4-12 S (118 eQTLs). This relatively low percentage was mainly due to the hotspot locus that did not entirely overlap the GLS severity QTL region (Figure 4.8). Twenty percent of the *trans*-eQTLs with peaks in HS 4-12 S (24 out of 118) belonged to genes whose expression profiles were significantly correlated to GLS severity. The network was constructed from 24 genes with *trans*-eQTL peaks and four genes with *cis*-eQTL peaks in HS 4-12 S (Table 4.5).

The four putative regulatory genes with *cis*-eQTLs encoded: a nudix family hydrolase domain-containing protein (A_92_P038137; NX), a protein binding gene with no other annotation information (A_92_P036978; PB), a DHHC-type zinc finger family protein (A_92_P039457; ZF) and a tetratricopeptide repeat (TPR) family protein (A_92_P040694; TPR). The TPR- and nudix domain-containing proteins had edges to 79% and 53% of the *trans*-eQTL genes (the largest subsets), respectively. TPRs are protein-protein interaction modules involved in regulation of different cellular func-

tions. Proteins containing TPRs have been identified as essential determinants for signal transduction pathways mediated by most plant hormones including ABA, ET, cytokinin, gibberellin and auxin (Schapire *et al.*, 2006). Interestingly, none of the *trans*-eQTL genes that were connected to the TPR *cis*-eQTL gene were part of the "GLS susceptibility" sub-network (Figure 4.9). Nudix domain-containing proteins are known to hydrolyse nucleotide derivatives. Ge *et al.* (2007) identified a nudix domain-containing protein in *Arabidopsis* to be a negative regulator of the basal defense response to infection by *Pseudomonas syringae*. They identified that this protein negatively regulates two distinct defense response pathways, one independent of and the other dependent on NPR1 and SA accumulation. The DHHC-type zinc finger family protein is involved in regulation of transcription. The DHHC domain is a highly conserved cysteine-rich motif (Putilina *et al.*, 1999), which helps to anchor proteins to cell membranes.

ZF seemed to regulate five genes, that were all well-correlated to GLS susceptibility. The two *trans*-eQTL genes with the highest GLS susceptibility correlation coefficients in Table 4.8 with links to the "GLS susceptibility" node as well as to ZF, were an F-box/kelch-repeat protein SKIP11-like (A_92_P037621; as1) and a glucan endo-1,3-beta-glucosidase (A_92_P036877; as2). The F-box/kelch-repeat protein SKIP11 is a component of Skp1 Cullin F-box (SCF) E3 ubiquitin ligase complexes, which may mediate the ubiquitination and subsequent proteasomal degradation of target proteins (see description on page 96). This pathway is essential to many processes in plants including hormone signalling, flower development and stress responses. Angot *et al.* (2006) showed that the phytopathogenic bacterium *Ralstonia solanacearum* requires F-box-like domain-containing type III effectors to promote disease on several host plants. They proposed that these effectors may act by hijacking the host SCF-type E3 ubiquitin ligases in order to interfere with the host ubiquitin/proteasome pathway and as a result promote disease. It can be hypotehsised that *C. zeina* uses a similar strategy to manipulate the host ubiquitin/proteasome pathway not to act on and degrade fungal toxins, but rather target specific host substrates. β-glucosidases in plants play important roles in diverse aspects of plant physiology including plant defense, e.g. formation of intermediates in cell wall lignification, activation of phytohormones and activation of chemical defense compounds (Morant *et al.*, 2008).

Four of the *trans*-eQTL genes in Figure 4.9 were annotated with the GO-term "re-

sponse to chitin" or "response to fungus", accroding to Blast2GO. These genes, mostly regulated by both NX and TPR, were: a glycolipid transfer protein (A_92_P041126; as3); a serine hydrolase domain containing protein (A_92_P035799; as10); a serine threonine-protein kinase OXI1-like protein (A_92_P038869; as11); and an uncharacterised protein family (A_92_P030773; as12). An additional GO-term of as3 was "negative regulation of defense response", which aligned with what Ge *et al.* (2007) reported regarding its potential regulator, NX (see the description above). Since as3 (glycolipid transfer protein) had links to both the "GLS susceptibility" node and the NX *cis*-eQTL reporter, it can be hypothesised that the expression of these two genes (NX and as3) are manipulated by *C. zeina* to negatively regulate the basal defense response.

Three defense-related genes with *trans*-eQTLs that appeared to be regulated by TPR, likely a post-translational regulator, was respectively involved in hormone metabolism, reduction-oxidation (redox) and abiotic stress. These genes respectively encoded: a 2-oxoglutarate and Fe(II)-dependent oxygenase (A_92_P041146; as15), which is involved in ethylene formation (also linked to NX); a glutaredoxin (A_92_P029218; as7), which is a catalyst of deglutathionylation/glutathionylation reactions and also involved in stress response and iron sulfur assembly reactions (Rouhier *et al.*, 2008); and a universal stress domain containing protein (A_92_P029298; as17).

Therefore, the QTL 4-11 regulatory network associated with GLS susceptibility included a few defense-related genes that seemed to be activated in response to fungal infection. It could be that either (i) the plants activated the correct defense response genes too late after infection started; (ii) the plants activated genes involved in less effective strategies against *C. zeina*, or (iii) the fungus manpulated plant gene expression to activate genes to either negatively regulate the basal defense response (for example the NX regulator) or to interfere with host pathway functionality (for example the ubiquitin/proteasome pathway).

QTL 9-5 regulatory network for genes associated with GLS susceptibility

Figure 4.10 and Tables 4.9 and 4.10 give the QTL 9-5 regulatory network for genes positively associated with GLS susceptibility. More than 95% of the *trans*-eQTLs in "QTL 9-5 S" had *trans*-eQTL peaks within "HS 9-6 S" and 39% of the *trans*-eQTLs with peaks in HS 9-6 S (121 out of 311) belonged to genes with genes whose expression profiles

were significantly correlated to GLS severity (Table 4.5). The network was constructed from 121 genes with *trans*-eQTL peaks and three genes with *cis*-eQTL peaks in HS 9-6 S.

The *cis*-eQTL gene with the highest correlation to GLS susceptibility (0.4), had links to 88% of the *trans*-eQTLs in Figure 4.10 and encoded an EF hand or calmodulin-related calcium sensor protein (A_92_P037035; EF). The calmodulin (CaM) family is a major class of calcium sensor proteins, which collectively play a crucial role in cellular signalling cascades through the regulation of numerous target proteins (Ranty *et al.*, 2006). The second *cis*-eQTL gene had links to 61% of the *trans*-eQTLs in Figure 4.10 and encoded a currently uncharacterised protein (A_92_P028216; PUF). The last *cis*-eQTL gene had links to 46% of the *trans*-eQTLs in Figure 4.10 and encoded an enoyl-CoA hydratase (ECH) family protein (A_92_P022781; ECH). ECH proteins are known to catalyse a step in the beta-oxidation pathway of fatty acid metabolism (Agnihotri and Liu, 2003) and is unlikely a regulator. Interestingly, none of the *cis*-eQTL genes were linked with GLS susceptibility, but the "GLS susceptibility" node had links to 38% of the *trans*-eQTLs in the network. Furthermore, 90% of all the *trans*-eQTL genes with links to the "GLS susceptibility" node also had links to the EF hand *cis*-eQTL gene (Figure 4.10).

At least 11 biotic stress-related genes were part of this network (see Table 4.9). The *trans*-eQTL gene with the best gene expression correlation with GLS susceptibility in this category, was a beta-1,3-glucanase / glycosyl hydrolases family 17 (matching *Arabidopsis* ortholog At4G16260). Doxey *et al.* (2007) studied the functional divergence in the *Arabidopsis* beta-1,3-glucanase gene family. They found that At4G16260 (a root-specific beta-1,3-glucanase) displayed a significant expression response to four pathogens (*Alternaria brassicicola*, *Botrytis cinerea*, *Erysiphe orontii* and *Phytophthora infestans*) and that it was highly expressed following treatment with ET (a major hormonal regulator of pathogenesis-related (PR)-responses). Importantly, two ET biosynthesis-related genes, namely 1-aminocyclopropane-1-carboxylate (ACC) synthase (A_92_P039018; bs29) and 1-aminocyclopropane-1-carboxylate oxidase (A_92_P026516; bs73), were also part of the network (see Table 4.10). Generally, ET is thought to be an important factor for the induction of defense responses against pathogen attack. However, Ohtsubo *et al.* (1999) illustrated that ET promoted necrotic lesion formation in tobacco mosaic virus (TMV)-infected tobacco. They further showed that at least two kinds of basic PR protein genes,

the PR-1 and proteinase inhibitor II genes, were positively regulated by ET. Since ET biosynthesis-related genes were higher expressed in susceptible plants (with lesions) it could be hypothesised that ET was associated with necrotic lesion formation and with positively regulated proteinase inhibitors in the network (mentioned in the next paragraph). A third hormone metabolism gene in the network, was an ABA synthesis-related gene (A_92_P018101; bs103). According to Mauch-Mani and Mauch (2005) it appears that ABA, the abiotic stress hormone, affects disease resistance mainly negatively by interfering at different levels with biotic stress signalling primarily controlled by SA, JA and ET.

Other biotic stress-related genes included a Bowman-birk type trypsin inhibitor (A_92_P006679; bs13), a maize proteinase inhibitor (A_92_P034379; bs37), a cysteine proteinase inhibitor (A_92_P032100; bs62), a OTU-like cysteine protease (A_92_P030068; bs84), a pathogenesis-related maize seed protein (A_92_P023606; bs46), two chitinases (A_92_P022755; bs15 and A_92_P001063; bs44), two GRAM domain-containing proteins (A_92_P009951; bs53 and A_92_P022862; bs71) and a dirigent-like protein (A_92_P030266; bs57). Most pathogens secrete extracellular enzymes and enzymes causing proteolytic digestion of proteins, which play important roles in pathogenesis. Plants defend themselves through the synthesis of various inhibitors that act against these proteolytic enzymes. Cordero *et al.* (1994) showed that maize proteinase inhibitors were induced in response to wounding and fungal infection. Chilosi *et al.* (2000) identified wheat trypsin inhibitors (WTIs), belonging to the Bowman Birk-type protease inhibitor family, to have a strong antifungal activity against a number of pathogenic fungi and to inhibit fungal trypsin-like activity. Chitin is a structural component of the cell wall of many phytopathogenic fungi and plant chitinases are digestive enzymes that break down glycosidic bonds in chitin, thereby playing a key role in the plant defense response against fungal pathogens (Punja and Zhang, 1993). GRAM-domain containing proteins have diverse functions, but generally function at or near membranes of cells or organelles. Lorrain *et al.* (2004) identified a GRAM domain-containing protein that was expressed in response to pathogen infection, which is involved in cell death and defense responses in vascular tissues. Dirigent proteins (DIRs) are extracellular glycoproteins that are thought to play important roles in plant secondary metabolism (Pickel and Schaller, 2013). Shi *et al.* (2012) reported that a cotton DIR gene was identified to be involved in cotton lig-

nification, which can block the spread of the fungal pathogen *Verticillium dahliae*. The two abiotic stress-related genes in the network was two glutathione S-transferases (GSTs) (A_92_P022951; bs29 and A_92_P022861; bs98). GSTs are involved in the detoxification of endogenous and xenobiotic compounds (compounds that are foreign to the plant) and in plant secondary metabolism. Glutathione-dependent reactions are known to play an important role in plant stress responses (Marrs, 1996).

Twelve signalling genes were included in the QTL 9-5 regulatory network associated with GLS susceptibility (Table 4.9). Apart from the EF-hand *cis*-eQTL gene, three other calcium signalling genes (A_92_P021227; bs7, A_92_P006094; bs27 and A_92_P006123, bs54), three guanine nucleotide-binding proteins (G-proteins) (A_92_P030257; bs45, A_92_P022078; bs50 and A_92_P019150; bs97) and four receptor kinases (A_92_P009091; bs23, A_92_P039626; bs43, A_92_P029167; bs49 and A_92_P020152; bs93) were part of this network. Pathogen infection causes significant ion fluxes across membranes in plants and increasing evidence implicates calcium signalling in plant defense responses (Ranty *et al.*, 2006). G-proteins belong to the larger group of enzymes called GTPases. They function to transduce signals from a variety of different stimuli from outside a cell to the inside. G-proteins were identified to play an important regulatory role in multiple physiological processes, including the plant immune response (Zhang *et al.*, 2012*b*). Receptor-like kinases (RLKs) are known to play a central role in signalling during pathogen recognition as well as in the subsequent activation of plant defense mechanisms (Afzal *et al.*, 2008).

Six genes were responsible for the enriched GO-term "secondary metabolic process" in Table 4.6, of the *trans*-eQTLs in "HS 9-6 S cor". Two of these genes were phenylpropanoids involved in lignin biosynthesis, namely PAL (A_92_P031017; bs19) and dihydroflavonol-4-reductase (A_92_P013002; bs80); and two were flavonoids, namely flavanone 3-hydroxylase (A_92_P017679; bs30) and anthocyanin 5-aromatic acyltransferase (A_92_P020690; bs65). Lignification makes cell walls more resistant to the mechanical pressure applied during fungal penetration and flavonoids are known to possess antibacterial activity. Two cytochrome P450 genes were also part of the network (A_92_P041061; bs77 and A_92_P036137; bs90), playing critical roles in the biosynthesis of plant secondary metabolites. According to analyses of the functional and metabolic pathways of mulberry cytochrome P450 genes, Ma *et al.* (2013) reported that these genes

may participate in the metabolism of lipids, other secondary metabolites, xenobiotics, amino acids, cofactors, vitamins, terpenoids, and polyketides.

Plant cell walls are comprised largely of the polysaccharides cellulose, hemicellulose and pectin, along with 10% protein and up to 40% lignin (Tan *et al.*, 2013). Two hemicellulose synthesis genes (A_92_P017529; bs24 and A_92_P025721; bs70) and one pectin esterase (A_92_P029498; bs33) were part of the network, as well as a few other genes annotated with the GO-terms "cell-wall" and "external encapsulating structure": beta-glucanase (A_92_P007649; bs9 mentioned above), xylanase inhibitor (A_92_P021132; bs12), two chitinases, pathogenesis-related maize seed proteina, cysteine proteinase inhibitor (bs15, bs45, bs46 and bs62 mentioned above) and aconitate hydratase (A_92_P031077; bs100). Flatman *et al.* (2002) showed that a xylanase inhibitor (Xip-I) from wheat inhibits fungal xylanases (fungal enzymes that degrade hemicellulose in plant cells). Igawa *et al.* (2005) further showed that the wheat Xip-I gene was significantly induced by biotic and abiotic signals that trigger plant defense. Also, the expression of Xip-1 was significantly elevated by treatment with methyl jasmonate (MeJA).

Furthermore, seven genes were involved in protein degradation, seven in transport and a variety of genes in other categories are not mentioned in the text (see Tables 4.9 and 4.10).

The genes in the QTL 9-5 regulatory network associated with GLS susceptibility seemed to be involved in a variety of different processes. The EF-hand *cis*-eQTL gene, being a calmodulin-related calcium sensor protein, appeared to act as a global transcriptional regulator to activate numerous target proteins. This could be due to the ion fluxes across membranes in response to pathogen infection, since many of the genes in this network encoded pathogenesis-related proteins. However, being associated with GLS susceptibility, these genes were likely activated too late after infection started or involved in strategies not effective against *C. zeina*.

QTL 10-10 regulatory network for genes associated with GLS susceptibility

Figure 4.11 and Table 4.11 give the QTL 10-10 regulatory network for genes positively associated with GLS susceptibility. Sixty-three percent of the *trans*-eQTLs that overlapped QTL 10-10 S had *trans*-eQTL peaks within HS 10-10 S (148 out of 234 eQTLs). Fifty

percent of the *trans*-eQTLs in HS 10-10 S belonged to genes whose expression profiles were significantly correlated to GLS severity (74 out of 148; Table 4.5). The network was constructed from 74 genes with *trans*-eQTL peaks and two genes with *cis*-eQTL peaks in HS 10-10 S. Only 33 *trans*-eQTL were genes significantly correlated with at least one of the two *cis*-eQTL genes, but 20 additional *trans*-eQTL genes had links exclusively to the "GLS susceptibility" node (Figure 4.11).

One of the *cis*-eQTL genes encoded a glycosyltransferase protein (A_92_P034905; GT), which had links to 55% of the *trans*-eQTL genes in Figure 4.11. GTs catalyse the transfer of sugars to a wide range of acceptor molecules by the formation of glycosidic bonds (Glombitza *et al.*, 2004). Glycosylation serves to change the stability and solubility of molecules. In plants, GTs are generally involved in biosynthesis of secondary metabolites and in the detoxification of xenobiotics. The other *cis*-eQTL gene, encoding an AAA (ATPases Associated with various cellular Activities) family ATPase peroxin 6, had links to only 19% of the *trans*-eQTLs in Figure 4.11 (A_92_P025529; AAA). Peroxins are required for peroxisome biogenesis. Plant peroxisomes are organelles involved in numerous processes, including oxidation reactions, primary and secondary metabolism, development as well as responses to abiotic and biotic stresses (Hu *et al.*, 2012). AAA ATPase peroxin 6 plays a role in the import of proteins into peroxisomes and peroxisome biogenesis. These two genes are likely not regulators that play a major role in GLS susceptibility. In contrast to the two *cis*-eQTL genes, the "GLS susceptibiltiy" node had links to 75% of the *trans*-eQTLs in Figure 4.11.

Three other glycosyltransferases (A_92_P013322; cs15, A_92_P026596; cs36 and A_92_P038403; cs53) had links to the above-mentioned *cis*-eQTL gene, GT, together with a GST (A_92_P022951; cs17), a cytochrome P450 (A_92_P041061; cs50) and two ATP-binding cassette (ABC) transporters and multidrug resistance related proteins (A_92_P030884; cs20 and A_92_P019947; cs34) (Table 4.11). These gene families encode enzymes acting in plant xenobiotic metabolism and pathogen defense (Glombitza *et al.*, 2004). ABC transporters are characterised by the presence of specific transmembrane and signature adenosine triphosphate (ATP)-binding cassette domains (Martinoia *et al.*, 2002). Apart from their involvement in plant growth and developmental processes, ABC transporters also play a key role in detoxification of xenobiotic conjugates.

The enriched GO-terms for "HS 10-10 S cor" (Table 4.6) included "transferase activ-

ity" as well as "secondary metabolic process". More specific secondary metabolic process terms were "phenylpropanoid biosynthetic process" and "phenylpropanoid metabolic process". Among various other functions in plants, phenylpropanoid compounds play key roles in resistance to pathogen attack (Dixon *et al.*, 2002). Phenylpropanoid compounds that are involved in plant defense include lignin, coumarins and flavonoids (Naoumkina *et al.*, 2010). Three *trans*-eQTL genes in Figure 4.11 encoded proteins that were involved in phenylpropanoid biosynthesis. Flavanone 3-hydroxylase (A_92_P017679; cs29) is involved in flavonoid biosynthesis, phenylalanine ammonia-lyase (PAL) (A_92_P000336; cs48) in lignin biosynthesis and hydroxycinnamoyl-coenzyme A shikimate/quinate hydroxycinnamoyl transferase (A_92_P014030; cs11) has a putative role also in lignin biosynthesis (Hoffmann *et al.*, 2004). Flavonoids are known to act as antimicrobial compounds and lignification makes cell walls more resistant to the mechanical pressure applied during fungal penetration.

Four biotic stress-related genes were part of the network, encoding a Bowman-birk type trypsin inhibitor (A_92_P006679; cs9), a chitinase (A_92_P022755; cs10), a beta-1,3-glucanase (A_92_P007388; cs27) and a C2 domain-containing protein (A_92_P018920; cs44). The three genes that were firstly mentioned, had links exclusively to the "GLS susceptibility" node (not to the *cis*-eQTL genes). Interestingly, the reporters encoding the Bowman-birk type trypsin inhibitor and the chitinase were also mentioned above in the QTL 9-5 regulatory network associated with GLS susceptibility (thus both reporters had *trans*-eQTLs in HS 9-6 and in HS 10-10). These may be indications that that QTLs are interacting via *cis-trans* relationships, which should give rise to epistasis. Even though beta-1,3-glucanase was also mentioned above in the QTL 9-5 regulatory network, different reporters (and maize genes) encoded the beta-1,3-glucanases in HS 10-10 and HS 9-6. However, both reporters matched the *Arabidopsis* ortholog AT4G1626 mentioned by Doxey *et al.* (2007) (discussed in the previous sub-section) that displayed a significant expression response to four pathogens and that was highly expressed following treatment with ET. Plant C2 domain proteins play important roles in diverse cellular processes including development, growth and membrane targeting. However, they also play a part as in abiotic and biotic stress adaptations via the sensing of intracellular calcium signals (Zhang *et al.*, 2013). Zhang *et al.* (2013) demonstrated that a wheat C2 domain protein might be involved in wheat defense responses against stripe

rust and abiotic stresses in an ABA-dependent signalling pathway.

Two calcium signalling genes were part of the network, encoding a calmodulin-binding heat-shock protein (A_92_P021227; cs3) and a calcium lipid binding protein (A_92_P002935; cs47). Furthermore, regulation of transcription-related genes in the network (Figure 4.11) encoded two WRKY transcription factors (A_92_P018873; cs2 and A_92_P015550; cs28) and a NAC domain transcription factor (A_92_P013324; cs13). Three cell wall-related genes encoded a xylanase inhibitor (A_92_P021132; cs8 also with a *trans*-eQTL in HS 9-6), a secondary cell wall-related glycosyltransferase (A_92_P017529; cs23 also with a *trans*-eQTL in HS 9-6) and a hydroxyproline-rich glycoprotein (A_92_P022159; cs45). Xylanase inhibitors (as mentioned in the previous sub-section) and hydroxyproline-rich glycoproteins are known to be induced by biotic stress. Hydroxyproline-rich glycoproteins generally contribute to the structural support of plant cell walls. García-Muniz *et al.* (1998) showed that mRNA accumulation of a cell wall hydroxyproline-rich glycoprotein gene in maize was induced by fungal elicitors. Overall, 40% of the *trans*-eQTL genes in this QTL 10-10 regulatory network were also part of the QTL 9-5 regulatory network associated with GLS susceptibility. Therefore, 21 out of the 53 *trans*-eQTL genes in this network had a *trans*-eQTL in both HS 9-6 and in HS 10-10, providing evidence of an underlying gene expression network.

In conclusion, the genes in the QTL 10-10 regulatory network associated with GLS susceptibility, seemed to be involved in detoxification of xenobiotic compounds, phenylpropanoid biosynthesis and protein degradation, amongst other processes. Thus for a more severe infection, it is apparent that more detoxification and protein degradation enzymes were needed to act against the toxins produced by *C. zeina*. However, possibly due to other key defense strategies that were lacking, this attempt was not sufficient.

QTL 4-11 regulatory network for genes associated with GLS resistance

Figure 4.12 and Table 4.12 give the QTL 4-11 regulatory network for genes positively associated with GLS resistance. Fifty-seven percent of the *trans*-eQTLs that overlapped QTL 4-11 R had *trans*-eQTL peaks within HS 4-12 R (86 out of 150 eQTLs). Twenty-six percent of the *trans*-eQTLs in HS 4-12 R belonged to genes whose expression profiles significantly correlated to GLS severity (22 out of 86; Table 4.5). The network was constructed from 22 genes with *trans*-eQTL peaks and one gene with a *cis*-eQTL peak in HS

4-12 R. Only one *trans*-eQTL gene, encoding an erythronate-4-phosphate dehydrogenase family protein, had a link to the "GLS resistance" node (A_92_P006115; ar1 in Figure 4.12).

The only *cis*-eQTL gene in Figure 4.12 encoded a rossmann-fold NAD(P)-binding domain-containing protein (A_92_P008436; SDR), which had links to 19 *trans*-eQTL genes. The rossmann-fold is a common protein structural motif found in proteins that bind nucleotides, for example the cofactor nicotinamide adenine dinucleotide (NAD). Rossmann fold proteins make up one of three main classes of proteins that belong to the alpha/beta structure proteins. Short chain dehydrogenases/reductases (SDRs) constitute a large family of NAD(P)(H)-dependent oxidoreductases, which shares the rossmann-fold motif for nucleotide binding (Kavanagha *et al.*, 2008). According to blast2GO, this gene is located in the chloroplast inner membrane. It is unlikely that this gene is a regulator involved in GLS resistance.

Four *trans*-eQTL genes in the network were involved in regulation of transcription, encoding a bZIP transcription factor family protein (A_92_P008422; ar4), an unclassified RNA recognition motif-containing protein (A_92_P000526; ar10), a C3H zinc finger family protein (A_92_P015590; ar12) and a coiled-coil domain-containing protein (A_92_P007314; ar16). A few other *trans*-eQTL transcripts in the network encoded: a ranBP1 domain containing protein that is involved in G-protein signalling (A_92_P005700; ar3); a metal transport cation efflux family protein (A_92_P005524; ar5); a bifunctional polymyxin resistance ArnA protein which is involved in arabinoxylans (a major component of graminaceous plant cell walls) biosynthesis (A_92_P007232; ar7); a peroxiredoxin antioxidant PER1-like family protein (A_92_P004318; ar8); and an ASC1-like protein that is involved in sphingolipids lipid metabolism (A_92_P012338; ar15). Spassieva *et al.* (2002) studied the nectrotrophic fungus *Alternaria alternata* f.sp. *lycopersici*, which infects tomato plants by utilising a host-selective toxin (AAL-toxin) that kills the host cells by inducing programmed cell death. They identified ASC1 (similar to ar15) as a plant disease resistance gene that prevented disruption of sphingolipid metabolism during ALL-toxin-induced programmed cell death. It can be hypothesised that this gene (ar15) could likewise be involved in defense against *C. zeina* toxins. In general, this regulatory network do not include genes that appear to play a key role in GLS resistance.

QTL 9-5 regulatory network for genes associated with GLS resistance

Figure 4.13 and Table 4.13 give the QTL 9-5 regulatory network for genes positively associated with GLS resistance. More than 95% of the *trans*-eQTLs in "QTL 9-5 R" had *trans*-eQTL peaks within "HS 9-6 S" and 35% of the *trans*-eQTLs with peaks in HS 9-6 R (58 out of 166) belonged to genes with genes whose expression profiles were significantly correlated to GLS severity (Table 4.5). The network was constructed from 58 genes with *trans*-eQTL peaks and four genes with *cis*-eQTL peaks in HS 9-6 R.

The *cis*-eQTL transcript with the best correlation to GLS resistance had links to 46% of the *trans*-eQTLs in Figure 4.13 and encoded a serine threonine-protein kinase (A_92_P033066; STK). Protein kinases and phosphatases play a key role in signalling mechanisms critical for responses to environmental stresses and attack by pathogens (Sessa and Martin, 2000). In signal transduction pathways, protein kinases modify other proteins by phosphorylation and phosphatases dephosphorylate proteins (which are post-translational modifications). In particular, serine/threonine protein kinases phosphorylate the OH group of serine or threonine, which have similar side-chains. Activity of these protein kinases can be regulated by specific events, for example DNA damage, or chemical signals such as Ca^{2+}/calmodulin. At least five *trans*-eQTL genes in this network were also involved in signalling/post-translational modification, encoding a cysteine-rich receptor-like protein kinase (A_92_P012575; cr7), a receptor-like serine threonine kinase (A_92_P008364; br29), two serine/threonine protein phosphatases (A_92_P010477; br28 and A_92_P010609; br42) and a protein kinase family protein (A_92_P017199; br32). The first four genes that were mentioned had links to the STK *cis*-eQTL gene (as well as the NAGLU *cis*-eQTL gene; see description below). Based on their structural characteristics, receptor-like kinases function as cell surface receptors and play a crucial role in defense signalling.

The *cis*-eQTL transcript with the second best correlation to GLS resistance had links to 64% of the *trans*-eQTLs in Figure 4.13 and encoded an alpha-N-acetylglucosaminidase (NAGLU) (A_92_P009668; NAG). Ronceret *et al.* (2008) confirmed that NAGLU plays an important role in plant reproductive development. Therefore, this gene is not likely a regulator for defense. The other two *cis*-eQTL reporters did not have functional annotations and had links to only 19% and 17% of the *trans*-eQTL genes in Figure 4.13, respectively. According to the microarray re-annotation analysis (see Chapter 2; Coet-

zer *et al.*, 2011), A_92_P001417 did not match any transcript sequence in the maize WGS, but it matched a gDNA position at this locus on chromosome 9. Furthermore, A_92_P001249 matched a maize transcript that was not yet annotated.

Three *trans*-eQTL genes in this network had absolute GLS severity correlation coefficients that were higher than 0.5. These genes included a chlorophyll synthase that is involved in photosynthesis (A_92_P011013; br1), an uncharacterised protein (A_92_P011630; br2) and a callose synthase that is involved in carbohydrate metabolism (A_92_P010785; br3). Interestingly, six *trans*-eQTL genes encoding proteins involved in photosynthesis were included in the network (Figure 4.13). This callose synthase was also identified as one of the best candidates linked to GLS resistance in the previous Chapter (see section 3.4.3 on page 107). It is well established that the local deposition of callose is induced by abiotic stress and wounding (Jacobs *et al.*, 2003). Callose deposits are thought to act as a physical barrier to impede microbial penetration. Mauch-Mani and Mauch (2005) reported that ABA considerably enhances plant resistance to fungal pathogens through its positive effect on callose deposition.

ECERIFERUM1 (CER1) is the product of one of the *trans*-eQTL genes in this QTL 9-5 regulatory nework (A_92_P034039; br26) with a link to the *cis*-eQTL gene STK. This protein plays a role in cuticular wax production, but could be an important player in defense. The cuticle, a hydrophobic layer that covers plant aerial organs, serves as a waterproof barrier protecting plants against desiccation, ultraviolet radiation and pathogens. It mainly consists of cuticular waxes in which very-long-chain alkanes are the major components. Bourdenx *et al.* (2011) reported that overexpression of *Arabidopsis* CER1 promoted wax very-long-chain alkane biosynthesis and influenced plant response to biotic and abiotic stresses. However, they found that CER1-overexpression increased susceptibility to bacterial and fungal pathogens. Importantly, *C. zeina* does not directly penetrate through the wax layer (since it enters through stomata), but wax composition could effect its success of either binding to the leaf or germination.

A ranBP1 domain-containing protein involved in G-protein signalling (A_92_P005700; br22 also with a *trans*-eQTL in HS 4-12 mentioned above) was encoded by a reporter in the network. At least two transcripts were involved in regulation of transcription, encoding a bZIP transcription factor family protein (A_92_P006669; br10) and a CCAAT box binding factor family protein (A_92_P039278; br38). In addition, two RNA bind-

ing proteins (A_92_P008762; br27 and A_92_P011285; br30) and a RNA processing protein (A_92_P005627; br41) appeared to be regulated from this locus. Furthermore, genes encoding proteins involved in calcium transport (A_92_P005246; br11), ion and metabolite transport (A_92_P006606; b15), metal transport (A_92_P002406; br20) and potassium transport (A_92_P014444; br47), were also part of this network.

The QTL 9-5 regulatory network associated with GLS resistance seemed to include a group of genes involved in signalling/post-translational modification, including the *cis*-eQTL gene STK, which could act as a post-translational global regulator. A few single genes had potential links to biotic stress. Since resistant plant material harvested at this time point would have fewer lesions, the gene expression most likely expected is rather due to constitutive defenses that are still present at the late stage of sampling, than due to induced defenses in response to *C. zeina* (since the fungus is likely to have attempted penetration weeks earlier and have been stopped). Furthermore, fewer lesions also imply more photosynthetic material, which could explain why photosynthesis-related genes were detected to have more transcripts compared to samples with lesions and a strong (negative) correlation with GLS severity scores. Therefore, the photosynthesis-related genes with eQTLs in this network are potentially spurious eQTLs.

QTL 9-7 regulatory network for genes associated with GLS resistance

Figure 4.14 and Table 4.14 give the QTL 9-7 regulatory network for genes positively associated with GLS resistance. Fifty-seven percent of the *trans*-eQTLs that overlapped QTL 9-7 R had *trans*-eQTL peaks within HS 9-7 R (163 out of 93 eQTLs). Twenty-seven percent of the *trans*-eQTLs in HS 9-7 R (25 out of 93) were associated with genes whose expression profiles significantly correlated to GLS severity (Table 4.5). The network was constructed from 25 genes with *trans*-eQTL peaks and 4 genes with *cis*-eQTL peaks in HS 9-7 R.

The four *cis*-eQTL transcripts within this network encoded a 3-dehydroquinate synthase (A_92_P009686; DQS), with links to 72% of the *trans*-eQTL genes, which is an enzyme that belongs to the family of lyases and participates in the biosynthesis of aromatic amino acids (e.g. phenylalanine, tyrosine and tryptophan) as part of the shikimate pathway (see description below); a GTP-binding protein (A_92_P014787; GTP), with links to 24% of the *trans*-eQTL genes, which plays a key role in the signal transduction

pathways for numerous hormones; and two uncharacterised proteins (A_92_P001310; PUF1 and A_92_P011516; PUF2), with links to 60% and 28% of the *trans*-eQTL genes, respectively. The shikimate pathway is the basis of many secondary metabolite pathways in plants, which could potentially play important roles in plant defense against biotic and abiotic stresses as well as environmental interactions (Tohge *et al.*, 2013).

The three *trans*-eQTL transcripts in the network with the strongest correlating gene expression profiles to GLS resistance encoded a PLAC8 family protein (A_92_P009740; cr1), the only gene in the network with a link to the "GLS resistance" node (see description below); a peptidyl-prolyl *cis-trans* isomerases (PPIases), which catalyses and facilitates protein folding (A_92_P017490; cr2) (Breimans *et al.*, 1992); and a histone H1 protein, which is a DNA sequence-dependent determinant of chromatin structure and of transcriptional activity in chromatin (A_92_P020271; cr3) (Sera and Wolffe, 1998). PLAC8 (placenta-specific gene 8 protein) motif-containing proteins form a large family, which was originally found in the spongiotrophoblast layer of the placenta of mammals. Despite this protein family's wide distribution, knowledge about their function is very limited. Two very different functions were previously associated with PLAC8 motif-containing proteins, namely a role in (i) the determination of fruit and plant size and (ii) transport of heavy metals such as cadmium or zinc (Song *et al.*, 2011). This family also includes the plant cadmium (an important environmental pollutant) resistance (PCR) proteins of plants (Song *et al.*, 2004).

Six out of the 22 *trans*-eQTL genes in this QTL 9-7 regulatory network were also part of the QTL 9-5 regulatory network associated with GLS resistance. These genes encoded a PLAC8 family protein (mentioned above, cr1); a ATP synthase mitochondrial F1 complex assembly factor 1, which is essential for the assembly of the mitochondrial F1 complex (A_92_P007992; cr4); a protein of unknown function (A_92_P015277; cr9); a serine/threonine protein phosphatase, which is involved in post-translational regulation (A_92_P010477; cr10); a RNA recognition motif containing protein (A_92_P011285; cr11); and a AP2-domain DRE binding factor DBF1, which is an abiotic stress-related transcription factor (A_92_P005298; cr14). Dimosthenis and Montserrat (2002) reported that the maize dehydration responsive element (DRE)-binding proteins, DBF1 and DBF2, are involved in rab17 (an ABA-responsive gene of maize) regulation through the drought-responsive element in an ABA-dependent pathway. Chen *et al.* (2007) con-

cluded that the soybean (*Glycine max* L.) DRE-binding transcriptional activator may be useful in improving plant tolerance to abiotic stresses.

Another potentially interesting *trans*-eQTL gene encoded a Topless-related protein containing WD repeats (A_92_P003738; cr20). According to Zhu *et al.* (2010), *Arabidopsis* resistance protein SNC1 (encoding a TIR-NB-LRR-type R protein) activates immune responses through association with a transcriptional corepressor, Topless-related 1 (TPR1). Among the target genes of TPR1 are "Defense no Death" 1 and 2, two known negative regulators of immunity that are repressed during pathogen infection. Zhu *et al.* (2010) suggested that TPR1 activates R protein-mediated immune responses through repression of negative regulators. It can be hypothesised that in the current study, this gene could act as a repressor of negative regulators of the plant immune response, which were activated due to *C. zeina* manipulated plant gene expression (e.g the nudix domain-containing protein associated with GLS susceptibiltiy described in sub-section 4.4.4).

Four transport-related genes were present in the network encoding an iron ABC superfamily transporter (A_92_P019715; cr5), a ABC transporter family protein (A_92_P006825; cr8), a protein-export membrane protein (A_92_P009675; cr18) and an adaptor protein complex AP1 (A_92_P012143; cr22). Interestingly, Krattinger *et al.* (2009) showed that a putative ABC transporter conferred durable resistance to multiple fungal pathogens in wheat. Apart from a few single genes potentially of interest, this regulatory network do not include genes that appear to play a key role in GLS resistance.

QTL 10-10 regulatory network for genes associated with GLS resistance

Figure 4.15 and Table 4.15 give the QTL 10-10 regulatory network for genes positively associated with GLS resistance. Sixty-one percent of the *trans*-eQTLs that overlapped QTL 10-10 R had *trans*-eQTL peaks within HS 10-10 R (122 out of 201 eQTLs). Seventeen percent of the *trans*-eQTLs in HS 10-10 R (21 out of 122) were associated with genes whose expression profiles significantly correlated to GLS severity (Table 4.5). The network was constructed from 21 genes with *trans*-eQTL peaks and 3 genes with *cis*-eQTL peaks in HS 10-10 R.

Two of the *cis*-eQTL transcripts in this network had exceptionally strong correlations to GLS resistance; both had absolute correlation coefficiets of 0.5 to GLS severity and had links to the "GLS resistance" node. Therefore, these genes could be components of an

underlying transcriptional network involved in GLS resistance. Interestingly, out of all the regulatory networks presented above, these were the only two genes with *cis*-eQTLs with a highly significant gene expression correlation to GLS severity (p-value <0.00001). These genes encoded a leucoanthocyanidin reductase (A_92_P009023; LAR) with links to 64% of the *trans*-eQTL genes in the network and a glutamyl-tRNA reductase (A_92_P006618; GTR) with links to 100% of the *trans*-eQTL genes in the network. LAR is an enzyme that participates in flavonoid biosynthesis and flavonoids may provide antioxidant activity as part of a general stress response (Winkel-Shirley, 2002). GTR catalyses the first step of tetrapyrrole biosynthesis in plants, which is involved in chlorophyll and heme (among other products) production. The third *cis*-eQTL gene encoded an uncharacterised protein (A_92_P001146; PUF) with links to 36% of the *trans*-eQTL genes. Not one of these *cis*-eQTL genes are likely global regulators.

Three *trans*-eQTL genes with links to the "GLS resistance" node encoded a lipid phosphate phosphatase (A_92_P017985; dr1), a cysteine-rich receptor-like protein kinase (A_92_P012575; dr2) and a RNA processing splicing factor (A_92_P018634; dr3). Interestingly, the two genes that were firstly mentioned were two of the three genes in this network that also had *trans*-eQTLs in HS 9-6 (Table 4.14).

Two transport-related *trans*-eQTL genes were present in the network encoded: an ABC transporter family protein (A_92_P025469; er4) and a a multidrug resistance (MRP)-type ATP binding protein (A_92_P017437; dr11). The MRP subfamily of plant ABC transporters are suggested to play a role in cellular detoxification by vacuolar sequestration of endogenous or exogenous toxic compounds. Stukkens *et al.* (2005) showed that NpPDR1, a pleiotropic drug resistance-type ABC transporter (another subfamily of plant ABC transporters) from tobacco (*Nicotiana plumbaginifolia*), plays a major role in pathogen resistance due to its involvement in both constitutive and JA-dependent induced defense. It can be hypothesised that the the ABC transporter family, via hormone signalling, play a part in GLS resistance.

Furthermore, two F-box proteins (A_92_P008221; dr9 and A_92_P015985; dr13) were also included in this network, which mediate ubiquitination and subsequent protein degradation. The COI1 F-box is an example of a component of the ubiquitination system that was shown to play a role in plant immunity (Xie *et al.*, 1998). COI1 controls defence pathways that are regulated by JAs, which are synthesised in response to

pathogen attack. Thomma *et al.* (1998) showed that *coi1* mutants, which are unable to relay the JA-signal, are more susceptible to necrotrophic pathogens. van den Burg *et al.* (2008) provided another example where F-box proteins were involved in defense. They reported that the F-Box protein ACRE189/ACIF1 (with an F-box domain that interacts with SKP1/CUL1/ F-box (SCF) subunits) regulated cell death and defense responses activated during pathogen recognition in tobacco and tomato. It could therefore be hypothesised that the F-box proteins in this network also play a role in disease resistance via plant hormone signalling.

Also, a gene encoding a cytochrome P450 (A_92_P039824; er10) were included in the network. Cytochrome P450s are known to catalyse most of the oxidation steps in plant secondary metabolism. Compounds metabolised by P450 enzymes can act as stress signals in plant defense or exert a direct antifungal activity. Li *et al.* (2010) showed that resistance to Fusarium head blight and seedling blight in wheat is associated with activation of a cytochrome P450 gene, where cytochrome P450 plays an important role in protecting plants against trichothecene mycotoxins. A P450 gene in potato was also identified as a molecular marker of resistance to fungal pathogen *Phytophthora infestans* (Trognitz *et al.*, 2002).

A few genes with potential associations to defense were part of the QTL 10-10 regulatory network associated with GLS resistance. It can be proposed that increased antifungal activity, due to higher levels of flavonoids and cytochrome P450s in resistant plants, could protect plants against toxins from *C. zeina.* Futhermore, the ABC transporter family or F-box proteins together with the ubiquitin system could play a role in disease resistance via plant hormone signalling.

4.5 Conclusion

The genetic basis for the response to *C. zeina* infection in the CML444×SC Malawi maize RIL population was studied and candidate genes and pathways associated with GLS resistance or susceptibility were identified. Various filtering steps were used to narrow down the list of potential candidates and the final output was hypotheses regarding genes and mechanisms that could explain the GLS severity QTLs. The analysis was based on the hypothesis that there is an underlying DNA polymorphism that gives rise

to a change in gene expression which in turn affects the phenotypic trait. Genetic control is not directly observed, rather genetic response to a purtubation, in this case a DNA polymorphism. Figure 4.16 is a Circos diagram (Krzywinski *et al.*, 2009) that displays the candidate genes with *trans*-eQTLs potentially involved in mechanisms associated with GLS resistance or susceptibility. These genes had expression profiles significantly correlating (positively or negatively) to GLS severity and their expression were (partially) explained by *trans*-eQTLs that were part of *trans*-eQTL hotspots coinciding with GLS severity QTLs.

Many of the genes in the proposed regulatory networks associated with GLS susceptibility (Figures 4.9, 4.10 and 4.11) had functional annotations that disclosed potential defense-related mechanisms. Since leaves were sampled during flowering when GLS lesions were evident, it was expected that cells around the lesions would be fighting the fungus and consequently that genes with a higher expression in susceptible plants would be activated in response to pathogen infection and damage to leaf cells. Susceptibility could be due to the plants either activating the response genes too late after infection started or activating genes involved in less effective strategies against *C. zeina*. Alternatively, it could be due to fungal manipulation of plant gene expression, for example activating genes that negatively regulate the basal defense response. Surprisingly, 21 genes with *trans*-eQTLs were shared between the QTL 9-5 and the QTL 10-10 regulatory networks for GLS susceptibility. This may indicate that these two QTLs are interacting, which could give rise to epistasis. Furthermore, these genes could be components of an underlying transcriptional network regulating the response to GLS disease. A few hypotheses regarding potential genes and pathways playing a role in GLS susceptibility are given below.

The EF-hand gene with a *cis*-eQTL in HS 9-6 S (a calmodulin-related calcium sensor protein) appears to act as a global regulator that activate numerous target proteins through calcium signalling, due to the ion fluxes across membranes in response to pathogen infection (Ranty *et al.*, 2006). Since 30% of genes displayed in Figure 4.16 had *trans*-eQTLs in HS 9-6 S, this locus appeared to play a significant role in GLS susceptibility. The genes with *trans*-eQTLs in HS 9-6 S seemed to be involved in a variety of different processes and included a strong signature of pathogenesis-related genes, genes involved in signalling, and secondary metabolism-related genes. It is hypothesised that a

PR beta-1,3-glucanase (with a *trans*-eQTL in HS 9-6 S) is activated too late after infection or that its activity was not effective against *C. zeina*. Due to two ET biosynthesis-related genes and three proteinase inhibitors (with *trans*-eQTLs in HS 9-6 S), it is hypothesised that ET promotes necrotic lesion formation, which was observed in susceptible plants, and simultaneously positively regulates proteinase inhibitors (similar to a result obtained from Ohtsubo *et al.* (1999) on TMV-infected tobacco).

Due to a gene encoding an F-box/kelch-repeat protein SKIP11 (with a *trans*-eQTL in HS 4-12 S), it is hypothesised that *C. zeina* could require F-box-like domain-containing type III effectors to promote disease and as a result manipulate the host ubiquitin/proteasome pathway. It is also hypothesised that the expression of two genes encoding a nudix family hydrolase domain-containing protein (with a *cis*-eQTL in HS 4-12 S) and a glycolipid transfer protein (with a *trans*-eQTL in HS 4-12 S) were manipulated by *C. zeina* to negatively regulate the basal defense response and promote disease.

The genes in the QTL 10-10 regulatory network associated with GLS susceptibility, seemed to be involved in detoxification of xenobiotic compounds, phenylpropanoid biosynthesis and protein degradation. It was hypothesised that the plant's attempt to detoxify and degrade toxins produced by *C. zeina* was not sufficient, probably due to other key defense strategies that were lacking. It was further hypothesised that these processes are explained by the presence of a gene (or a few tightly linked genes) within the HS 10-10 locus that were not detected by this study for two reasons: (i) the two genes with *cis*-eQTLs in the network were not likely regulators of the mentioned processed and (ii) the 20 out of 53 genes had expression values that highly correlated with GLS susceptibility, did not have links to genes with *cis*-eQTLs. It could be that the additional regulator(s) was not present on the microarray, its *cis*-eQTL effect was too small to be detected, its *cis*-eQTL peak was just outside the hotspot region, or its gene expression profile did not correlate well with GLS severity.

The gene expression in resistant plants with fewer lesions is most likely due to constitutive defense mechanisms that are still present at the late stage of sampling, rather than due to induced defenses in response to *C. zeina*, since the fungus has likely attempted penetration weeks earlier and has been stopped in resistant plants. Noteworthy, six genes with *trans*-eQTLs were shared between the QTL 9-5 and the QTL 9-7 regulatory networks for GLS resistance, indicating potential interaction between these QTLs. A few

hypotheses regarding potential genes and pathways playing a role in GLS resistance are given below.

A serine threonine-protein kinase (with a *cis*-eQTL in HS 9-6 R) is hypothesised to act as a global regulator through post-translational modification. Due to at least five *trans*-eQTL genes (with *trans*-eQTLs in HS 9-6 R) that were also involved in signalling/post-translational modification, it is hypothesised that the phosphatases/kinases in the QTL 9-5 regulatory network associated with GLS resistance could be involved in post-translational modifications, which could modulate the expression of genes responding to fungal elicitors and play an important role in defense. It is further hypothesised that the gene encoding a callose synthase (with a *trans*-eQTL in HS 9-6 R), plays a role in GLS resistance via deposition of callose in the form of local cell wall thickenings to block fungal penetration. Furthermore, the gene encoding CER1 (with a *trans*-eQTL in HS 9-6 R), which is involved in cuticular wax production, is hypothesised to play a role in GLS resistance, since wax composition could affect *C. zeina*'s success of binding to the leaf or germination.

It is hypothesised that the gene encoding an ASC1-like protein in HS 4-12 R, act as a plant disease resistance gene by preventing disruption of sphingolipid metabolism during *C. zeina* toxin-induced programmed cell death (similar to what was reported by Spassieva *et al.*, 2002). A Topless-related protein containing WD repeats (with a *trans*-eQTL in HS 9-7 R) is hypothesised to act as a repressor of negative regulators of the plant immune response, which could have been activated due to *C. zeina* manipulation of plant gene expression (similar to what was reported by Zhu *et al.*, 2010). Furthermore, it is hypothesised that members of the ABC transporter family (two genes with *trans*-eQTLs in HS 10-10 R) could play a role in disease resistance via plant hormone signalling, as well as that the F-box proteins together with the plant ubiquitin system (two additional genes with *trans*-eQTLs in HS 10-10 R) could be involved in GLS resistance, through hormone signalling.

The overlap of eQTLs with GLS severity QTLs, including additional filtering steps, were used to identify candidate genes and pathways that could be involved in the disease response of maize to GLS. The findings in this chapter were mainly hypotheses, which need to be validated with further studies, for example via the generation of overexpression or knockout lines. However, this chapter indicates that that there is a genetic basis

for the response to *C. zeina* infection in the CML444×SC Malawi maize RIL population. It further reveals the complex genetic architecture of transcript level variation in maize and confirms that determining the molecular mechanisms underlying complex phenotypic traits generally remains a bottleneck. Once the genetic basis (the causal genes and/or polymorphisms) is determined, it can be put to practical use in crop improvement. Chapter 5 will incorporate gene co-expression networks (from Chapter 3) with the QTL/eQTL analyses (from Chapter 4) in a systems genetics context to determine whether there is a genetic basis for the coordinated expression responses to GLS disease.

4.6 Acknowledgement of data contributions

I would like to acknowledge the following people for contributing to make the analysis of this chapter possible:

- Ms. Jeanne Korsman for constructing the linkage map.
- A bioinformaticist (who preferred to stay anonymous) employed in the Maize eQTL project for developing the first version of a customised computer script for QTL and eQTL identification.

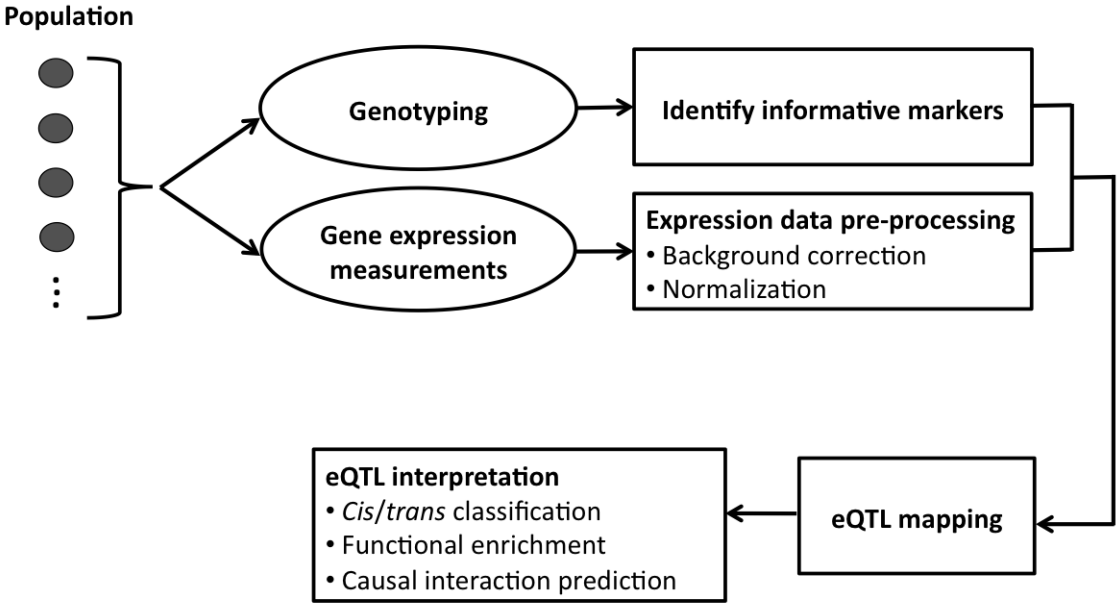


Figure 4.1: The steps in an eQTL study. Adapted from Michaelson *et al.* (2009). A set of individuals in a population is genotyped (using for example SNP arrays) and markers that are polymorphic in the study population are selected for the QTL analysis. Gene expression, in the same individuals, is measured (using microarrays or RNA sequencing) and the expression data are pre-processed using standard procedures. eQTL mapping consists of selecting markers that explain the expression variation in the population. Significant eQTLs are interpreted, for example by checking groups of genes with eQTLs at a common locus for functional enrichment or by using network analysis for inferring causal relationships.

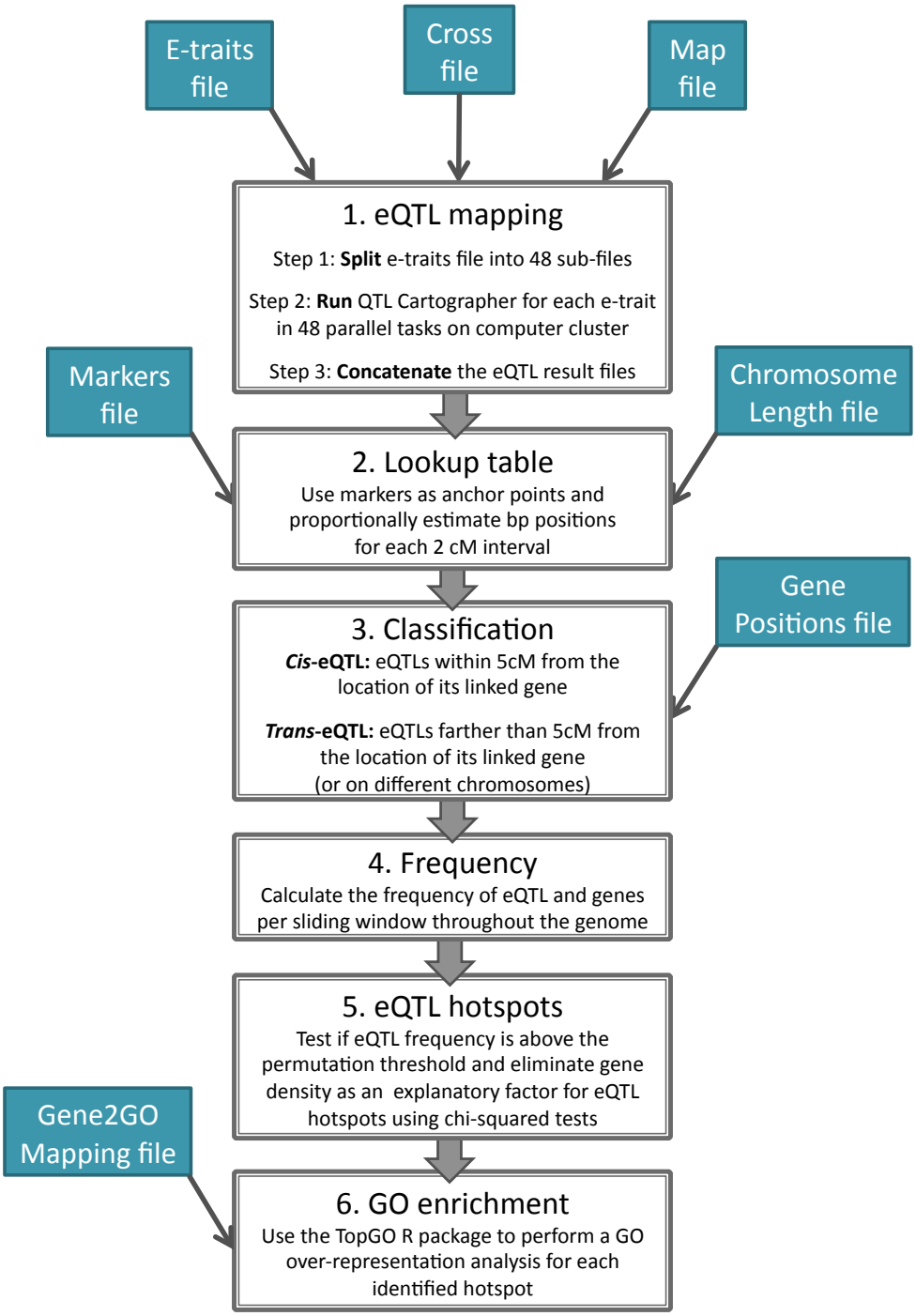


Figure 4.2: eQTL data analysis pipeline implemented in Galaxy. The pipeline consists of six modules and seven input files are required (shown as blue blocks). The first module runs QTL Cartographer for each expression trait (e-trait) as 48 parallel tasks using a computer cluster in order to map eQTLs. The second module links the genetic and physical maps, where markers are used as anchor points to proportionally estimate bp positions for each 2 cM interval. The third module classifies eQTLs as *cis* or *trans*. The fourth module calculates the frequency of eQTL and genes per sliding window throughout the genome. The fifth module identifies significant unbiased eQTL hotspots. The last module performs a GO over-representation analysis on each identified hotspot using the TopGO R package.

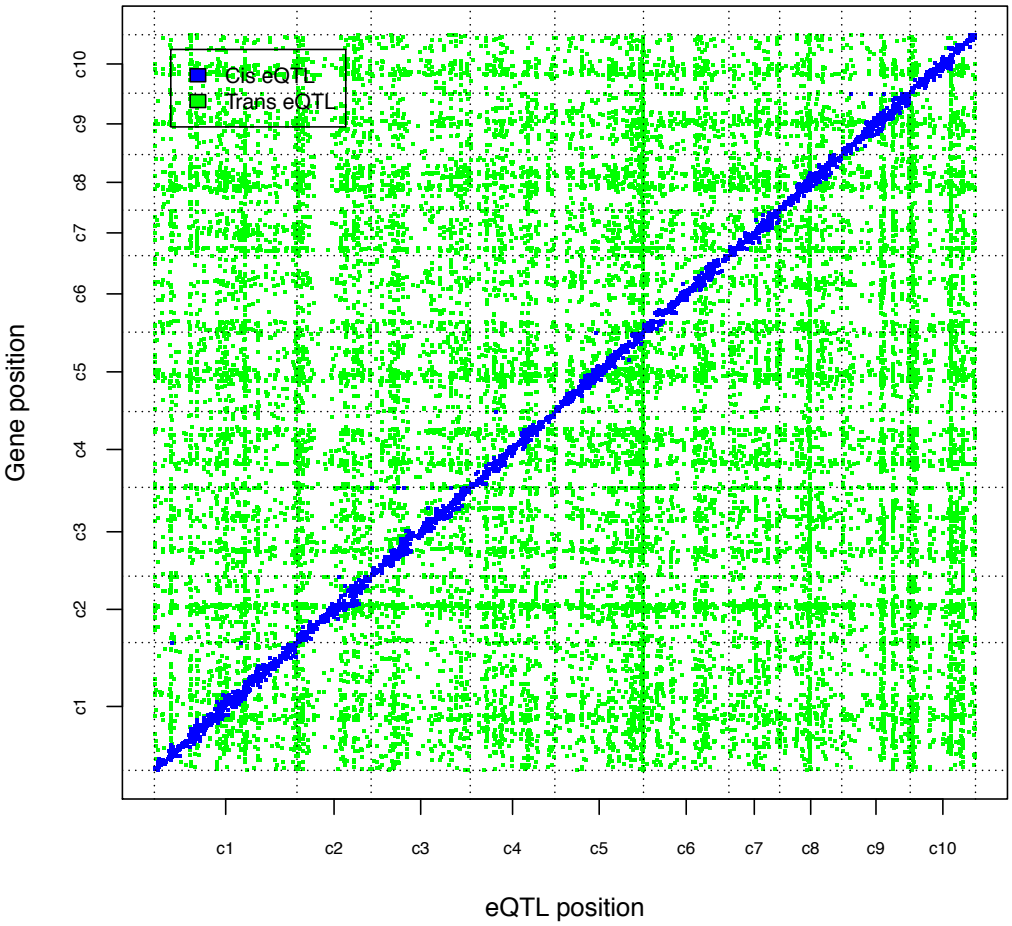


Figure 4.3: A scatter plot giving the genomic relationships between eQTL positions (x-axis) and the corresponding e-trait gene positions (y-axis) across the maize genome. The figure was generated by the "classification" module of the eQTL data analysis pipeline in Galaxy. The eQTL and e-trait positions corresponded to the 1009 bins (from the lookup table) across the genome. Each bin is linked to a cM and bp position. The color-key distinguishes between *cis*- (blue) and *trans*-eQTLs (green). The ten maize chromosomes are separated by grey dashed lines and tick-marks indicate the middle of each chromosome.

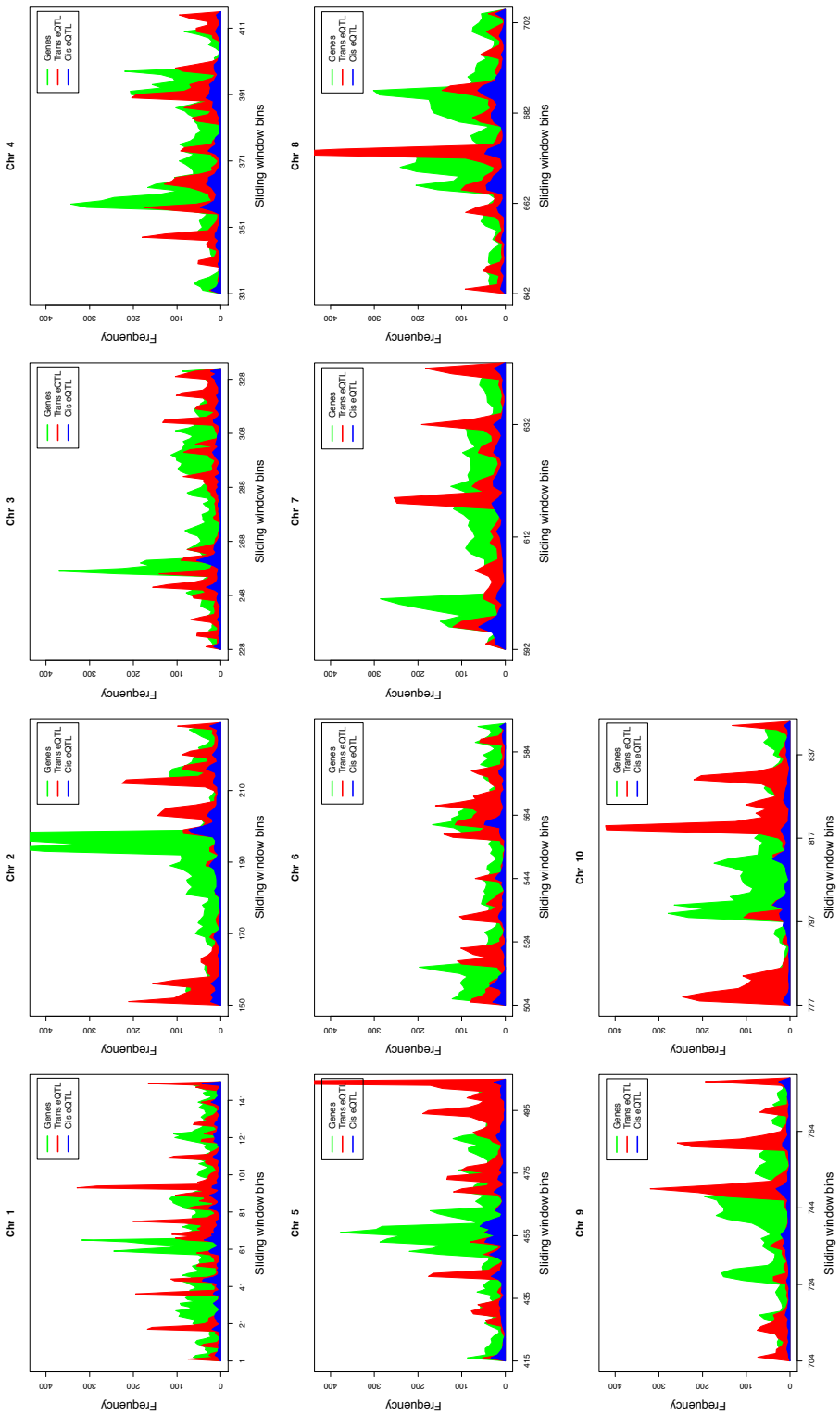


Figure 4.4: Frequency distribution of genes, *cis*- and *trans*-eQTLs across the maize genome. This figure was generated by the "frequency" module of the eQTL data analysis pipeline in Galaxy. The numbers of genes, *cis*- and *trans*-eQTLs (y-axis) were plotted against the genomic locations per chromosome (sliding window bins on the x-axis). The color-key distinguishes between gene density (green), *cis*- (blue) and *trans*-eQTLs (green).

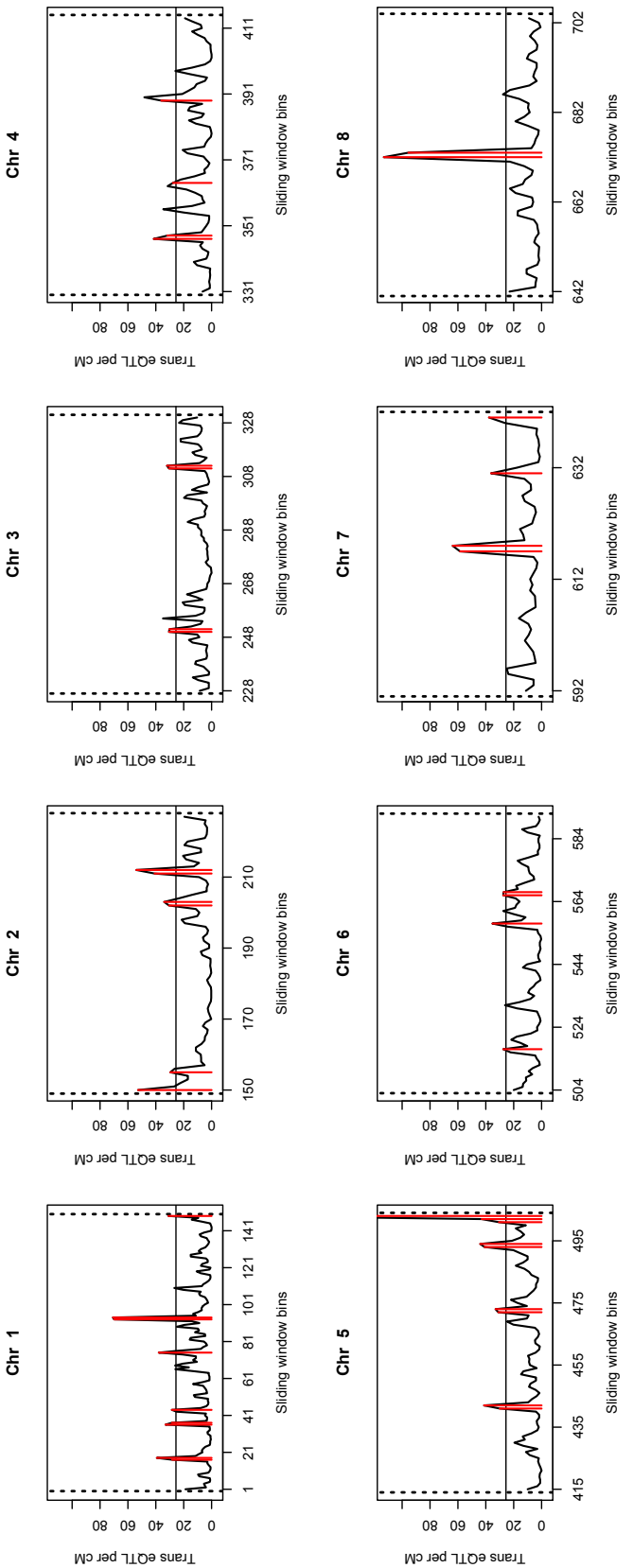


Figure 4.5: Distribution of the *trans*-eQTLs per cM across the 10 maize chromosomes. This figure was generated by the "hotspots" module of the eQTL data analysis pipeline in Galaxy. The y-axis gives the number of *trans*-eQTLs per cM that was calculated per sliding window bin (x-axis). The horizontal line shows the permuted threshold (p-value < 0.05) for detection of *trans*-eQTL hotspots. Sliding window bins where the number of eQTLs per cM was above the permutation threshold and with significant eQTL excess compared to gene number (chi-squared test p-value< 0.0001), were declared as "unbiased" eQTL hotspots and marked in red.

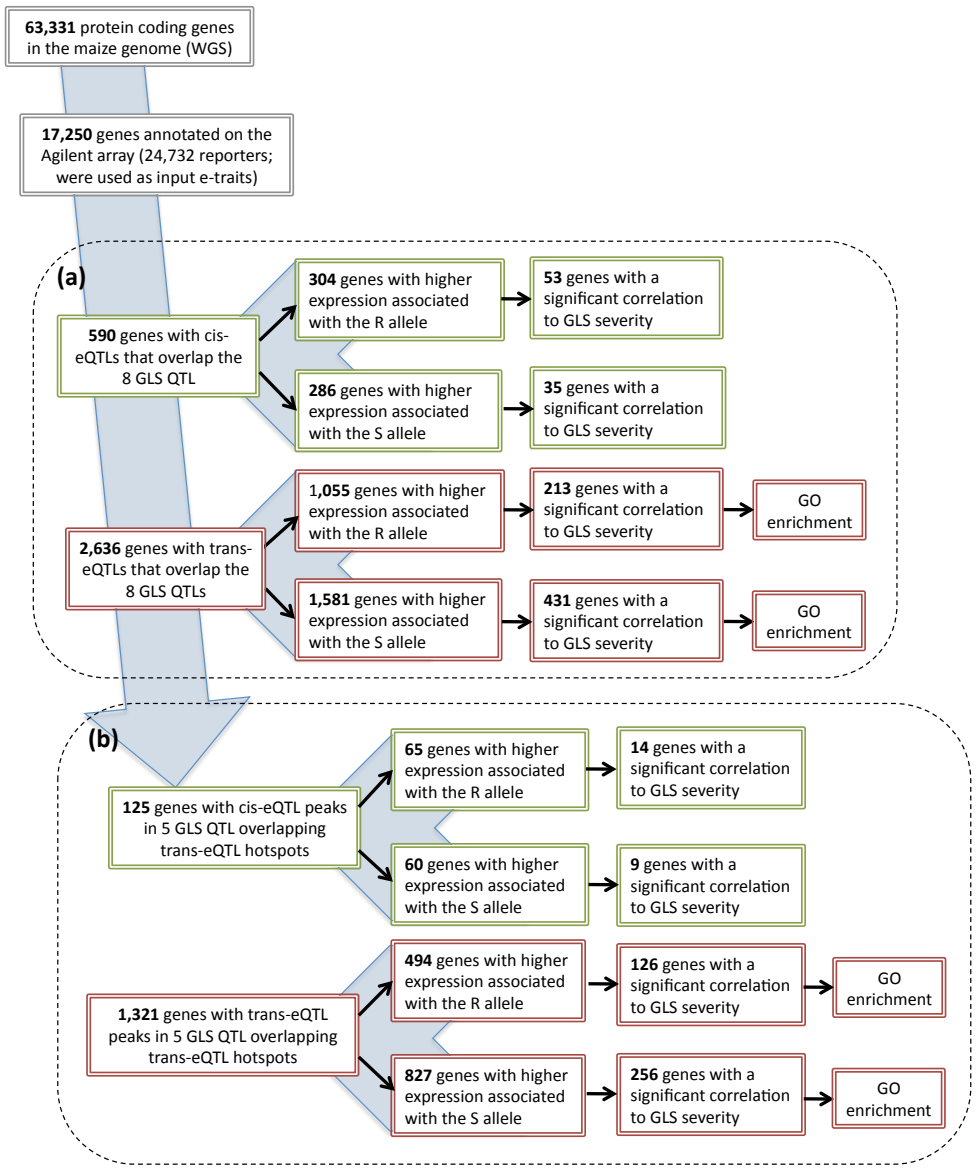


Figure 4.6: Flow diagram of the QTL/eQTL overlap strategy to identify the genes and processes associated with GLS resistance or susceptibility. Table 4.1 states whether CML444 or SC Malawi was the parent with the resistance associated allele for each GLS severity QTL. "R" and "S" in the diagram refer to the resistance (R) or susceptibility (S) associated allele. **(a)** Genes with *cis*- (green) and *trans*-eQTLs (red) that overlapped the GLS severity QTLs were identified and divided into groups based on parent associated with higher expression. Subsequently, genes with expression profiles that significantly correlated with the phenotype values (GLS severity scores) across the RILs were identified (p-value < 0.01). **(b)** Genes with *cis*- (green) and *trans*-eQTL (red) peaks within the five previously identified *trans*-eQTL hotspot intervals that overlapped the GLS severity QTLs, were identified and divided into groups based on parent associated with higher expression. Subsequently, genes with expression profiles that significantly correlated with the GLS severity scores were identified (p-value < 0.01). GO enrichment analyses were performed on the resulting sets of genes with *trans*-eQTL to reveal biological processes potentially associated with GLS disease.

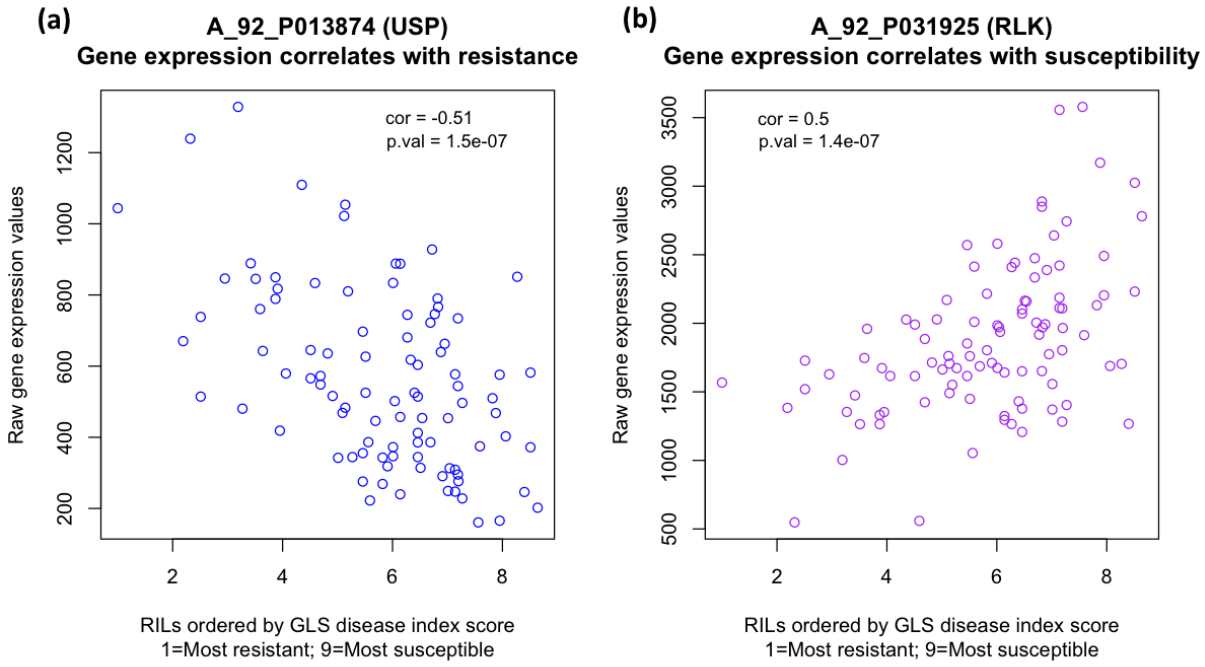


Figure 4.7: Scatter plots illustrating a negative and a positive gene expression correlation with GLS disease severity scores. **(a)** Scatter plot of the gene expression values across the RILs, for a gene encoding an ubiquitin-specific protease (USP) with a *cis*-eQTL that overlapped GLS severity QTL 10-10 (Table S4.2 in the electronic Appendix). The negative correlation indicates that RILs for which this gene have high gene expression values, have low disease severity scores (i.e. resistant RILs); whereas RILs for which this gene have low gene expression values, have high disease severity scores (i.e. susceptible RILs). Therefore, this gene's expression values correlate with GLS resistance. **(b)** Scatter plot of the gene expression values across the RILs, of a receptor-like cytosolic serine threonine-protein kinase (RLK) with a *cis*-eQTL that overlapped GLS severity QTL 9-5 (Table S4.1 in the electronic Appendix). The positive correlation indicates that RILs for which this gene have high gene expression values, also have high disease severity scores (i.e. susceptible RILs); and RILs for which this gene have low gene expression values, also have low disease severity scores (i.e. resistant RILs). Therefore, this gene's expression values correlate with GLS susceptibility.

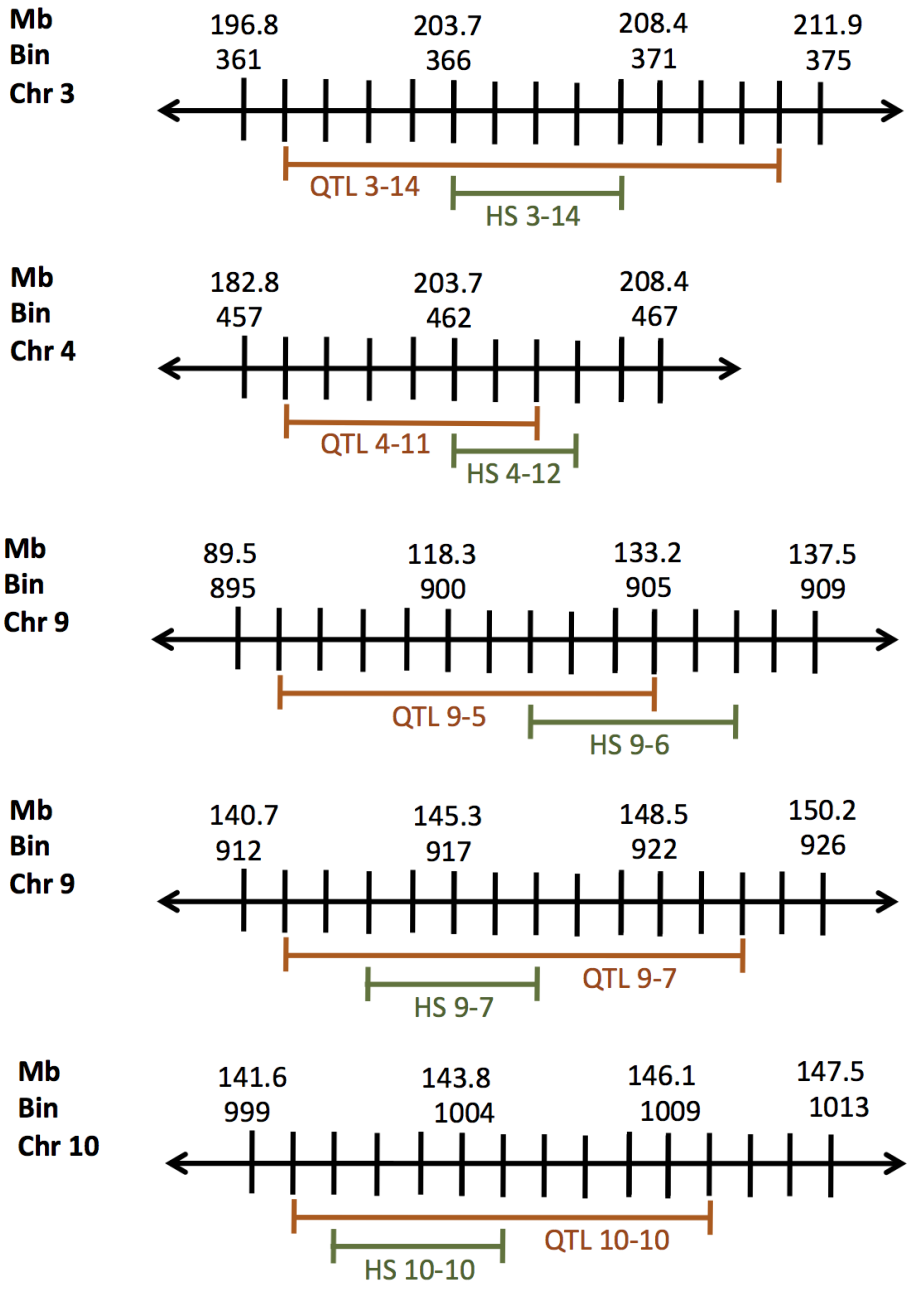


Figure 4.8: *Trans*-eQTL hotspot regions (green) coinciding with GLS severity QTLs (brown). Summary statistics per *trans*-eQTL hotspot that overlapped a GLS severity QTL are given in Table 4.5. The Mb position, 2 cM bin number (see lookup table in electronic Appendix) and chromosome number are given for each of the five GLS severity QTL overlapping *trans*-eQTL hotspot regions. QTLs and *trans*-eQTL hotspots were named based on the chromosome and the number of the start marker on QMap 2.0.

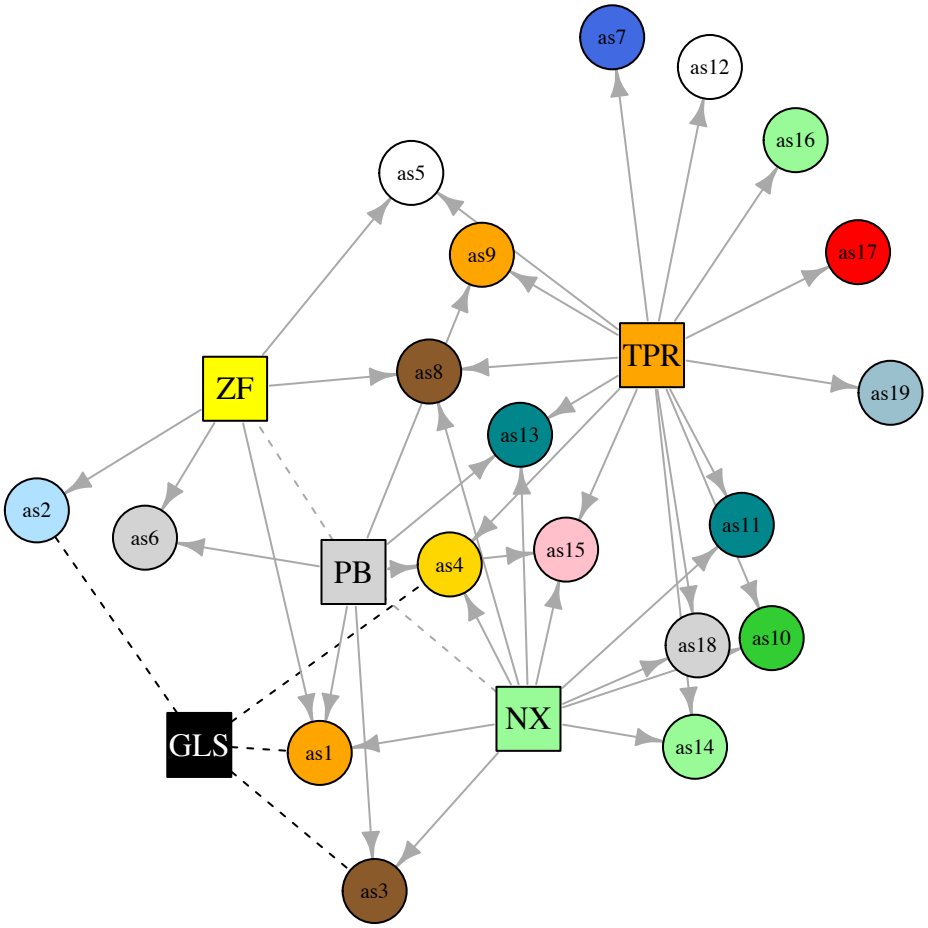


Figure 4.9: QTL 4-11 regulatory network model for genes associated with GLS susceptibility. The black square represents the phenotypic trait "GLS susceptibility". Black dotted lines indicate a strong correlation (p-value <0.00001) between the GLS severity profile and the expression profiles of genes with eQTLs in HS 4-12 S. Coloured square nodes represent *cis*-eQTL genes (ZF=DHHC-type zinc finger family protein; NX=NUDIX family domain containing protein; TPR=TPR repeat region family protein; PB=protein-binding protein) and round nodes *trans*-eQTL genes. The *trans*-eQTL nodes are sequentially numbered from "as1" to "as19" where "as1" represents the *trans*-eQTL gene with the strongest correlation to GLS susceptibility (see Table 4.8 for node annotations). Node colours correspond to the categories in Table 4.7. Solid edges represent gene co-expression between *cis*- and *trans*-eQTL genes (p-value <0.00001); and arrows indicate direction of regulation, assuming that *cis* variation explains gene expression differences for genes in *trans*. Grey dotted lines indicate a strong correlation (p-value <0.00001) between the gene expression profiles of pairs of *cis*-eQTL genes.

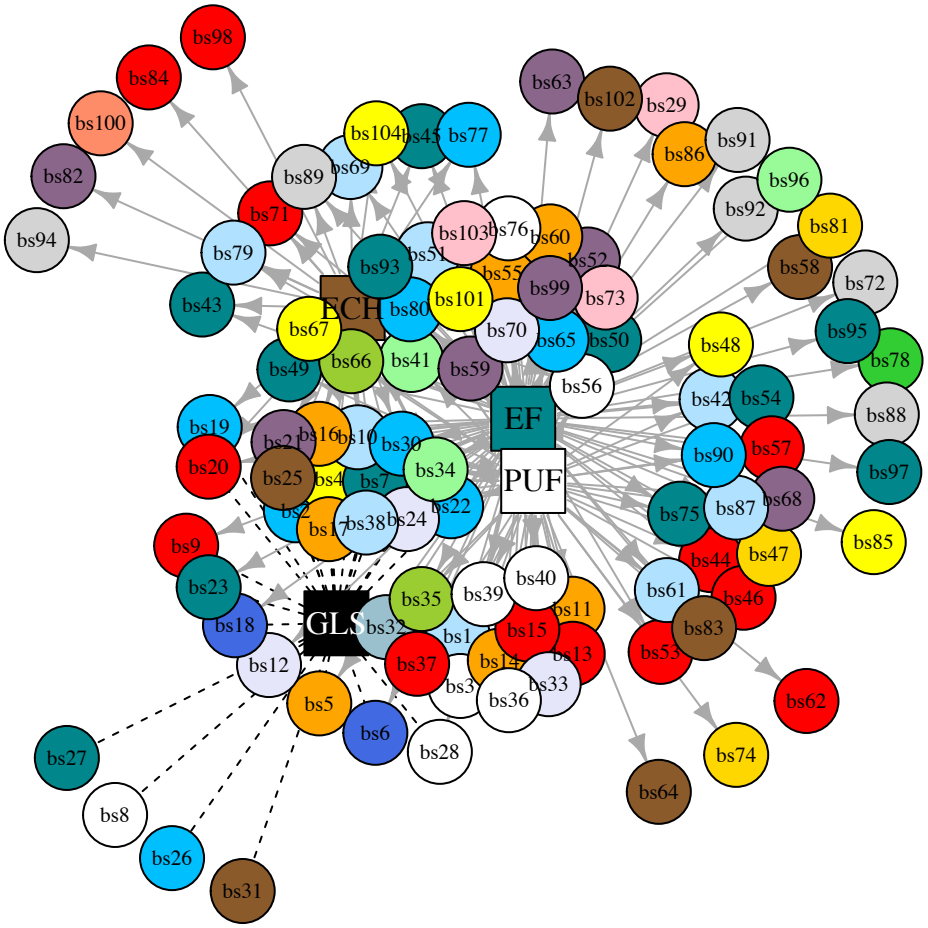


Figure 4.10: QTL 9-5 regulatory network model for genes associated with GLS susceptibility. The black square represents the phenotypic trait "GLS susceptibility". Black dotted lines indicate a strong correlation (p-value <0.00001) between the GLS severity profile and the expression profiles of genes with eQTLs in HS 4-12 S. Coloured square nodes represent *cis*-eQTL genes (EF=EF hand calmodulin-related protein; ECH=Enoyl-CoA hydratase family protein; PUF=protein of unknonw function) and round nodes *trans*-eQTL genes. The *trans*-eQTL nodes are sequentially numbered from "bs1" to "bs104" where "bs1" represents the *trans*-eQTL gene with the strongest correlation to GLS susceptibility (see Tables 4.9 and 4.10 for node annotations). Node colours correspond to the categories in Table 4.7. Solid edges represent gene co-expression between *cis*- and *trans*-eQTL genes (p-value <0.00001); and arrows indicate direction of regulation, assuming that *cis* variation explains gene expression differences for genes in *trans*. Grey dotted lines indicate a strong correlation (p-value <0.00001) between the gene expression profiles of pairs of *cis*-eQTL genes.

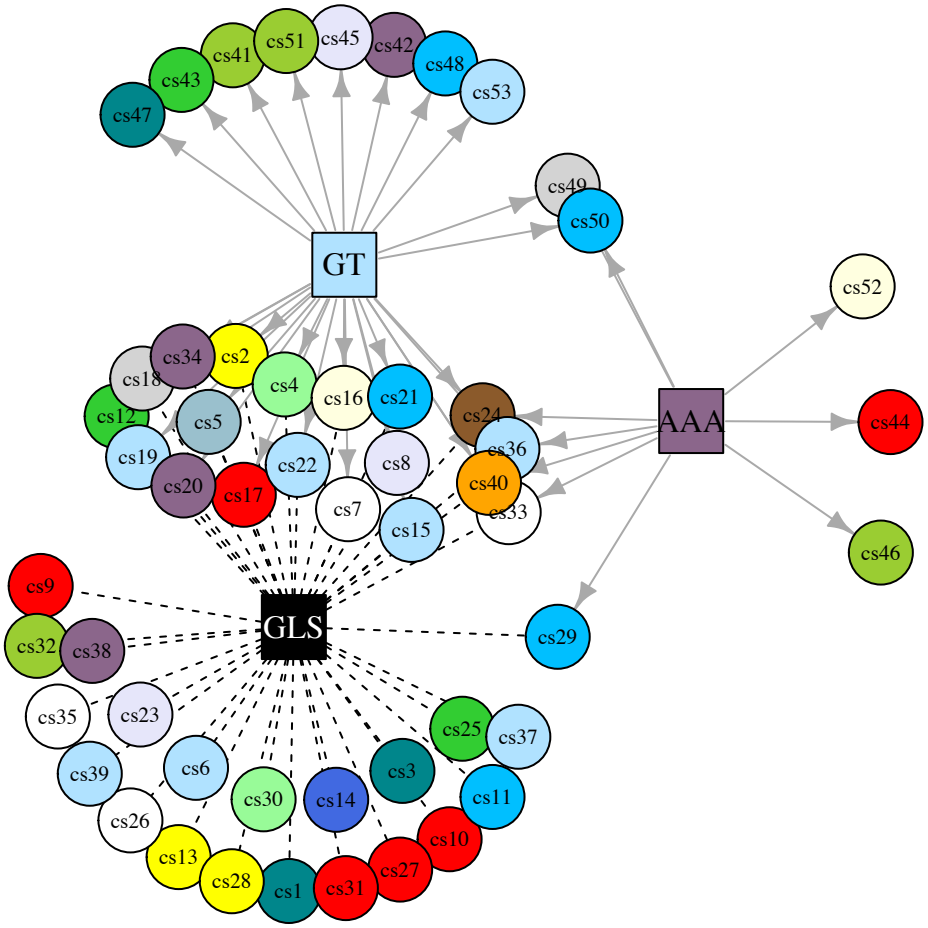


Figure 4.11: QTL 10-10 regulatory network model for genes associated with GLS susceptibility. The black square represents the phenotypic trait "GLS susceptibility". Black dotted lines indicate a strong correlation (p-value <0.00001) between the GLS severity profile and the expression profiles of genes with eQTLs in HS 4-12 S. Coloured square nodes represent *cis*-eQTL genes (GT=glycosyltransferase; AAA=AAA family ATPase) and round nodes *trans*-eQTL genes. The *trans*-eQTL nodes are sequentially numbered from "cs1" to "cs53" where "cs1" represents the *trans*-eQTL gene with the strongest correlation to GLS susceptibility (see Table 4.11 for node annotations). Node colours correspond to the categories in Table 4.7. Solid edges represent gene co-expression between *cis*- and *trans*-eQTL genes (p-value <0.00001); and arrows indicate direction of regulation, assuming that *cis* variation explains gene expression differences for genes in *trans*. Grey dotted lines indicate a strong correlation (p-value <0.00001) between the gene expression profiles of pairs of *cis*-eQTL genes.

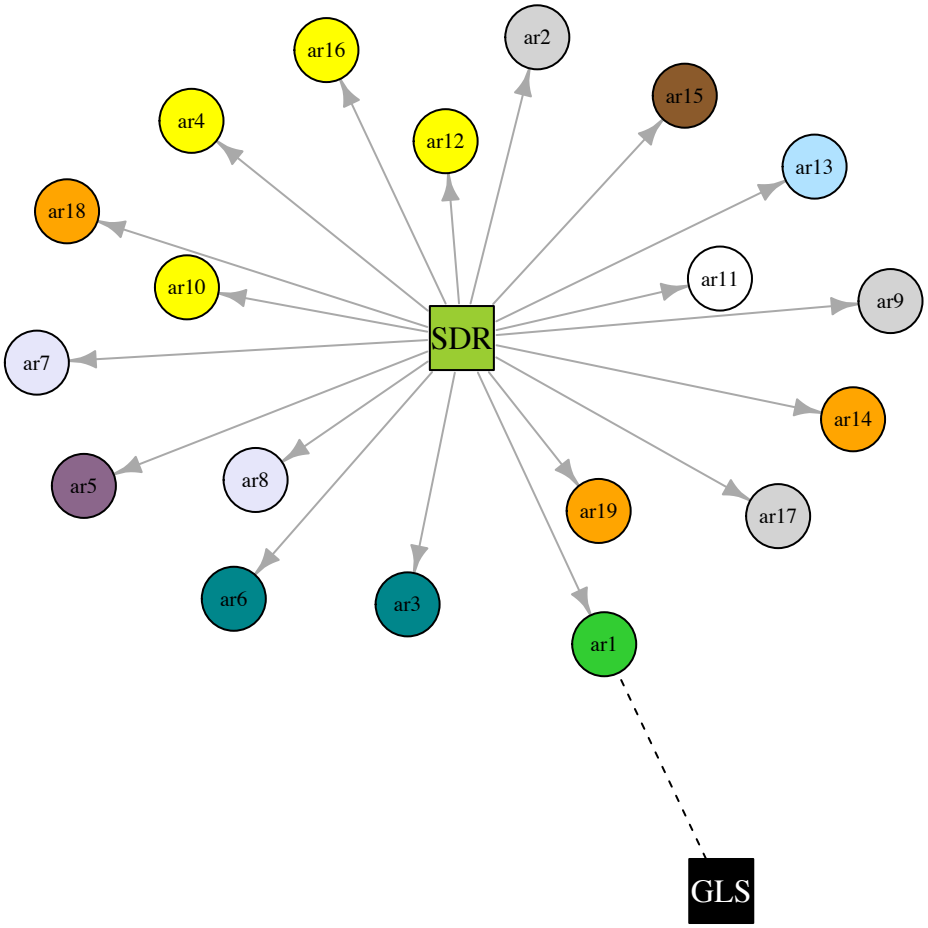


Figure 4.12: QTL 4-11 regulatory network model for genes associated with GLS resistance. The black square represents the phenotypic trait "GLS susceptibility". Black dotted lines indicate a strong correlation (p-value <0.00001) between the GLS severity profile and the expression profiles of genes with eQTLs in HS 4-12 S. Coloured square nodes represent *cis*-eQTL genes (SDR=short-chain dehydrogenase/reductase or rossmann-fold NAD(P)-binding domain-containing protein) and round nodes *trans*-eQTL genes. The *trans*-eQTL nodes are sequentially numbered from "ar1" to "ar19" where "ar1" represents the *trans*-eQTL gene with the strongest correlation to GLS susceptibility (see Table 4.12 for node annotations). Node colours correspond to the categories in Table 4.7. Solid edges represent gene co-expression between *cis*- and *trans*-eQTL genes (p-value <0.00001); and arrows indicate direction of regulation, assuming that *cis* variation explains gene expression differences for genes in *trans*. Grey dotted lines indicate a strong correlation (p-value <0.00001) between the gene expression profiles of pairs of *cis*-eQTL genes.

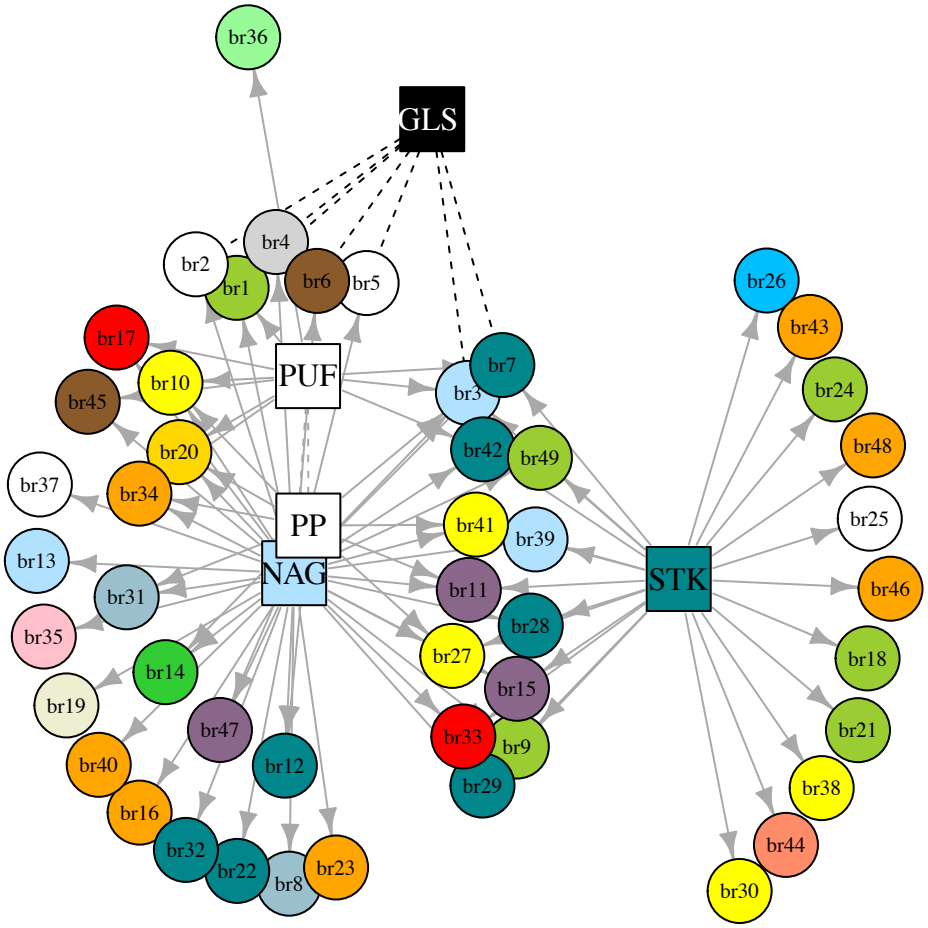


Figure 4.13: QTL 9-5 regulatory network model for genes associated with GLS resistance. The black square represents the phenotypic trait "GLS susceptibility". Black dotted lines indicate a strong correlation (p-value <0.00001) between the GLS severity profile and the expression profiles of genes with eQTLs in HS 4-12 S. Coloured square nodes represent *cis*-eQTL genes (STK=serine threonine-protein kinase; NAG=NAGLU family protein; PP=putative protein; PUF=protein of unknown function) and round nodes *trans*-eQTL genes. The *trans*-eQTL nodes are sequentially numbered from "br1" to "br49" where "br1" represents the *trans*-eQTL gene with the strongest correlation to GLS susceptibility (see Table 4.13 for node annotations). Node colours correspond to the categories in Table 4.7. Solid edges represent gene co-expression between *cis*- and *trans*-eQTL genes (p-value <0.00001); and arrows indicate direction of regulation, assuming that *cis* variation explains gene expression differences for genes in *trans*. Grey dotted lines indicate a strong correlation (p-value <0.00001) between the gene expression profiles of pairs of *cis*-eQTL genes.

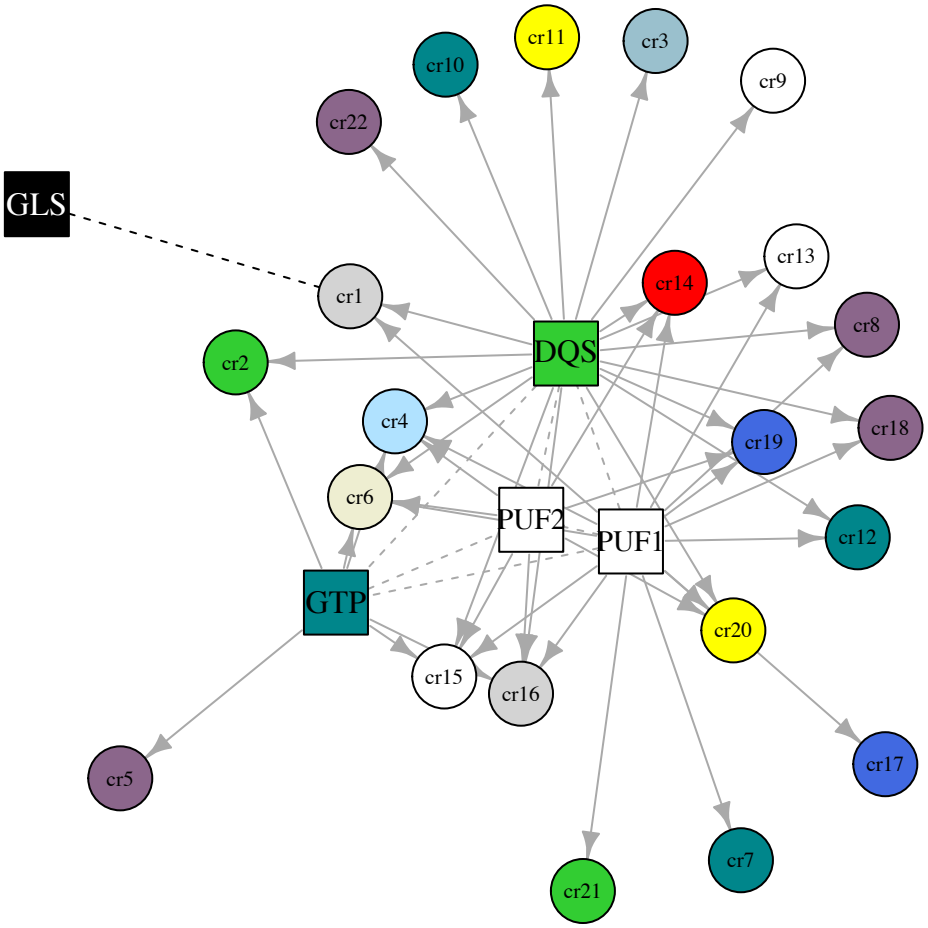


Figure 4.14: QTL 9-7 regulatory network model for genes associated with GLS resistance. The black square represents the phenotypic trait "GLS susceptibility". Black dotted lines indicate a strong correlation (p-value <0.00001) between the GLS severity profile and the expression profiles of genes with eQTLs in HS 4-12 S. Coloured square nodes represent *cis*-eQTL genes (DQS=3-dehydroquinate synthase; GTP=GTP-binding protein; PUF=protein of unknown function) and round nodes *trans*-eQTL genes. The *trans*-eQTL nodes are sequentially numbered from "cr1" to "cr22" where "cr1" represents the *trans*-eQTL gene with the strongest correlation to GLS susceptibility (see Table 4.14 for node annotations). Node colours correspond to the categories in Table 4.7. Solid edges represent gene co-expression between *cis*- and *trans*-eQTL genes (p-value <0.00001); and arrows indicate direction of regulation, assuming that *cis* variation explains gene expression differences for genes in *trans*. Grey dotted lines indicate a strong correlation (p-value <0.00001) between the gene expression profiles of pairs of *cis*-eQTL genes.

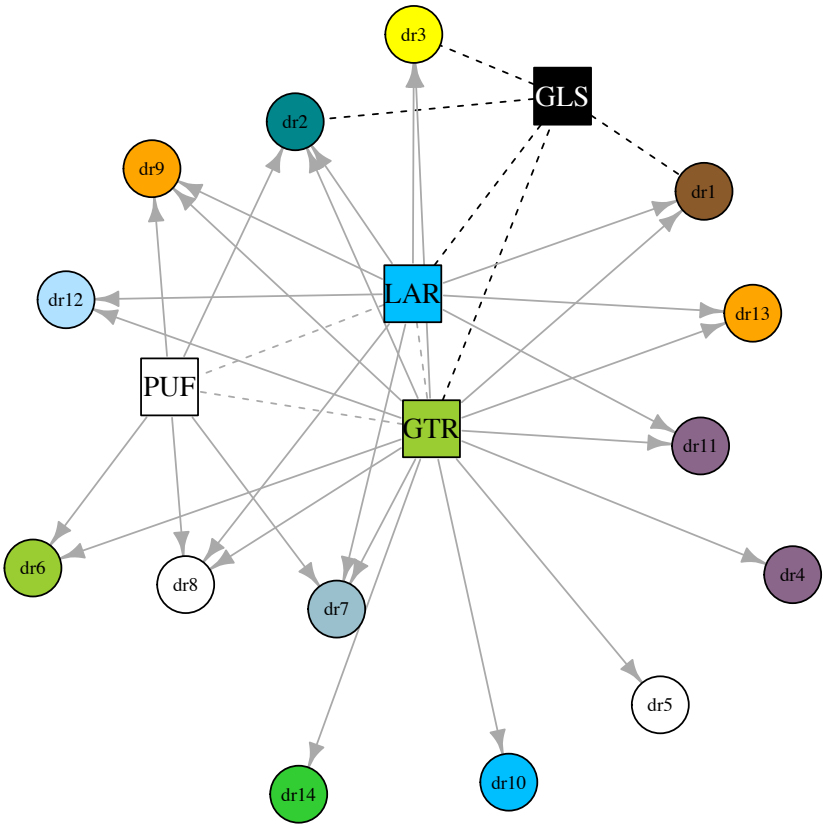


Figure 4.15: QTL 10-10 regulatory network model for genes associated with GLS resistance. The black square represents the phenotypic trait "GLS susceptibility". Black dotted lines indicate a strong correlation (p-value <0.00001) between the GLS severity profile and the expression profiles of genes with eQTLs in HS 4-12 S. Coloured square nodes represent *cis*-eQTL genes (LAR=leucoanthocyanidin reductase; GTR=glutamyl-tRNA reductase; PUF=protein of unknown function) and round nodes *trans*-eQTL genes. The *trans*-eQTL nodes are sequentially numbered from "dr1" to "dr14" where "dr1" represents the *trans*-eQTL gene with the strongest correlation to GLS susceptibility (see Table 4.15 for node annotations). Node colours correspond to the categories in Table 4.7. Solid edges represent gene co-expression between *cis*- and *trans*-eQTL genes (p-value <0.00001); and arrows indicate direction of regulation, assuming that *cis* variation explains gene expression differences for genes in *trans*. Grey dotted lines indicate a strong correlation (p-value <0.00001) between the gene expression profiles of pairs of *cis*-eQTL genes.

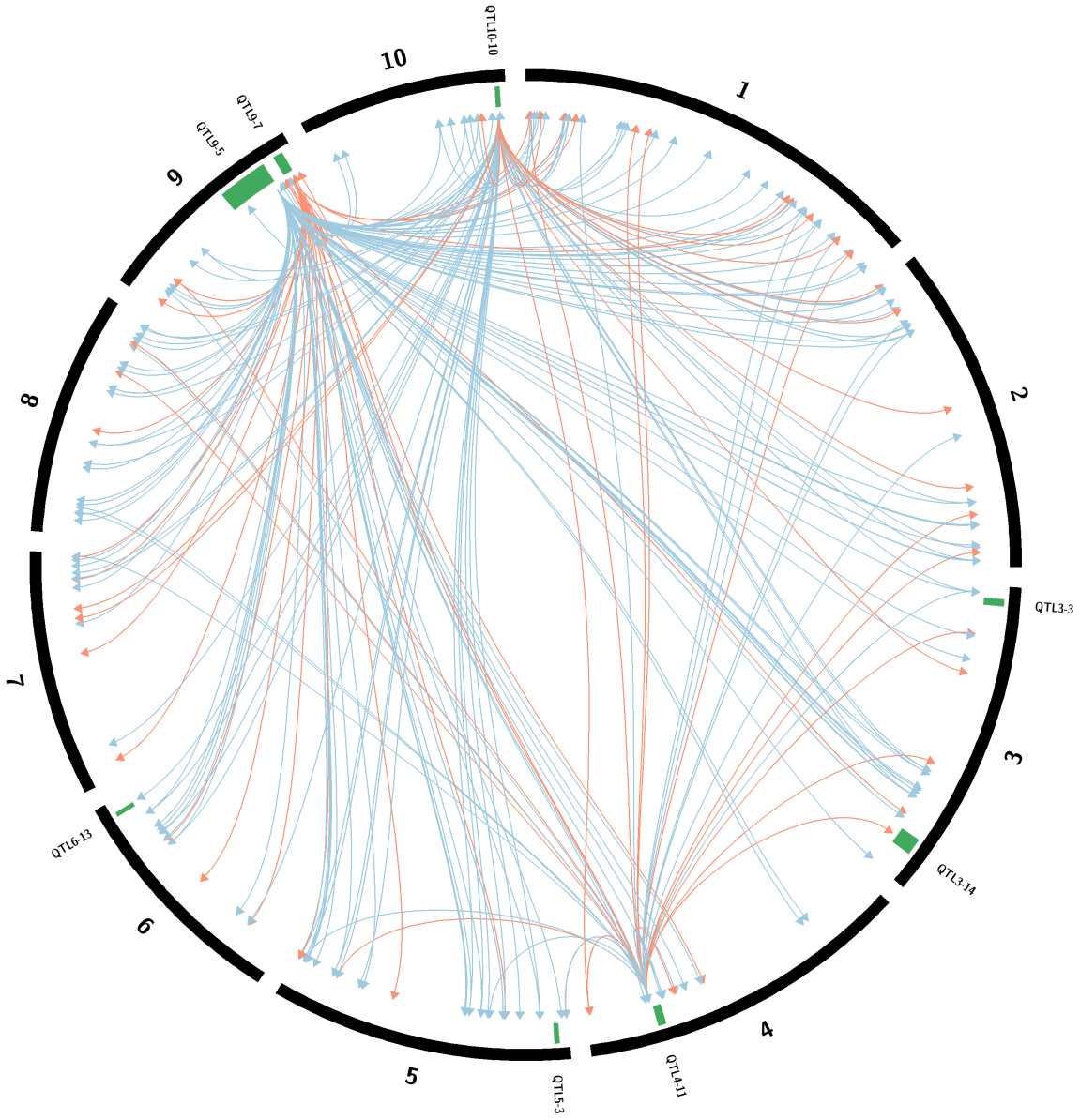


Figure 4.16: Circos plot of candidate genes with *trans*-eQTLs potentially involved in mechanisms associated with GLS resistance or susceptibility. The ten maize chromosomes are indicated in black. Green blocks indicate the eight GLS severity QTLs. Each arrow links a *trans*-eQTL position, coinciding with a GLS severity QTL, with the position of the gene whose expression level is affected by the *trans*-eQTL. Arrowheads indicate gene positions. Only genes with *trans*-eQTLs in the five *trans*-eQTL hotspots that were identified to overlap GLS severity QTLs 3-14, 4-11, 9-5, 9-7 and 10-10, and whose expression profiles were significantly correlated with the GLS severity scores (p-value$<$0.01), are indicated. Red lines link *trans*-eQTLs for which the resistance associated allele (of the coinciding GLS severity QTL) was the allele associated with higher expression. Blue lines link *trans*-eQTLs for which the susceptibility associated allele (of the coinciding GLS severity QTL) was the allele associated with higher expression.

Table 4.1: Eight GLS severity QTLs identified for the CML444×SC Malawi maize RIL population based on GLS severity scores from the field trial at Baynesfield Estate during the 2008/2009 season.

QTL name[a]	Chr[b]	Start marker[c]	Start interval[d]	Start core bin[e]	Peak marker[f]	Peak interval[g]	Peak core bin[h]	End marker[i]	End interval[j]	End core bin[k]	Peak LR[l]	$r^{2\,m}$	$r_t^{2\,n}$	Parent with resistance associated allele[o]	Rating[p]
QTL 3-3	3	3	0.356	3.02	3	0.396	3.03	4	0.4414	3.03	12.68	0.09	0.36	CML 444	1,3
QTL 3-14	3	14	1.6722	3.07	15	1.7957	3.08	15	1.8757	3.08	16.15	0.10	0.35	CML 444	1,2
QTL 4-11	4	11	1.1729	4.08	12	1.2226	4.08	12	1.2626	4.08	14.10	0.07	0.41	CML 444	1
QTL 5-3	5	3	0.3003	5.01	4	0.3759	5.02	5	0.4024	5.02	12.82	0.07	0.33	CML 444	1,2
QTL 6-13	6	13	1.4339	6.06	13	1.5139	6.06	14	1.6177	6.06	17.86	0.18	0.38	SC Malawi	3,4
QTL 9-5	9	5	0.8106	9.03	5	0.8906	9.04	6	0.9614	9.05	13.67	0.09	0.36	CML444	1
QTL 9-7	9	7	1.1382	9.06	8	1.2052	9.06	9	1.3079	9.07	22.99	0.12	0.33	SC Malawi	1,2,3
QTL 10-10	10	10	1.1261	10.06	11	1.2364	10.07	11	1.2964	10.07	19.67	0.14	0.32	CML 444	3,4

[a]QTLs were named based on the chromosome and the number of the start marker on QMap 2.0.
[b]Chromosome number.
[c]Start marker number on QMap 2.0; closest marker to left the QTL region.
[d]Start of leftmost interval of QTL region (distance in Morgans from start of chromosome).
[e]Maize core bin that includes the start interval.
[f]Peak marker number on QMap 2.0; closest marker to left of, or at the QTL peak.
[g]Peak interval of the QTL region (distance in Morgans from start of chromosome).
[h]Maize core bin that includes the peak interval.
[i]End marker number on QMap 2.0; closest marker to left of, or at the end of the QTL region.
[j]End interval of the QTL region (distance in Morgans from start of chromosome).
[k]Maize core bin that includes the end interval.
[l]Peak likelihood ratio (ie. LR> 11.5; LOD > 2.4) – measure of the significance of the marker-trait association.
[m]r^2 is the proportion of the total variation explained by the QTL alone (calculated by QTL Cartographer), converted to percentage.
[n]r_t^2 is the proportion of the total variation explained by the QTL and the background markers and any explanatory variables (calculated in QTL Cartographer), converted to percentage.
[o]Genotype with the GLS resistance associated allele: based on QTL cartographer "z file" output (additive (a) value for the peak interval, i.e. if (a) is negative, genotype = B (SC Malawi); if (a) is positive, genotype = A (CML444)).
[p]GLS rating for which QTL was observed; 1, 2, 3, 4 refer to 92, 99, 109 and 116 days after planting.

Table 4.2: A summary of the numbers of markers, 2 cM bins, reporters and eQTLs per maize chromosome. This table was generated by the "classification" module of the eQTL data analysis pipeline in Galaxy after classification of eQTLs as eQTL *cis* or *trans*.

Chromosome	Number of markers	cM size	Mb size	Number of 2cM bins[a]	Number of Agilent reporters[b]	Number of eQTLs	*Number of cis*-eQTLs	*Number of trans*-eQTLs	Number of unclassified eQTLs[c]
1	27	324.4	301.4	176	5185	4752	800 (17%)	3191 (67%)	761 (16%)
2	14	169.1	237.1	91	3782	2919	538 (18%)	1871 (64%)	510 (17%)
3	19	221.3	232.1	122	3608	3181	483 (15%)	2084 (66%)	614 (19%)
4	20	188.4	241.5	104	3389	2990	502 (17%)	1997 (67%)	491 (16%)
5	20	203.8	217.9	109	3964	4496	670 (15%)	3064 (68%)	762 (17%)
6	17	195.9	169.2	105	2744	2862	415 (15%)	1931 (67%)	516 (18%)
7	12	111.2	176.8	62	2721	2165	413 (19%)	1424 (66%)	328 (15%)
8	14	141.3	175.8	76	2869	2766	514 (19%)	1828 (66%)	424 (15%)
9	11	158.8	156.8	84	2416	2326	291 (13%)	1689 (73%)	346 (15%)
10	13	147.5	150.2	80	2228	3092	353 (11%)	2233 (72%)	506 (16%)
Total	**167**	**1861.6**	**2058.6**	**1009**	**32906**	**31549**	**4979 (16%)**	**21312 (68%)**	**5258 (17%)**

[a]Bins refer to 2 cM intervals used in CIM. Bins were 2 cM in size, except for the last bin before each new marker.
[b]The number of Agilent reporters with known positions on the maize genome, as re-annotated by Coetzer *et al.* (2011).
[c]Some eQTLs could not be classified as *cis* or *trans,* due to uncertainty concerning the genomic position of its linked genes.

Table 4.3: A summary of the 32 *trans*-eQTL hotspots identified across the maize genome. This table summarises a list that was generated by the "GO enrichment" module of the eQTL data analysis pipeline in Galaxy.

Trans-eQTL hotspot name[a]	Chr	Number of 2 cM bins[b]	Number of *trans*-eQTLs in hotspot[c]	Number of *trans*-eQTLs where SC Malawi allele was associated with higher expression (%)[d]	Number of *trans*-eQTLs where CML444 allele was associated with higher expression (%)[e]	Enriched GO-terms (p.val <0.01)[f]	Enriched GO-terms (adj.p.val <0.05)[g]	Directional bias (p.val < 0.05)[h]
HS 1-3	1	4	192	110 (57.3%)	82 (42.7%)	3	0	no
HS 1-8	1	4	202	189 (93.6%)	13 (6.4%)	24	0	yes
HS 1-10	1	2	114	30 (26.3%)	84 (73.7%)	1	0	yes
HS 1-14	1	3	201	122 (60.7%)	79 (39.3%)	4	0	yes
HS 1-18	1	4	347	146 (42.1%)	201 (57.9%)	20	2	yes
HS 1-26	1	3	166	58 (34.9%)	108 (65.1%)	7	0	yes
HS 2-1a	2	2	211	182 (86.3%)	29 (13.7%)	2	0	yes
HS 2-1b	2	3	156	92 (59%)	64 (41%)	1	0	yes
HS 2-7	2	4	176	96 (54.5%)	80 (45.5%)	8	0	no
HS 2-9	2	4	260	208 (80%)	52 (20%)	22	0	yes
HS 3-4	3	4	195	33 (16.9%)	162 (83.1%)	16	0	yes
HS 3-14	3	4	153	112 (73.2%)	41 (26.8%)	1	0	yes
HS 4-3	4	4	197	138 (70.1%)	59 (29.9%)	14	0	yes
HS 4-7	4	2	111	37 (33.3%)	74 (66.7%)	7	0	yes
HS 4-12	4	3	204	118 (57.8%)	86 (42.2%)	3	0	yes
HS 5-6	5	4	199	49 (24.6%)	150 (75.4%)	18	0	yes
HS 5-13	5	4	162	73 (45.1%)	89 (54.9%)	17	0	no
HS 5-16	5	4	239	74 (31%)	165 (69%)	22	0	yes
HS 5-18	5	6	1060	321 (30.3%)	739 (69.7%)	21	0	yes
HS 6-4	6	2	110	84 (76.4%)	26 (23.6%)	14	0	yes
HS 6-11	6	2	141	10 (7.1%)	131 (92.9%)	6	0	yes
HS 6-12	6	4	190	19 (10%)	171 (90%)	22	0	yes
HS 7-6	7	4	283	163 (57.6%)	120 (42.4%)	50	5	yes
HS 7-9	7	3	192	100 (52.1%)	92 (47.9%)	28	0	no
HS 7-11	7	3	183	104 (56.8%)	79 (43.2%)	17	0	no
HS 8-5	8	4	621	373 (60.1%)	248 (39.9%)	5	0	yes
HS 9-6	9	5	477	311 (65.2%)	166 (34.8%)	12	2	yes
HS 9-7	9	4	271	93 (34.3%)	178 (65.7%)	5	0	yes
HS 9-10	9	3	194	88 (45.4%)	106 (54.6%)	6	0	no
HS 10-1	10	5	448	243 (54.2%)	205 (45.8%)	8	0	no
HS 10-7	10	4	545	306 (56.1%)	239 (43.9%)	13	0	yes
HS 10-10	10	4	270	148 (54.8%)	122 (45.2%)	6	0	no

[a]Hotspots were named based on the chromosome and the number of the interval start marker on QMap 2.0; [b]Bins were 2 cM in size, except for the last bin before each new marker. Multiplying the number of bins by 2 thus gives the upper-limit cM size of the *trans*-eQTL hotspot; [c]The number of *trans*-eQTLs with eQTL peaks in the hotspot region; [d]The number of *trans*-eQTLs for which the SC Malawi allele was associated with higher expression; [e]The number of *trans*-eQTLs for which the CML444 allele was associated with higher expression; [f]The number of enriched GO-terms identified by TopGO using a classic Fisher exact test unadjusted (for multiple testing) p-value of 0.01; [g]The number of enriched GO-terms identified by TopGO using a classic Fisher's exact test adjusted (for multiple testing) p-value of 0.05; [h]Pearson's chi-squared tests for excess of positive alleles from one parent (with a p-value cut-off of 0.05); "yes" indicates a significant directional bias.

Table 4.4: Summary statistics of the eQTLs per GLS severity QTL. The table provides the numbers of *cis*- and *trans*-eQTLs that overlapped each GLS severity QTL after: (i) eQTLs were divided into groups based whether higher expression was associated with the resistance or susceptibility associated allele and (ii) genes with expression profiles that significantly correlated with the phenotype values (GLS severity scores) were identified (p-value <0.01). Table 4.1 states whether CML444 or SC Malawi was the parent with the resistance associated allele for each GLS severity QTL.

	QTL 3-3	QTL 3-14	QTL 4-11	QTL 5-3	QTL 6-13	QTL 9-5	QTL 9-7	QTL 10-10	Total
Chromosome	3	3	4	5	6	9	9	10	.
Size in cM	10.54	21.32	10.62	12.21	18.92	16.76	18.97	17.93	127.27
Size in Mb	5.7	12.8	6.4	4.1	3.6	37.9	7.6	4.2	82.3
Number of genes (FGS)	223	539	269	197	227	1196	356	237	3244
Number of eQTL	315	498	570	309	539	695	562	572	4060
Cis-eQTL (% of eQTL)	63 (20%)	84 (17%)	119 (21%)	69 (22%)	59 (11%)	117 (17%)	80 (14%)	63 (11%)	654 (16%)
Trans-eQTL (% of eQTL)	197 (63%)	299 (60%)	351 (62%)	188 (61%)	404 (75%)	480 (69%)	412 (73%)	435 (76%)	2766 (68%)
Not classified (% of eQTL)	55 (17%)	115 (23%)	100 (18%)	52 (17%)	76 (14%)	98 (14%)	70 (12%)	74 (13%)	640 (16%)
***Cis*-eQTL:**									
Resistance associated allele associated with higher expression (% of *cis*-eQTL)	35 (56%)	45 (54%)	66 (55%)	31 (45%)	28 (47%)	62 (53%)	42 (53%)	34 (54%)	343 (52%)
Resistance associated allele associated with higher expression + significant GLS correlation (% of *cis*-eQTL)	7 (11%)	7 (8%)	8 (7%)	5 (7%)	2 (3%)	12 (10%)	5 (6%)	13 (21%)	59 (9%)
Susceptiblity associated allele associated with higher expression (% of *cis*-eQTL)	28 (44%)	39 (46%)	53 (45%)	38 (55%)	31 (53%)	55 (47%)	38 (48%)	29 (46%)	311 (48%)
Susceptiblity associated allele associated with higher expression + significant GLS correlation (% of *cis*-eQTL)	1 (2%)	2 (2%)	10 (8%)	2 (3%)	6 (10%)	8 (7%)	3 (4%)	4 (6%)	36 (6%)
***Trans*-eQTLs:**									
Resistance associated allele associated with higher expression (% of *trans*-eQTL)	127 (64%)	96 (32%)	150 (43%)	100 (53%)	97 (24%)	169 (35%)	163 (40%)	201 (46%)	1103 (40%)
Resistance associated allele associated with higher expression + significant GLS correlation (% of *trans*-eQTL)	9 (5%)	13 (4%)	23 (7%)	17 (9%)	26 (6%)	63 (13%)	36 (9%)	33 (8%)	220 (8%)
Susceptiblity associated allele associated with higher expression (% of *trans*-eQTL)	70 (36%)	203 (68%)	201 (57%)	88 (47%)	307 (76%)	311 (65%)	249 (60%)	234 (54%)	1663 (60%)
Susceptiblity associated allele associated with higher expression + significant GLS correlation (% of *trans*-eQTL)	20 (10%)	28 (9%)	33 (9%)	23 (12%)	91 (23%)	120 (25%)	45 (11%)	88 (20%)	448 (16%)

Table 4.5: Summary statistics of the eQTLs per *trans*-eQTL hotspot that overlapped GLS severity QTLs. The table provides the numbers of *cis*- and *trans*-eQTLs with peaks in each QTL-overlapping *trans*-eQTL hotspot interval after: (i) eQTLs were divided into groups based whether higher expression was associated with the resistance or susceptibility associated allele and (ii) genes with expression profiles that significantly correlated to the phenotype values (GLS severity scores) were identified (p-value <0.01). Table 4.1 states whether CML444 or SC Malawi was the parent with the resistance associated allele for each GLS severity QTL.

	HS 3-14	**HS 4-12**	**HS 9-6**	**HS 9-7**	**HS 10-10**	**Total**
Chromosome	3	4	9	9	10	.
Size in cM	6.35	5.65	9.58	6.7	7.03	35.31
Size in Mb	3.7	6.9	8.1	3.6	1.6	23.9
Number of gene models	148	249	297	182	99	975
Number of eQTL	227	288	604	326	311	1756
Cis-eQTL (% of eQTL)	21 (9%)	34 (12%)	47 (8%)	22 (7%)	8 (3%)	132 (8%)
Trans-eQTL (% of eQTL)	153 (67%)	204 (71%)	477 (79%)	271 (83%)	270 (87%)	1375 (78%)
Not classified (% of eQTL)	53 (23%)	50 (17%)	80 (13%)	33 (10%)	33 (11%)	249 (14%)
***Cis*-eQTL:**						
Resistance associated allele associated with higher expression (% of *cis*-eQTL)	12 (57%)	21 (62%)	22 (47%)	11 (50%)	5 (63%)	71 (54%)
Resistance associated allele associated with higher expression + significant GLS correlation (% of *cis*-eQTL)	2 (10%)	1 (3%)	4 (9%)	4 (18%)	3 (38%)	14 (11%)
Susceptiblity associated allele associated with higher expression (% of *cis*-eQTL)	9 (43%)	13 (38%)	25 (53%)	11 (50%)	3 (38%)	61 (46%)
Susceptiblity associated allele associated with higher expression + significant GLS correlation (% of *cis*-eQTL)	0 (0%)	4 (12%)	3 (6%)	0 (0%)	2 (25%)	9 (7%)
***Trans*-eQTLs:**						
Resistance associated allele associated with higher expression (% of *trans*-eQTL)	41 (27%)	86 (42%)	166 (35%)	93 (34%)	122 (45%)	508 (37%)
Resistance associated allele associated with higher expression + significant GLS correlation (% of *trans*-eQTL)	3 (2%)	22 (11%)	58 (12%)	25 (9%)	21 (8%)	129 (9%)
Susceptiblity associated allele associated with higher expression (% of *trans*-eQTL)	112 (73%)	118 (58%)	311 (65%)	178 (66%)	148 (55%)	867 (63%)
Susceptiblity associated allele associated with higher expression + significant GLS correlation (% of *trans*-eQTL)	5 (3%)	24 (12%)	121 (25%)	37 (14%)	74 (27%)	261 (19%)

Table 4.6: Enriched GO-terms from BiNGO analyses of the *trans*-eQTLs with peaks in the *trans*-eQTL hotspots that overlapped the GLS severity QTLs. All *trans*-eQTLs that belonged to the relevant hotspot were included in the first analysis ("All"), whereafter the subsets of these *trans*-eQTLs (given in Table 4.5) were included in subsequent analyses. "S" and "R" refers to the trans-eQTLs where the parent associated with higher expression was the susceptibility (S) or resistance (R) associated allele, respectively; and "cor" refers to the subset of trans-eQTLs with gene expression profiles that significantly correlated with the GLS severity scores (p-value <0.01).

	GO-ID	Description[a]	p-val[b]	corr p-val[c]	Cluster freq[d]	Total freq[e]
HS 4-12 R	10102	lateral root morphogenesis	1.12E-04	3.07E-02	3/72 4.1%	15/11134 0.1%
	10101	post-embryonic root morphogenesis	1.12E-04	3.07E-02	3/72 4.1%	15/11134 0.1%
HS 4-12 S	42895	antibiotic transporter activity	7.83E-05	2.48E-02	2/99 2.0%	2/11134 0.0%
	8493	tetracycline transporter activity	7.83E-05	2.48E-02	2/99 2.0%	2/11134 0.0%

	GO-ID	Description[a]	p-val[b]	corr p-val[c]	Cluster freq[d]	Total freq[e]
HS 9-6 All *	5886	plasma membrane	4.19E-05	3.81E-03	56/391 14.3%	930/11134 8.3%
	3824	catalytic activity	3.18E-04	1.40E-02	167/391 42.7%	3831/11134 34.4%
	8152	metabolic process	5.65E-04	1.40E-02	146/391 37.3%	3305/11134 29.6%
	16020	membrane	6.14E-04	1.40E-02	94/391 24.0%	1957/11134 17.5%
	3674	molecular_function	1.88E-03	3.42E-02	367/391 93.8%	9975/11134 89.5%
	5975	carbohydrate metabolic process	3.12E-03	4.31E-02	25/391 6.3%	394/11134 3.5%
	19725	cellular homeostasis	3.31E-03	4.31E-02	8/391 2.0%	71/11134 0.6%
HS 9-6 R *	5886	plasma membrane	2.15E-04	1.75E-02	25/140 17.8%	930/11134 8.3%
HS 9-6 S cor *	3824	catalytic activity	2.79E-07	2.17E-05	57/95 60.0%	3831/11134 34.4%
	8152	metabolic process	9.04E-04	2.56E-02	43/95 45.2%	3305/11134 29.6%
	19748	secondary metabolic process	9.83E-04	2.56E-02	6/95 6.3%	134/11134 1.2%
	5618	cell wall	2.93E-03	4.03E-02	7/95 7.3%	223/11134 2.0%
	16740	transferase activity	3.08E-03	4.03E-02	20/95 21.0%	1225/11134 11.0%
	30312	external encapsulating structure	3.15E-03	4.03E-02	7/95 7.3%	226/11134 2.0%
	9058	biosynthetic process	3.62E-03	4.03E-02	19/95 20.0%	1154/11134 10.3%

	GO-ID	Description[a]	p-val[b]	corr p-val[c]	Cluster freq[d]	Total freq[e]
HS 10-10 All *	3824	catalytic activity	9.27E-05	7.88E-03	104/223 46.6%	3831/11134 34.4%
	9607	response to biotic stimulus	1.85E-03	3.33E-02	15/223 6.7%	318/11134 2.8%
	16301	kinase activity	1.93E-03	3.33E-02	22/223 9.8%	560/11134 5.0%
	16740	transferase activity	2.23E-03	3.33E-02	39/223 17.4%	1225/11134 11.0%
	16265	death	2.35E-03	3.33E-02	6/223 2.6%	68/11134 0.6%
	8219	cell death	2.35E-03	3.33E-02	6/223 2.6%	68/11134 0.6%
HS 10-10 S	19200	carbohydrate kinase activity	4.99E-05	2.08E-02	4/124 3.2%	19/11134 0.1%
	51704	multi-organism process	1.11E-04	2.08E-02	14/124 11.2%	391/11134 3.5%
	3824	catalytic activity	1.16E-04	2.08E-02	63/124 50.8%	3831/11134 34.4%
	8865	fructokinase activity	1.23E-04	2.08E-02	2/124 1.6%	2/11134 0.0%
	4340	glucokinase activity	1.23E-04	2.08E-02	2/124 1.6%	2/11134 0.0%
	51707	response to other organism	1.48E-04	2.08E-02	12/124 9.6%	305/11134 2.7%
	9607	response to biotic stimulus	2.18E-04	2.63E-02	12/124 9.6%	318/11134 2.8%
	4396	hexokinase activity	3.66E-04	3.87E-02	2/124 1.6%	3/11134 0.0%
HS 10-10 S cor	3824	catalytic activity	4.33E-06	2.40E-03	38/60 63.3%	3831/11134 34.4%
	19748	secondary metabolic process	7.98E-05	1.63E-02	6/60 10.0%	134/11134 1.2%
	9699	phenylpropanoid biosynthetic process	8.81E-05	1.63E-02	4/60 6.6%	44/11134 0.3%
	16740	transferase activity	1.75E-04	2.43E-02	17/60 28.3%	1225/11134 11.0%
	9698	phenylpropanoid metabolic process	2.78E-04	3.09E-02	4/60 6.6%	59/11134 0.5%

*The analysis was based on the plant-specific GO slim ontology as opposed to the full GO hierarchy.

[a]The GO-term definition that corresponds to the GO-ID.

[b]The hypergeometric test unadjusted (for multiple testing) p-value for GO enrichment.

[c]Adjusted p-value, based on Benjamini and Hochberg correction applied on hypergeometric test.

[d]Cluster frequency represents the total number of genes annotated to that GO term divided by total number of genes in the test set.

[e]Total frequency represents the total number of genes annotated to that GO term divided by total number of genes in the reference set.

Table 4.7: Functional categories (based on MapMan bins) and the associated colours of genes referred to in Tables 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.14 and 4.15.

Category name	Category description	Colour and acronym
Amino acid metabolism	Amino acid synthesis and degradation	AAM
Carbohydrate metabolism	Major and minor carbohydrate metabolism (including UDP glucosyl transferases, Beta-1,3-glucan hydrolases)	CHO
Cell	Cell organisation, division, cycle, growth	CELL
Cell wall	Cell wall synthesis, degradation and modification	CW
DNA	DNA synthesis/chromatin structure and repair	DNA
Hormone metabolism	Hormone (auxin, GA, cytokinin, ABA, ET, JA, SA) synthesis and sensing	HM
Lipid metabolism	Fatty acid synthesis, elongation, desaturation and lipid degradation	LM
Metal handling	Metal binding, chelation and storage	MH
Nucleotide metabolism	Nucleotide synthesis, degradation and salvage	NM
Photosynthesis	Lightreactions, photorespiration, chlorophyll synthase	PS
Protein	Ribosomal / mitochondrial protein processing and synthesis; Post-translational modification; Protein degradation through proteases, ubiquitination etc.	PROT
Redox	Heme, thioreredoxin, ascorbate and glutathione, glutaredoxis, periredoxin, dismutase/catalase	RX
RNA	RNA processing, transcription and regulation of transcription	RNA
Secondary metabolism	Biosynthesis of secondary metabolites (including phenylpropanoids, flavonoids)	SM
Signalling	Proteins with signalling functions e.g. receptor kinases, G-proteins, calcium signaling proteins, phosphatases, kinases	SIGN
Stress	Biotic stress (BS); Abiotic stress (AS) e.g. glutathione S transferases	BS/AS
Transport	Hormone and membrane system transport; vesicle transport; ABC transporters and multidrug resistance systems	TR
Tricarboxylic acid (TCA) cycle	Organic acid transformations, tetrapyrrole	TCA
Not assigned	Other / Non-specific annotation, no ontology	NA
Protein of unknown function	Protein of unknown function; Putative protein	PUF/PP

Table 4.8: The node annotations for the QTL 4-11 regulatory network model for genes associated with GLS susceptibility (Figure 4.9 on page 200). For a description of the categories in the "MM (MapMan) bin" column refer to Table 4.7.

Symbol	Agilent ID	Cis/ Trans	GLS cor	Annotation (AT, Rice, B2G)	MM bin
ZF	A_92_P039457	cis	0.34	DHHC-type zinc finger family protein	RNA
NX	A_92_P038137	cis	0.44	Hydrolase, NUDIX family, domain containing protein	NM
TPR	A_92_P040694	cis	0.26	TPR repeat region family protein	PROT
PB	A_92_P036978	cis	0.37	Protein binding	NA
as1	A_92_P037621	trans	0.60	F-box/kelch-repeat protein SKIP11-like	PROT
as9	A_92_P041707	trans	0.36	Mitochondrial-processing peptidase subunit beta-like	PROT
as3	A_92_P041126	trans	0.44	Glycolipid transfer protein (GLTP) family protein	LM
as8	A_92_P041531	trans	0.38	Phosphatidylinositol transfer protein 1-like	LM
as11	A_92_P038869	trans	0.34	Serine threonine-protein kinase OXI1-like	SIGN
as13	A_92_P028767	trans	0.31	DNA-binding protein phosphatase 1	SIGN
as14	A_92_P029615	trans	0.31	Polyadenylate-binding cytoplasmic and nuclear-like	NM
as16	A_92_P035084	trans	0.30	Nicotianamine synthase 3	NM
as2	A_92_P036877	trans	0.48	Glucan endo-1,3-beta-glucosidase precursor	CHO
as4	A_92_P036813	trans	0.44	Oxidoreductase / transition metal ion binding protein	MH
as7	A_92_P029218	trans	0.39	Glutaredoxin subgroup I	RX
as10	A_92_P035799	trans	0.34	Serine hydrolase domain containing protein	AAM
as15	A_92_P041146	trans	0.30	2-oxoglutarate and Fe(II)-dependent oxygenase superfamily	HM
as17	A_92_P029298	trans	0.29	Universal stress protein domain containing protein	AS
as19	A_92_P039285	trans	0.27	Type I site-specific deoxyribonuclease	DNA
as6	A_92_P029256	trans	0.40	VQ motif-containing protein	NA
as18	A_92_P035773	trans	0.29	ACT domain containing protein	NA
as5	A_92_P041670	trans	0.40	Protein of unknown function	PUF
as12	A_92_P030773	trans	0.32	Protein of unknown function	PUF

Table 4.9: The node annotations (part I) for the QTL 9-5 regulatory network model for genes associated with GLS susceptibility (Figure 4.10 on page 201). For a description of the categories in the "MM (MapMan) bin" column refer to Table 4.7.

Symbol	Agilent ID	Cis/ Trans	GLS cor	Annotation (AT, Rice, B2G)	MM bin
EF	A_92_P037035	cis	0.40	EF hand / calmodulin-related calcium sensor protein	SIGN
ECH	A_92_P022781	cis	0.26	Enoyl-CoA hydratase/isomerase family protein	LM
PUF	A_92_P028216	cis	0.31	Protein of unknonw function	PUF
bs9	A_92_P007649	trans	0.54	Beta-glucanase / glycosyl hydrolases family 17	BS
bs13	A_92_P006679	trans	0.52	Bowman-birk type trypsin inhibitor	BS
bs15	A_92_P022755	trans	0.52	Chitinase	BS
bs37	A_92_P034379	trans	0.45	Maize proteinase inhibitor	BS
bs44	A_92_P001063	trans	0.41	Class III chitinase	BS
bs46	A_92_P023606	trans	0.40	Pathogenesis-related maize seed protein	BS
bs53	A_92_P009951	trans	0.39	C2 calcium/lipid-binding and GRAM domain containing	BS
bs57	A_92_P030266	trans	0.38	Dirigent-like protein	BS
bs62	A_92_P032100	trans	0.37	Cysteine proteinase inhibitor 8 precursor	BS
bs71	A_92_P022862	trans	0.35	GRAM domain-containing protein 1A-like	BS
bs84	A_92_P030068	trans	0.32	OTU-like cysteine protease-like	BS
bs20	A_92_P022951	trans	0.50	Glutathione S-transferase	AS
bs98	A_92_P022861	trans	0.28	Glutathione S-transferase / Integral membrane protein	AS
bs7	A_92_P021227	trans	0.55	Calmodulin-binding heat-shock protein	SIGN
bs23	A_92_P009091	trans	0.49	Cysteine-rich receptor-like protein kinase 10-like	SIGN
bs27	A_92_P006094	trans	0.48	Calmodulin-binding protein	SIGN
bs43	A_92_P039626	trans	0.42	Receptor-like serine threonine protein kinase ark3	SIGN
bs45	A_92_P030257	trans	0.41	TBC domain containing protein	SIGN
bs49	A_92_P029167	trans	0.40	LRR protein kinase family protein	SIGN
bs50	A_92_P022078	trans	0.39	Chloride channel family	SIGN
bs54	A_92_P006123	trans	0.39	Calcium-dependent protein kinase	SIGN
bs75	A_92_P017691	trans	0.34	Protein strubbelig-receptor family 3-like	SIGN
bs93	A_92_P020152	trans	0.29	Probable LRR receptor-like serine threonine-protein kinase	SIGN
bs95	A_92_P039597	trans	0.29	Protein phosphatase 2c	SIGN
bs97	A_92_P019150	trans	0.28	Small GTP-binding protein	SIGN
bs2	A_92_P023090	trans	0.56	Cytochrome P450 / Transposon	SM
bs19	A_92_P031017	trans	0.50	Phenylalanine ammonia-lyase (PAL)	SM
bs22	A_92_P032474	trans	0.49	Catalytic LigB subunit of aromatic ring-opening dioxygenase	SM
bs26	A_92_P007140	trans	0.48	Laccase 7	SM
bs30	A_92_P017679	trans	0.46	Flavanone 3-hydroxylase	SM
bs65	A_92_P020690	trans	0.36	HXXXD-type acyl-transferase family protein	SM
bs77	A_92_P041061	trans	0.34	Cytochrome P450	SM
bs80	A_92_P013002	trans	0.33	Dihydroflavonol-4-reductase	SM
bs90	A_92_P036137	trans	0.30	Cytochrome P450 / ABA 8-hydroxylase 2	SM
bs1	A_92_P035928	trans	0.62	Dihydrolipoyllysine-residue acetyltransferase	CHO
bs10	A_92_P022183	trans	0.53	Lactate dehydrogenase	CHO
bs38	A_92_P026596	trans	0.45	Glycosyltransferase	CHO
bs42	A_92_P018154	trans	0.43	Cytokinin-o-glucosyltransferase 2	CHO
bs51	A_92_P041308	trans	0.39	Trichome birefringence-like 11	CHO
bs61	A_92_P018852	trans	0.37	Glycosyl hydrolase	CHO
bs69	A_92_P019555	trans	0.36	Core-2/I-branching beta-N-acetylglucosaminyltransferase	CHO
bs79	A_92_P034521	trans	0.33	NADP-dependent oxidoreductase	CHO
bs87	A_92_P038403	trans	0.31	Core-2/I-branching beta-N-acetylglucosaminyltransferase	CHO
bs5	A_92_P018938	trans	0.55	RING finger and CHY zinc finger domain-containing protein	PROT
bs11	A_92_P030102	trans	0.53	Outer envelope membrane protein 7	PROT
bs14	A_92_P026115	trans	0.52	Kelch motif family protein	PROT
bs16	A_92_P021296	trans	0.51	Ring-H2 finger protein	PROT
bs17	A_92_P020857	trans	0.50	E3 ubiquitin-protein ligase RGLG2	PROT
bs55	A_92_P023478	trans	0.39	Peptidase M16 family protein	PROT
bs60	A_92_P031777	trans	0.37	Signal peptide peptidase 2B-like	PROT
bs86	A_92_P037279	trans	0.31	50S ribosomal protein L12-2	PROT

Table 4.10: The node annotations (part II) for the QTL 9-5 regulatory network model for genes associated with GLS susceptibility (Figure 4.10 on page 201). For a description of the categories in the "MM (MapMan) bin" column refer to Table 4.7.

Symbol	Agilent ID	Cis/ Trans	GLS cor	Annotation (AT, Rice, B2G)	MM bin
bs21	A_92_P030884	trans	0.50	ABC transporter C family member 8-like	TR
bs52	A_92_P020294	trans	0.39	ATP binding protein	TR
bs59	A_92_P028689	trans	0.38	Sugar phosphate phosphate translocator	TR
bs63	A_92_P020891	trans	0.37	Pollen-specific SF21 / N-MYC downregulated-like	TR
bs68	A_92_P021776	trans	0.36	ABC transporter B family member 19-like	TR
bs82	A_92_P035625	trans	0.32	Integral membrane protein / nodulin MtN21	TR
bs99	A_92_P019916	trans	0.28	Glycerol 3-phosphate permease	TR
bs4	A_92_P018873	trans	0.56	WRKY transcription factor	RNA
bs48	A_92_P039696	trans	0.40	NAC domain-containing protein 68	RNA
bs67	A_92_P023369	trans	0.36	Leucine zipper	RNA
bs85	A_92_P038241	trans	0.31	Heat stress transcription factor A-5-like	RNA
bs101	A_92_P025044	trans	0.28	CCR4-not transcription complex subunit 7	RNA
bs104	A_92_P011891	trans	0.26	bZIP transcription factor bZIP28	RNA
bs25	A_92_P032943	trans	0.48	3-ketoacyl-synthase 1	LM
bs31	A_92_P012775	trans	0.46	Peroxisomal 3-ketoacyl-CoA thiolase 3	LM
bs58	A_92_P039784	trans	0.38	AMP-binding protein	LM
bs64	A_92_P024923	trans	0.37	Omega-6 fatty acid endoplasmic reticulum isozyme 2	LM
bs83	A_92_P011631	trans	0.32	Phosphatidic acid phosphatase-like protein	LM
bs102	A_92_P035839	trans	0.27	Acyl carrier protein	LM
bs12	A_92_P021132	trans	0.52	Xylanase inhibitor	CW
bs24	A_92_P017529	trans	0.49	Secondary cell wall-related glycosyltransferase family 8	CW
bs33	A_92_P029498	trans	0.46	Pectin methylesterase	CW
bs70	A_92_P025721	trans	0.36	Galactoside 2-alpha-l-fucosyltransferase-like	CW
bs47	A_92_P003729	trans	0.40	Metallothionein 2A	MH
bs74	A_92_P007799	trans	0.34	Molybdopterin cofactor	MH
bs81	A_92_P032564	trans	0.33	SelT-like protein precursor	MH
bs29	A_92_P039018	trans	0.46	ACC synthase	HM
bs73	A_92_P026516	trans	0.35	1-aminocyclopropane-1-carboxylate oxidase 1	HM
bs103	A_92_P018101	trans	0.27	Aldehyde oxidase	HM
bs34	A_92_P021328	trans	0.46	Nicotinate phosphoribosyltransferase	NM
bs41	A_92_P038137	trans	0.44	Hydrolase, NUDIX family, domain containing protein	NM
bs96	A_92_P032260	trans	0.29	Cytidine/deoxycytidylate deaminase family protein	NM
bs6	A_92_P030005	trans	0.55	Monodehydroascorbate reductase	RX
bs18	A_92_P022165	trans	0.50	GMC oxidoreductase family / Alcohol oxidase	RX
bs35	A_92_P009408	trans	0.46	COP1-interacting protein 7-like protein	PS
bs66	A_92_P022882	trans	0.36	Protein kinase chloroplast	PS
bs32	A_92_P016791	trans	0.46	Chromatin assembly factor-1	DNA
bs78	A_92_P026520	trans	0.33	Glutamate decarboxylase	AAM
bs100	A_92_P031077	trans	0.28	Aconitate hydratase 1	TCA
bs72	A_92_P009914	trans	0.35	HD domain containing / RELA/SPOT homolog 3	NA
bs88	A_92_P031177	trans	0.31	Autophagy-related protein 13	NA
bs89	A_92_P036534	trans	0.31	GPI-C transferase complex	NA
bs91	A_92_P008773	trans	0.30	Phosphoribosylaminoimidazole catalytic subunit	NA
bs92	A_92_P014552	trans	0.30	Nitrate reductase	NA
bs94	A_92_P021234	trans	0.29	SAD1/UNC-84 domain protein 1	NA
bs3	A_92_P013177	trans	0.56	Protein of unknown function	PUF
bs8	A_92_P001316	trans	0.55	Protein of unknown function	PUF
bs28	A_92_P013182	trans	0.47	Protein of unknown function	PUF
bs36	A_92_P013412	trans	0.45	Protein of unknown function	PUF
bs76	A_92_P008269	trans	0.34	Protein of unknown function	PUF
bs39	A_92_P007024	trans	0.45	Putative protein	PP
bs40	A_92_P007880	trans	0.44	Putative protein	PP
bs56	A_92_P016055	trans	0.38	Putative protein	PP

Table 4.11: The node annotations for the QTL 10-10 regulatory network model for genes associated with GLS susceptibility (Figure 4.11 on page 202). For a description of the categories in the "MM (MapMan) bin" column refer to Table 4.7.

Symbol	Agilent ID	Cis/ Trans	GLS cor	Annotation (AT, Rice, B2G)	MM bin
GT	A_92_P034905	cis	0.40	Glycosyltransferase protein A	CHO
AAA	A_92_P025529	cis	0.31	AAA family ATPase peroxin 6	TR
cs6	A_92_P022183	trans	0.53	Lactate/malate dehydrogenase family protein	CHO
cs15	A_92_P013322	trans	0.50	Glycosyltransferase family 61 protein	CHO
cs19	A_92_P018607	trans	0.50	Fructokinase 1	CHO
cs22	A_92_P032638	trans	0.49	Hexokinase 2	CHO
cs36	A_92_P026596	trans	0.45	Glycosyltransferase	CHO
cs37	A_92_P005512	trans	0.44	Glycosyl hydrolase	CHO
cs39	A_92_P016465	trans	0.44	Lactate dehydrogenase	CHO
cs53	A_92_P038403	trans	0.31	Core-2/I-branching beta-N-acetylglucosaminyltransferase	CHO
cs9	A_92_P006679	trans	0.52	Bowman-birk type trypsin inhibitor	BS
cs10	A_92_P022755	trans	0.52	Chitinase	BS
cs27	A_92_P007388	trans	0.46	Beta-glucanase / glycosyl hydrolases family 17	BS
cs44	A_92_P018920	trans	0.38	C2 domain-containing	BS
cs17	A_92_P022951	trans	0.50	Glutathione S-transferase	AS
cs31	A_92_P019132	trans	0.46	Heat shock transcription factor HSF4	AS
cs11	A_92_P014030	trans	0.51	Hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferase	SM
cs21	A_92_P032474	trans	0.49	Catalytic LigB subunit of aromatic ring-opening dioxygenase	SM
cs29	A_92_P017679	trans	0.46	Flavanone 3-hydroxylase	SM
cs48	A_92_P000336	trans	0.35	Phenylalanine ammonia-lyase (PAL)	SM
cs50	A_92_P041061	trans	0.34	Cytochrome P450	SM
cs20	A_92_P030884	trans	0.50	ABC transporter c family member 8-like	TR
cs34	A_92_P019947	trans	0.45	White-brown-complex ABC transporter family	TR
cs38	A_92_P002836	trans	0.44	NSP (nuclear shuttle protein)-interacting GTPase	TR
cs42	A_92_P020294	trans	0.39	ATP binding protein	TR
cs32	A_92_P009408	trans	0.46	Cop1-interacting protein 7-like protein	PS
cs41	A_92_P002419	trans	0.42	Chloroplastic group IIA intron splicing facilitator chloroplast	PS
cs46	A_92_P022882	trans	0.36	Protein kinase chloroplast	PS
cs51	A_92_P023400	trans	0.34	Protein thylakoid chloroplastic-like	PS
cs1	A_92_P014596	trans	0.58	Two-component system sensor kinase	SIGN
cs3	A_92_P021227	trans	0.55	Calmodulin-binding heat-shock protein	SIGN
cs47	A_92_P002935	trans	0.36	Calcium lipid binding	SIGN
cs2	A_92_P018873	trans	0.56	WRKY transcription factor	RNA
cs13	A_92_P013324	trans	0.50	NAC domain transcription factor	RNA
cs28	A_92_P015550	trans	0.46	WRKY transcription factor	RNA
cs8	A_92_P021132	trans	0.52	Xylanase inhibitor	CW
cs23	A_92_P017529	trans	0.49	Secondary cell wall-related glycosyltransferase family 8	CW
cs45	A_92_P022159	trans	0.38	Hydroxyproline-rich glycoprotein	CW
cs12	A_92_P021296	trans	0.51	Ring-H2 finger protein	AAM
cs25	A_92_P010667	trans	0.48	Branched-chain-amino-acid aminotransferase chloroplast	AAM
cs43	A_92_P023961	trans	0.38	Alanine aminotransferase	AAM
cs4	A_92_P038520	trans	0.54	Cytidine deaminase	NM
cs30	A_92_P021328	trans	0.46	Nicotinate phosphoribosyltransferase	NM
cs16	A_92_P025853	trans	0.50	Cyclin-dependent kinase D1-like	CELL
cs52	A_92_P023068	trans	0.34	Cell division protein AAA ATPase family	CELL
cs5	A_92_P030660	trans	0.53	Nuclease I	DNA
cs14	A_92_P022165	trans	0.50	GMC oxidoreductase family / Alcohol oxidase	RX
cs24	A_92_P032943	trans	0.48	3-ketoacyl-synthase 1	LM
cs40	A_92_P031757	trans	0.43	Aspartic proteinase-like protein 1-like	PROT
cs18	A_92_P021490	trans	0.50	WPP domain-interacting protein 1-like	NA
cs49	A_92_P023035	trans	0.34	Alpha/beta-hydrolases superfamily protein	NA
cs7	A_92_P038555	trans	0.52	Protein of unknown function	PUF
cs26	A_92_P008339	trans	0.48	Protein of unknown function	PUF
cs33	A_92_P016734	trans	0.46	Protein of unknown function	PUF
cs35	A_92_P019435	trans	0.45	Protein of unknown function	PUF

Table 4.12: The node annotations for the QTL 4-11 regulatory network model for genes associated with GLS resistance (Figure 4.12 on page 203). For a description of the categories in the "MM (MapMan) bin" column refer to Table 4.7.

Symbol	Agilent ID	Cis/ Trans	GLS cor	Annotation (AT, Rice, B2G)	MM bin
SDR	A_92_P008436	cis	-0.36	Rossmann-fold NAD(P)-binding domain-containing	PS
ar4	A_92_P008422	trans	-0.33	Transcription factor HBP-1a	RNA
ar10	A_92_P000526	trans	-0.28	RNA recognition motif-containing protein	RNA
ar12	A_92_P015590	trans	-0.28	Zinc finger CCCH type domain-containing protein	RNA
ar16	A_92_P007314	trans	-0.27	Coiled-coil domain-containing protein	RNA
ar14	A_92_P008657	trans	-0.27	WWE protein-protein interaction domain protein family	PROT
ar18	A_92_P000772	trans	-0.26	tRNA methyltransferase	PROT
ar19	A_92_P005501	trans	-0.26	Zinc RING finger protein	PROT
ar3	A_92_P005700	trans	-0.33	RanBP1 domain containing protein	SIGN
ar6	A_92_P001026	trans	-0.31	Adenylyl-sulfate kinase	SIGN
ar7	A_92_P007232	trans	-0.30	Bifunctional polymyxin resistance arnA protein	CW
ar8	A_92_P004318	trans	-0.30	Post-GPI attachment to proteins / PER-1-like	CW
ar1	A_92_P006115	trans	-0.45	Erythronate-4-phosphate dehydrogenase family protein	AAM
ar5	A_92_P005524	trans	-0.32	Cation efflux family protein	TR
ar13	A_92_P013690	trans	-0.27	Endonuclease or glycosyl hydrolase	CHO
ar15	A_92_P012338	trans	-0.27	ASC1-like protein	LM
ar2	A_92_P009439	trans	-0.35	BSD domain-containing protein	NA
ar9	A_92_P000691	trans	-0.29	Nuclear protein E3-isoform	NA
ar17	A_92_P012710	trans	-0.26	XAP5 family protein	NA
ar11	A_92_P008901	trans	-0.28	Protein of unknown function	PUF

Table 4.13: The node annotations for the QTL 9-5 regulatory network model for genes associated with GLS resistance (Figure 4.13 on page 204). For a description of the categories in the "MM (MapMan) bin" column refer to Table 4.7.

Symbol	Agilent ID	Cis/ Trans	GLS cor	Annotation (AT, Rice, B2G)	MM bin
STK	A_92_P033066	cis	-0.39	Serine threonine-protein kinase SAPK2	SIGN
NAG	A_92_P009668	cis	-0.38	Alpha-N-acetylglucosaminidase family / NAGLU family	CHO
PP	A_92_P001417	cis	-0.34	Putative protein	PP
PUF	A_92_P001249	cis	-0.33	Protein of unknown function	PUF
br7	A_92_P012575	trans	-0.44	Cysteine-rich receptor-like protein kinase 25-like	SIGN
br12	A_92_P009393	trans	-0.37	Phosphatidylinositol 3- and 4-kinase family-like	SIGN
br22	A_92_P005700	trans	-0.33	RanBP1 domain containing protein	SIGN
br28	A_92_P010477	trans	-0.31	Serine threonine protein phosphatase	SIGN
br29	A_92_P008364	trans	-0.31	Receptor-like serine threonine kinase	SIGN
br32	A_92_P017199	trans	-0.31	Protein kinase family protein	SIGN
br42	A_92_P010609	trans	-0.28	Serine threonine-protein phosphatase BSL2-like	SIGN
br13	A_92_P007992	trans	-0.36	ATP synthase mitochondrial F1 complex assembly factor 1	PROT
br16	A_92_P008892	trans	-0.34	Outer membrane OMP85 family protein	PROT
br23	A_92_P013016	trans	-0.33	50S ribosomal protein L14	PROT
br34	A_92_P002189	trans	-0.31	Ribosomal protein S24 S35	PROT
br48	A_92_P006377	trans	-0.27	Ribosomal protein L17e family protein	PROT
br40	A_92_P008077	trans	-0.29	Transferring glycosyl groups / F-box domain containing	PROT
br43	A_92_P017437	trans	-0.28	Protein MRP homolog	PROT
br46	A_92_P017150	trans	-0.27	Clp-like energy-dependent protease	PROT
br1	A_92_P011013	trans	-0.55	Chlorophyll synthase / Prenyltransferase	PS
br9	A_92_P006916	trans	-0.39	Translocase subunit chloroplastic-like SECA2	PS
br18	A_92_P016376	trans	-0.33	Glycolate oxidase	PS
br24	A_92_P008576	trans	-0.33	Phototropin 1	PS
br21	A_92_P024571	trans	-0.33	NADP-dependent glyceraldehyde-3-phosphate dehydrogenase	PS
br49	A_92_P005648	trans	-0.27	Rossmann-fold NAD(P)-binding domain-containing protein	PS
br10	A_92_P006669	trans	-0.39	Transcription factor bZIP63	RNA
br27	A_92_P008762	trans	-0.32	RNA recognition motif-containing protein	RNA
br30	A_92_P011285	trans	-0.31	RNA recognition motif containing protein	RNA
br38	A_92_P039278	trans	-0.30	Nuclear transcription factor Y subunit A-10	RNA
br41	A_92_P005627	trans	-0.28	Splicing factor 4-like protein / Gamma response I protein	RNA
br11	A_92_P005246	trans	-0.37	Sodium/calcium exchanger protein	TR
br15	A_92_P006606	trans	-0.34	Plasma membrane H+-ATPase	TR
br47	A_92_P014444	trans	-0.27	Outward-rectifying potassium channel	TR
br3	A_92_P010785	trans	-0.50	Callose synthase 3-like / Glucan synthase-like 12	CHO
br39	A_92_P009042	trans	-0.29	Adenine nucleotide translocator / ADP/ATP carrier 2	CHO
br6	A_92_P017985	trans	-0.44	Lipid phosphate phosphatase 3	LM
br45	A_92_P014105	trans	-0.28	Dentin sialophospho protein	LM
br8	A_92_P003011	trans	-0.42	Histone H1	DNA
br31	A_92_P006252	trans	-0.31	MIP18 family protein	DNA
br17	A_92_P014428	trans	-0.34	Resistance to phytophthora 1 protein	BS
br33	A_92_P005298	trans	-0.31	AP2-domain DRE binding factor DBF1	AS
br19	A_92_P011464	trans	-0.33	BTB/POZ domain-containing protein	CELL
br14	A_92_P011373	trans	-0.34	DDT domain-containing protein	AAM
br20	A_92_P002406	trans	-0.33	Metal tolerance protein	MH
br26	A_92_P034039	trans	-0.32	CER1 / WAX2	SM
br36	A_92_P000561	trans	-0.30	Nucleotide pyrophosphatase phosphodiesterase	NM
br35	A_92_P015107	trans	-0.30	Ethylene receptor	HM
br44	A_92_P011520	trans	-0.28	NADP-dependant malate dehydrogenase	TCA
br4	A_92_P009740	trans	-0.48	PLAC8 family protein	NA
br2	A_92_P011630	trans	-0.54	Protein of unknown function	PUF
br25	A_92_P015277	trans	-0.32	Protein of unknown function	PUF
br37	A_92_P003770	trans	-0.30	Protein of unknown function	PUF
br5	A_92_P010266	trans	-0.47	Putative protein	PP

Table 4.14: The node annotations for the QTL 9-7 regulatory network model for genes associated with GLS resistance (Figure 4.14 on page 205). For a description of the categories in the "MM (MapMan) bin" column refer to Table 4.7.

Symbol	Agilent ID	Cis/ Trans	GLS cor	Annotation (AT, Rice, B2G)	MM bin
PUF1	A_92_P001310	cis	-0.37	Protein of unknown function	PUF
DQS	A_92_P009686	cis	-0.35	3-dehydroquinate synthase	AAM
GTP	A_92_P014787	cis	-0.33	GTP-binding protein	SIGN
PUF2	A_92_P011516	cis	-0.30	Protein of unknown function	PUF
cr5	A_92_P019715	trans	-0.36	Iron ABC superfamily transporter	TR
cr8	A_92_P006825	trans	-0.35	ABC transporter family protein	TR
cr18	A_92_P009675	trans	-0.28	Protein-export membrane protein	TR
cr22	A_92_P012143	trans	-0.26	Adaptor protein complex AP1, gamma subunit	TR
cr7	A_92_P001231	trans	-0.35	Calcium-binding protein	SIGN
cr10	A_92_P010477	trans	-0.31	Serine/threonine protein phosphatase	SIGN
cr12	A_92_P009517	trans	-0.31	Adenylate cyclase	SIGN
cr2	A_92_P017490	trans	-0.40	Peptidyl-prolyl *cis-trans* isomerase	AAM
cr21	A_92_P015567	trans	-0.27	Membrane-anchored ubiquitin-fold protein	AAM
cr17	A_92_P011451	trans	-0.29	Thioredoxin chloroplastic-like	RX
cr19	A_92_P001302	trans	-0.28	Metallo-hydrolase/oxidoreductase superfamily protein	RX
cr11	A_92_P011285	trans	-0.31	RNA recognition motif containing protein	RNA
cr20	A_92_P003738	trans	-0.27	TOPLESS-related protein / WD repeat-containing	RNA
cr3	A_92_P020271	trans	-0.37	Histone H1 / winged-helix DNA-binding transcription factor	DNA
cr4	A_92_P007992	trans	-0.36	ATP synthase mitochondrial F1 complex assembly factor 1	PROT
cr6	A_92_P004942	trans	-0.36	Phosphatidylinositolglycan-related protein	CELL
cr14	A_92_P005298	trans	-0.31	AP2-domain DRE binding factor DBF1	AS
cr1	A_92_P009740	trans	-0.48	PLAC8 family protein	NA
cr16	A_92_P007798	trans	-0.31	Cupin, RmlC-type	NA
cr9	A_92_P015277	trans	-0.32	Protein of unknown function	PUF
cr13	A_92_P003861	trans	-0.31	Protein of unknown function	PUF
cr15	A_92_P008549	trans	-0.31	Protein of unknown function	PUF

Table 4.15: The node annotations for the QTL 10-10 regulatory network model for genes associated with GLS resistance (Figure 4.15 on page 206). For a description of the categories in the "MM (MapMan) bin" column refer to Table 4.7.

Symbol	Agilent ID	Cis/ Trans	GLS cor	Annotation (AT, Rice, B2G)	MM bin
LAR	A_92_P009023	cis	-0.50	Leucoanthocyanidin reductase	SM
GTR	A_92_P006618	cis	-0.50	Glutamyl-tRNA reductase	PS
PUF	A_92_P001146	cis	-0.30	Protein of unknown function	PUF
dr4	A_92_P025469	trans	-0.38	ABC transporter G family member 22-like	TR
dr11	A_92_P017437	trans	-0.28	Multidrug resistance (MRP)-type ATP binding protein	TR
dr9	A_92_P008221	trans	-0.34	F-box protein	PROT
dr13	A_92_P015985	trans	-0.27	F-box and tubby domain containing protein	PROT
dr1	A_92_P017985	trans	-0.44	Lipid phosphate phosphatase 3	LM
dr2	A_92_P012575	trans	-0.44	Cysteine-rich receptor-like protein kinase 25-like	SIGN
dr3	A_92_P018634	trans	-0.43	Splicing factor U2af 38 kDa subunit	RNA
dr6	A_92_P007674	trans	-0.37	Protein acclimation of photosynthesis to environment	PS
dr7	A_92_P011271	trans	-0.35	DNA mismatch repair protein 2	DNA
dr10	A_92_P039824	trans	-0.29	Cytochrome P450	SM
dr12	A_92_P018107	trans	-0.28	Cytosolic phosphoglucomutase	CHO
dr14	A_92_P016094	trans	-0.26	Glutamine amidotransferase subunit	AAM
dr5	A_92_P020223	trans	-0.37	Protein of unknown function	PUF
dr8	A_92_P013500	trans	-0.35	Protein of unknown function	PUF

Chapter 5

Meta-analysis: Combining co-expression network analysis with QTL/eQTL overlap analysis

5.1 Introduction

A systems genetics strategy, outlined in Figure 5.1, was developed to incorporate the analysis of gene co-expression with phenotypic QTL and eQTL mapping, to identify candidate genes and pathways associated with GLS disease.

As an initial analysis in Chapter 3, gene expression profiles across the individuals (*C. zeina*-infected maize plants) of the CML444×SC Malawi RIL population was used in a correlation analysis to identify 42 gene co-expression modules (Figure 5.1 Box A). The GLS severity measurements (phenotype data) across the same individuals were incorporated to identify eight modules significantly correlated to GLS severity (an output of Chapter 3). Advantages of this analysis were that central nodes of the relevant gene co-expression modules could be identified as potential global regulators, or drivers, and functional enrichment analyses could be used to potentially link biological processes to the GLS disease response. However, the analysis provided no reference to the genetic basis for either gene expression variation or the response to GLS disease.

Subsequently, a global eQTL analysis was performed to identify associations between genotype and gene expression (Figure 5.1 Box B). The expression profile of each gene, across the individuals in the above-mentioned population, was treated as a quantitative

trait to identify regions of the genome where genetic variation was associated with gene expression variation among the RIL lines. In Chapter 4 the 31,549 identified eQTLs were classified as *cis*- or *trans*-acting and 32 *trans*-eQTL hotspots were identified. A biologically meaningful *trans*-eQTL hotspot was assumed to reveal a significant number of response genes under common transcriptional control and acting in the same pathway. Genes that belong to such a hotspot are typically regulated by one or more regulatory gene(s) located in the eQTL hotspot, which in turn may have a *cis*-eQTL at the hotspot locus. However, it is also possible that the cis-polymorphism is based on DNA variation in coding regions leading to a change in protein sequence or structure rather than gene expression variation. Even though *C. zeina*-infected maize plants were used for expression profiling, no conclusions regarding the genetic basis for the response to GLS disease could be inferred from the eQTL analysis results alone.

In order to unravel the genetic basis for the quantitative disease response to GLS, a QTL analysis for GLS severity was performed in Chapter 4 on the same field trial that was sampled for the expression study (Figure 5.1 Box C). Eight regions of the maize genome were identified, where genetic variation was significantly associated with phenotypic variation. However, the identified regions spanned broad genetic intervals including hundreds of genes, any of which might contain a polymorphism affecting the phenotype.

A QTL/eQTL overlap analysis was performed in Chapter 4, including only genes with eQTLs that belonged to *trans*-eQTL hotspots overlapping GLS severity QTLs (Figure 5.1 Box D). A hotspot may be significant if it points to a polymorphism in a gene that affects the expression of many related genes which, in turn, affects GLS severity and therefore underlies a GLS severity QTL. QTL-overlapping *trans*-eQTL hotspots in this study were on average 60% smaller than GLS severity QTLs themselves and only genes with eQTL peaks in these hotspots were included in the analysis. Two data management steps were performed on the identified *cis*- and *trans*-eQTL candidate genes: (i) candidates were split based on whether higher expression was associated with the resistance or susceptibility linked allele; and (ii) correlation analysis between phenotype values (GLS severity scores) and the gene expression values were used to filter the resulting gene lists. This QTL/eQTL overlap analysis assumed that: (i) there is an underlying DNA polymorphism that gives rise to a change in gene expression which in turn affects GLS severity; (ii) the allele

associated with a higher expression was linked to the trait of interest (i.e. GLS resistance or susceptibility, depending on the underlying polymorphism); (iii) the causal gene's expression profile correlates with the quantitative trait of interest. Additional correlation analyses between the expression values of each identified gene with a *cis*-eQTL and all the genes with *trans*-eQTLs in a linked *trans*-eQTL hotspot, were used to construct putative regulatory network models per GLS severity QTL with an overlapping *trans*-eQTL hotspot. Thus gene expression data, genotype data and phenotype data were integrated to construct regulatory network models (Figures 4.9-4.15). A limitation of this analysis was that information relating to the gene co-expression modules (see Figure 5.1 Box A) was not incorporated.

One way to incorporate the identified gene co-expression modules would be to ask whether it was possible to identify a genetic basis for the observed coordinated expression responses to GLS disease, i.e. if there was significant concurrence between the genes in co-expression modules correlating with GLS disease (Table 3.1) and the genes in *trans*-eQTL hotspots that overlapped the GLS severity QTLs (Table 4.5). This question will be further explored in the this chapter.

5.2 Methods

Fisher's exact tests were used to determine whether each co-expression module was enriched for genes with eQTLs in a common *trans*-eQTL hotspot (Figure 5.1 Box E), using a customised script in R (available in the electronic Appendix). Thus for a specific co-expression module, the proportion of genes with eQTLs in a given *trans*-eQTL hotspot was compared to the proportion of genes included in the analysis with eQTLs in that hotspot. Therefore, for each of the 42 co-expression modules, 64 tests were performed (32 hotspots were split according to the parental allele associated with higher expression). The p-values were adjusted for multiple testing by controlling the false discovery rate (Benjamini and Hochberg, 1995). The resulting p-values per co-expression module are given in the electronic Appendix. Note that the *trans*-eQTL hotspots were named such that "HS 10-10 R" referred to the *trans*-eQTL hotspot (on chromosome 10, starting at marker 10) overlapping a GLS severity QTL where the resistance associated allele (that was determined by the overlapping GLS severity QTL; Table 4.1) was associated with a

higher expression.

As an example, the proportion of reporters in the greenyellow module with eQTLs in HS 9-6 S [39 reporters in the greenyellow module with eQTLs in HS 9-6 S (Tables 5.1 and 5.2) out of 185 reporters in the greenyellow module (Table 3.1), thus 39/185 = 0.21, i.e. 21%] was compared to the proportion of total reporters that was included in the co-expression module analysis with eQTLs in HS 9-6 S [296 reporters with eQTLs in HS 9-6 S (Table 4.3; note that 296 out of the 311 were included in the co-expression module analysis) out of 19,281 reporters in the co-expression module analysis (see section 3.4.2), thus 296/19,281=0.015, i.e. 1.5%]. The p-value calculated using the Fisher's exact test was $3.5e^{-33}$ (adjusted p-value = $1.4e^{-30}$), which indicated that genes with eQTLs in HS 9-6 S were significantly overrepresented in the greenyellow module.

MapMan was used to functionally classify genes into predefined bins (Thimm *et al.*, 2004). This classification, together with manual revision, were used to construct Figures 5.2, 5.3 and 5.4, summarising the functional annotations of the genes per co-expression module enriched for genes with eQTLs in QTL-overlapping *trans*-eQTL hotspots.

The first module in the eQTL data analysis pipeline (Figure 4.2) was used to map "*a priori* network eQTLs" as defined by Hansen *et al.* (2008). The module eigengene expression profiles of the eight gene co-expression modules correlating with GLS severity (Table 3.1), were extracted using functions from the weighted gene co-expression network analysis (WGCNA) package in R. An "e-traits file", consisting of the eight module eigengene expression profiles, together with the "map file" and "cross file" that were previously used for eQTL mapping, were used as input files in Galaxy. QTL mapping was performed using the parameters mentioned previously for eQTL mapping, i.e. forward and backward stepwise regression (p-value = 0.1), a 2 cM walking speed and CIM was implemented. The LR threshold was set to 11.5. A customised R script was used to visualise the results (Figure 5.5).

5.3 Results and Discussion

5.3.1 Over-representation analysis

Out of five co-expression modules with a significant correlation to GLS susceptibility (greenyellow, paleturquoise, blue, yellowgreen and magenta), an enrichment analysis re-

vealed that only the greenyellow module was significantly enriched for genes with eQTLs in *trans*-eQTL hotspots that overlapped with GLS severity QTLs (Figure 5.1 Box F). In total, the greenyellow module was enriched for genes with eQTLs in five hotspots, of which two overlapped GLS severity phenotypic QTLs 9-5 and 10-10; and in both cases the hotspot's parental allele associated with increased expression matched the GLS severity QTL parental allele associated with susceptibility. Tables 5.1 and 5.2 give the 58 gene models in the greenyellow module with eQTLs either in HS 9-6 S or in HS 10-10 S (or with eQTLs in both hotspots) and Figure 5.2 provides an overview of the functional categories of these genes. All 58 gene models (100%) had significant individual correlation coefficients (p-value <0.01) when gene expression values were correlated with the GLS severity scores. Since the module eigengene of the greenyellow co-expression module was highly correlated (with a correlation coefficient of 0.71) with GLS severity, it was not unexpected that the genes with correlated expression values in this co-expression module were also correlated to GLS severity. Out of all the *trans*-eQTLs in HS 9-6 S and HS 10-10 S, 12.6% and 22.3% respectively, belonged to genes in the greenyellow module.

Furthermore, out of the three co-expression modules with a significant correlation to GLS resistance (turquoise, darkred and yellow), the turquoise and yellow modules were significantly enriched for genes with eQTLs in *trans*-eQTL hotspots that overlapped the GLS severity QTLs (Figure 5.1 Box F). The turquoise module was enriched for genes with eQTLs in nine hotspots, of which two overlapped GLS severity phenotypic QTLs 9-5 and 9-7; and in both cases the hotspot's parental allele with increased expression matched the GLS severity QTL parental allele associated with resistance. Tables 5.3, 5.4 and 5.5 give the 74 gene models in the turquoise module with eQTLs either in HS 9-6 R or in HS 9-7 R (or with eQTLs in both hotspots) and Figure 5.3 provides an overview of the functional categories of these genes. Thirty out of the 74 gene models (41%) had significant individual correlation coefficients (p-value <0.01) when gene expression values were correlated with the GLS severity scores. The yellow module was enriched for genes with eQTLs in four hotspots, of which one overlapped GLS disease QTL 4-11; and this hotspot's parental allele with increased expression also matched the GLS severity QTL parental allele associated with resistance. Tables 5.6 and 5.7 give the 41 gene models in the yellow module with eQTLs in HS 4-12 R and Figure 5.4 provides an overview of the functional categories of these genes. Fourteen out of the 41 gene models (34%) had

significant individual correlation coefficients (p-value <0.01) when gene expression values were correlated with the GLS severity scores.

5.3.2 Network eQTL analysis

An *a priori* network eQTL analysis (according to terminology used by Hansen *et al.*, 2008) was performed where the module eigengene profiles of the gene co-expression modules were used as the traits in a QTL analysis. The aim was to identify genomic regions where genetic variation influences entire co-expression modules. For the greenyellow module eigengene, three network eQTLs were mapped, of which two overlapped GLS severity phenotypic QTLs 9-5 and 10-10 (Figure 5.5). In both cases, the module eigengene network eQTL's parental allele with increased expression matched the GLS severity QTL parental allele associated with susceptibility (the SC Malawi allele was associated with susceptibility for QTL 9-5 and QTL 10-10). Importantly, this supported the above-mentioned result (Figure 5.1 Box F) that the genes in the greenyellow module had a significant number of eQTLs in the *trans*-eQTL hotspots that overlapped GLS severity phenotypic QTLs 9-5 and 10-10. For the turquoise module eigengene, only one network eQTL was identified on chromosome 7, not overlapping a GLS severity QTL (Figure 5.5). However, not surprisingly it overlapped one of the *trans*-eQTL hotspots for which the turquoise module was strongly enriched, i.e. the hotspot with the lowest Fisher's exact test p-value for the turquoise module (HS 7-6 SC with an adjusted p-value of $4.8e^{-49}$; data available in electronic Appendix). For the yellow module eigengene, one network eQTL was identified on chromosome 4 and it overlapped GLS severity QTL 4-11 (Figure 5.5). Here, the module eigengene network eQTL's parental allele with increased expression matched the GLS severity QTL parental allele associated with resistance (the CML444 allele was associated with resistance for QTL 4-11). This finding also supported the above-mentioned result (Figure 5.1 Box F) that the genes in the yellow module had a significant number of eQTLs in the *trans*-eQTL hotspot that overlapped GLS severity phenotypic QTL 4-11.

Therefore, three of the gene co-expression modules that were previously linked to GLS disease were identified to have module eigengene network eQTLs overlapping GLS severity QTLs. Furthermore, depending on whether the co-expression module was positively associated with either resistance or susceptibility, the module eigengene network

eQTL's parental allele with increased expression matched the overlapping GLS severity QTL parental allele associated with the same phenotype (either resistance or susceptibility). Also, in most cases, the genes in a co-expression module had a significant number of eQTLs that belonged to a *trans*-eQTL hotspot coinciding with its module eigengene network eQTL. In such cases, the module eigengene network eQTL's parental allele associated with higher expression matched that of the coinciding *trans*-eQTL hotspot (which was split by parental allele associated with higher expression).

5.3.3 Final hypotheses regarding genes and processes underlying the GLS disease response in maize

In this study, loci that affected GLS severity were identified and consequently several mechanisms of defense were hypothesised. Due to the complexity of disease resistance mechanisms, the plant molecular mechanisms that control quantitative disease resistance (or susceptibility) are generally poorly understood. Since quantitative disease resistance is usually conditioned by many loci with small effects, no single hypothesis can fully explain the scope of quantitative disease resistance (Poland *et al.*, 2009). It is anticipated that each of the eight GLS severity QTLs that were identified in this population offers a different mechanism of defense.

The RIL population was sampled during flowering when GLS lesions were evident and therefore sampling was more suited to measure the plant's responses to the spreading disease within the leaf (i.e. susceptibility). The greenyellow co-expression module (consisting of 185 genes) had an exceptionally strong correlation with GLS susceptibility (correlation coefficient of 0.71; Table 3.1). The expression of the bulk of genes in this co-expression module seemed to be explained by the two *trans*-eQTL hotspot loci, HS 9-6 S and HS 10-10 S, coinciding with the GLS severity phenotypic QTLs 9-5 and 10-10. These genes are co-expressed, they belong to the greenyellow module which is associated with GLS susceptibility and their expression is explained by the two loci (HS 9-6 S and HS 10-10 S) (Figure 5.2, together with Tables 5.1 and 5.2, provides a summary of these candidate genes and their functional categories). Several final hypotheses of potential mechanisms underlying quantitative disease susceptibility, for GLS severity QTLs 9-5 and 10-10, are given below.

The first group of hypotheses for mechanisms or genes involved in quantitative disease

susceptibility was associated with defense signal transduction. It is hypothesised that a calmodulin-related calcium sensor gene with a *cis*-eQTL that coincided with HS 9-6 S (Table 5.2) encodes a protein that acts as a transcriptional regulator to activate numerous target proteins through signalling due to the ion fluxes across membranes in response to pathogen infection. It is hypothesised that calcium signalling plays a role in the regulation of induced defense-related signalling cascades and plant adaptation to fungal attack. Since the expression of these genes were mainly explained by HS 9-6 S (Table 5.2), this could be a mechanism conferred by the overlapping GLS severity QTL 9-5. It is further hypothesised that the higher expression in susceptible plants of two WRKY transcription factors with *trans*-eQTLs in HS 9-6 S and HS 10-10 S (Tables 5.1 and 5.2), respectively, aid in the regulation of host responses in reaction to pathogen challenge typically via hormone signalling (Pandey and Somssich, 2009). Furthermore, it is hypothesised that ethylene (Aminocyclopropane-1-carboxylic acid synthase in Table 5.1 with a *trans*-eQTL in HS 9-6 S), which is an important factor for the induction of defense responses against pathogen attack, promoted necrotic lesion formation in susceptible plants.

A second group of hypotheses for mechanisms or genes involved in quantitative disease susceptibility was associated with chemical warfare. It is hypothesised that increased antifungal activity was needed to act against the toxins produced by *C. zeina* in susceptible plants. This activity resulted from genes involved in secondary metabolism with *trans*-eQTLs in HS 9-6 S (Table 5.1). Similarly, it is hypothesised that genes involved in cellular detoxification of xenobiotic conjugates were activated in response to *C. zeina* infection. This mechanism linked to GLS severity QTL 10-10, where the *trans*-eQTLs of a few adenosine triphosphate (ATP)-binding cassette (ABC) transporters as well as a glutathione S-transferase were localised in HS 10-10 S (Table 5.2).

Finally, there was a strong signature of pathogenesis-related (PR) and biotic stress response genes with *trans*-eQTLs mainly coinciding with QTL 9-5 (Table 5.2). These genes, which appeared to be associated with GLS susceptibility, included two proteinase inhibitors, a PR beta-1,3-glucanase, two chitinases and a GRAM-domain containing protein. The mentioned genes were either activated too late after infection started or its activity was overcome or suppressed by *C. zeina* effectors. It is hypothesised that *C. zeina* could be protected against plant chitinases due to chitin-binding effectors for example Avr4 or Ecp6 (extracellular protein 6) (van den Burg *et al.*, 2006; de Jonge, 2010).

In contrast to the susceptible response, gene expression in resistant plants were expected to portray constitutive defenses that were still present at the late stage of sampling, rather than early induced defenses in response to *C. zeina*. The turquoise co-expression module (a large module, consisting of 1,564 genes) was the module with the strongest correlation with GLS resistance. The expression of a group of genes in this co-expression module seemed to be explained by the two *trans*-eQTL hotspot loci, HS 9-6 R and HS 9-7 R, coinciding with the GLS severity phenotypic QTLs 9-5 and 9-7. These genes are co-expressed, they belong to the turquoise module which is associated with GLS resistance and their expression is explained by the two loci (HS 9-6 R and HS 9-7 R) (Figure 5.3, together with Tables 5.3, 5.4 and 5.5, provides a summary of these candidate genes and their functional categories). Several final hypotheses of potential mechanisms underlying quantitative disease resistance, for GLS severity QTLs 9-5 and 9-7, are given below.

The main hypothesis for quantitative disease resistance relate to defense signal transduction. It is hypothesised that polymorphisms at QTLs 9-5 and 9-7 could differentially regulate the expression of phosphatases/kinases that are involved in post-translational modifications, which promote defense mechanisms that threaten the survival of fungal cells. Interestingly, out of the nine phosphatases/kinases that were part of the above-mentioned list of candidate genes (Table 5.4), only one (A_92_P010477) was included in the previously constructed QTL 9-5 and 9-7 regulatory network models (Tables 4.13 and 4.14). Importantly, signalling/post-translational modification was also identified as a main hypothesis for the QTL 9-5 regulatory network model, however most of the contributing genes were not part of the turquoise module. Therefore, the eight remaining genes in the turquoise module with this same functionality (that were previously removed through filtering), confirmed signalling/post-translational modification as an important potential mechanism of resistance conferred by GLS severity QTLs 9-5 and 9-7.

Other hypotheses for quantitative disease resistance, linked to QTLs 9-5 and 9-7, relate to genes or mechanisms involved in basal defense. It is hypothesised that the two receptor-like kinases (Table 5.4), with *trans*-eQTLs in HS 9-6 R, function as cell surface receptors and play a crucial role in perception of fungal pathogen-associated molecular patterns (PAMPs) and defense signal transduction. In addition, it is hypothesised that the gene encoding a TOPLESS-related protein containing WD repeats (involved in regulation of transcription), with a strong correlation to GLS resistance and with a *trans*-eQTL

in HS 9-7 R (Table 5.4), could act as a repressor of negative regulators of the plant immune response. Such negative regulators of the plant immune response will typically strongly correlate with GLS susceptibility and can potentially be activated due to *C. zeina* manipulated plant gene expression. An example of a potential negative regulator associated with GLS susceptibiltiy from this study, was the NUDIX domain-containing protein (Table 5.1) with a *cis*-eQTL in HS 4-11 S and a *trans*-eQTL in HS 9-6 S. Finally, it is hypothesised that the gene encoding a callose synthase (with a *trans*-eQTL in HS 9-6 R and a strong gene expression correlation to GLS resistance; Table 5.3), plays a role in GLS resistance via depositions of callose in the form of local cell wall thickenings to block fungal penetration (similar to the *Arabidopsis* callose synthase studied by Jacobs *et al.*, 2003).

The yellow co-expression module (another large module, consisting of 1,170 genes) was also associated with GLS resistance. The expression of a group of genes in this co-expression module was explained by the *trans*-eQTL hotspot locus, HS 4-12 R, coinciding with the GLS severity QTL 4-11. These genes are co-expressed, they belong to the yellow module which is associated with GLS resistance and their expression is explained by a single locus (HS 4-12 R) (Figure 5.4, together with Tables 5.6 and 5.7, provides a summary of these candidate genes and their functional categories). Several final hypotheses of potential mechanisms underlying quantitative disease resistance, for GLS severity QTL 4-11, are given below.

It is hypothesised that the ubiquitination system and the 26S proteasome play a key role in conferring resistance against *C. zeina* (Table 5.6), possibly via the regulation of processes such as the oxidative burst, hormone signalling, gene induction, and programmed cell death (Trujillo and Shirasu, 2010). Seven genes in the above-mentioned list of candidate genes encoded proteins involved in ubiquitin-mediated protein degradation, of which two were previously also included in the QTL 4-11 regulatory network (Table 4.12). It is further hypothesised that maize heterotrimeric G-proteins successfully control defense responses to *C. zeina* (Table 5.6).

Finally, it is hypothesised that R-gene-mediated defense could potentially account for a proportion of the quantitative disease resistance against the necrotrophic fungus *C. zeina*. An R-gene encoding a NBS-LRR resistance protein (Table 5.7), located on chromosome 3, but with a *trans*-eQTL in HS 4-12 R, was part of this group of candidate

genes. Poland *et al.* (2009) noted that although it is commonly believed that R-genes confer complete race-specific resistance and QTLs confer partial race non-specific resistance, there are examples of R-genes that condition partial resistance. This hypothesis would be much stronger if there was an R-gene with a *cis*-eQTL coinciding with a GLS severity QTL.

5.4 Conclusion

This chapter concluded the systems genetics strategy which was developed to incorporate the analysis of gene co-expression with phenotypic QTL and eQTL mapping, in order to identify candidate genes and biological processes associated with GLS disease in maize. Three of the co-expression modules correlating with GLS disease were enriched for genes with eQTLs in *trans*-eQTL hotspots that overlapped the GLS severity phenotypic QTLs. Module eigengene network eQTLs were identified to verify these findings. It was possible, at least in some cases, to identify a genetic basis for the coordinated expression responses to GLS disease. In two cases, the genes in a co-expression module correlating with GLS severity were identified to have a significant number of *trans*-eQTLs in two different QTL-overlapping hotspots. This indicates that more than one regulatory gene likely participates in explaining the expression variation in the same endpoint genes/mechanisms.

Hypotheses regarding genes and processes associated with GLS resistance/susceptibility were concluded, firstly from the co-expression modules correlating with GLS disease in Chapter 3 and secondly from the *trans*-eQTL hotspots that overlapped the GLS severity QTLs in Chapter 4. The approach outlined in this chapter brought the above-mentioned separate analyses together by studying the overlap between the co-expression modules correlating with GLS disease and the *trans*-eQTL hotspots that overlapped the GLS severity QTLs. Thus, this meta-analysis highlighted the most likely defense mechanisms utilised by maize plants in this RIL population against the fungus *Cercospora zein*a.

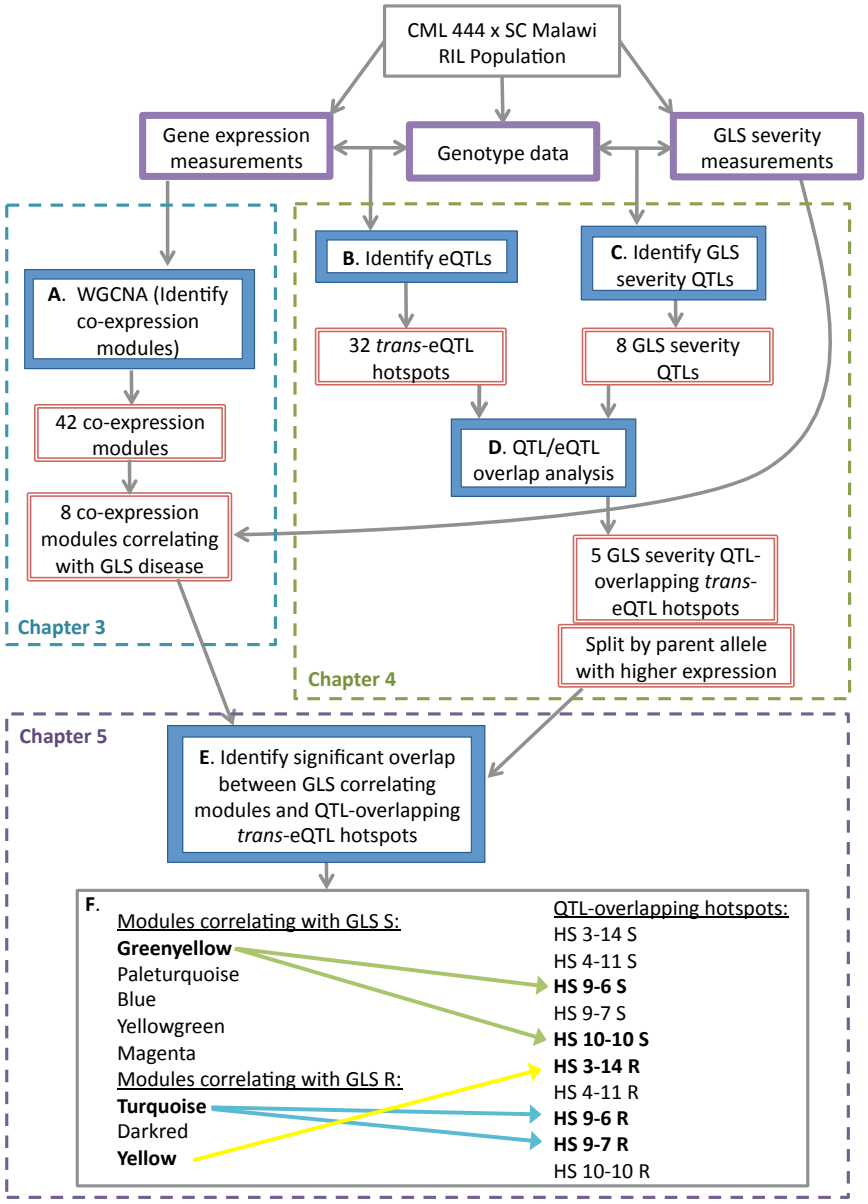


Figure 5.1: Overview of the strategy that was followed in this study. Gene expression data, genotype data and GLS severity data from a CML444×SC Malawi maize RIL population were used in a systems genetics approach to dissect quantitative disease resistance and susceptibility of GLS disease. These data types were used to identify gene co-expression modules (Box A), expression QTLs (Box B) and GLS severity QTLs (Box C). To prioritise genes and pathways associated with GLS disease, co-expression modules correlating with GLS disease were identified in Chapter 3 and eQTLs overlapping with GLS severity QTLs (Box D) in Chapter 4. Finally, in Chapter 5, candidate genes in the GLS-correlating co-expression modules and in the QTL-overlapping *trans*-eQTL hotspots were compared (Box E) to determine the genetic basis of the co-expression modules. Box F gives a summary of these results.

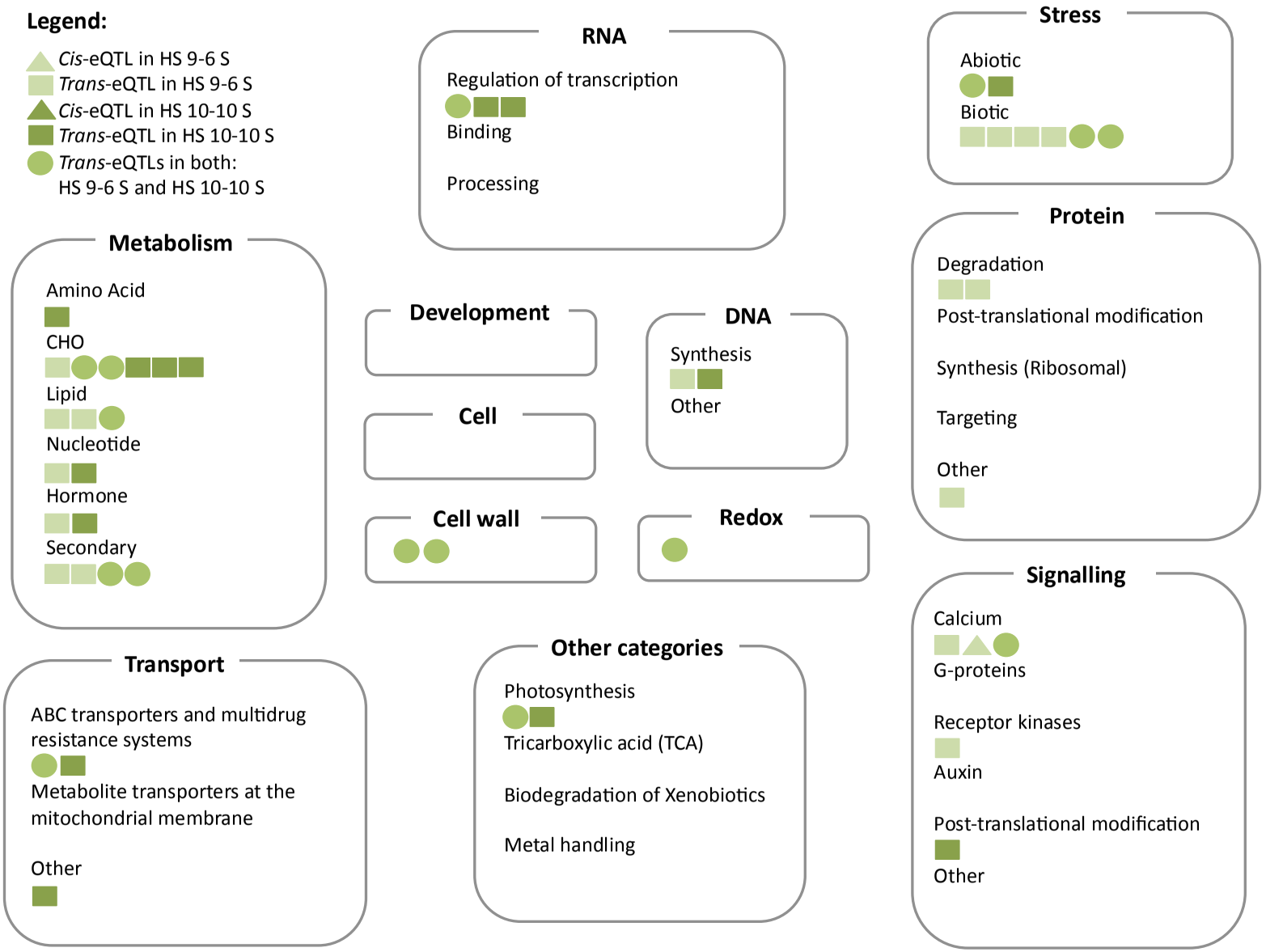


Figure 5.2: An overview of the functional categories of the genes in the greenyellow module with eQTLs in the HS 9-6 S and HS 10-10 S. MapMan BINs were used as a basis to group the genes into categories. Each block represent a maize gene model.

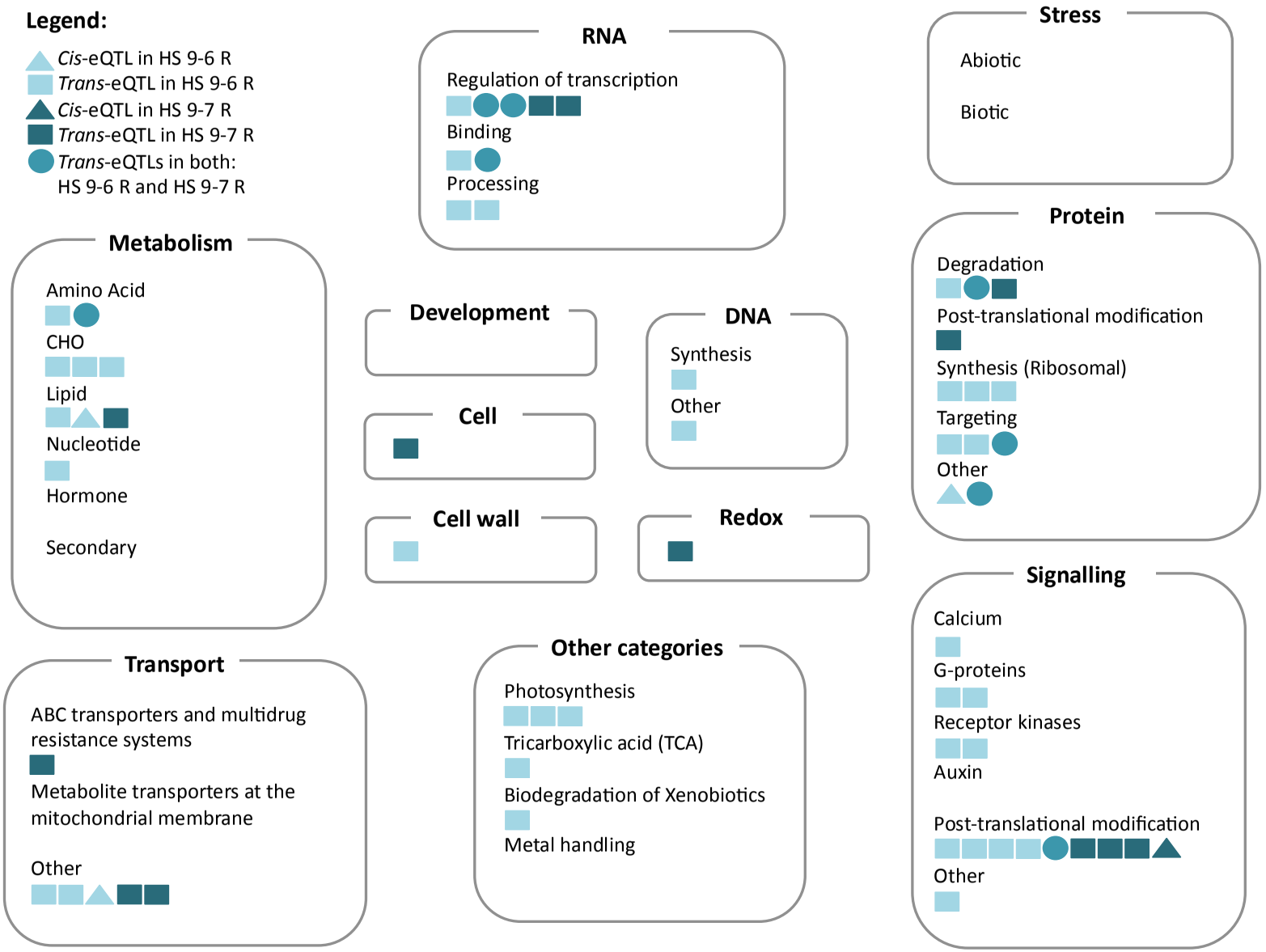


Figure 5.3: An overview of the functional categories of the genes in the turquoise module with eQTLs in the HS 9-6 R and HS 9-7 R. MapMan BINs were used as a basis to group the genes into categories. Each block represent a maize gene model.

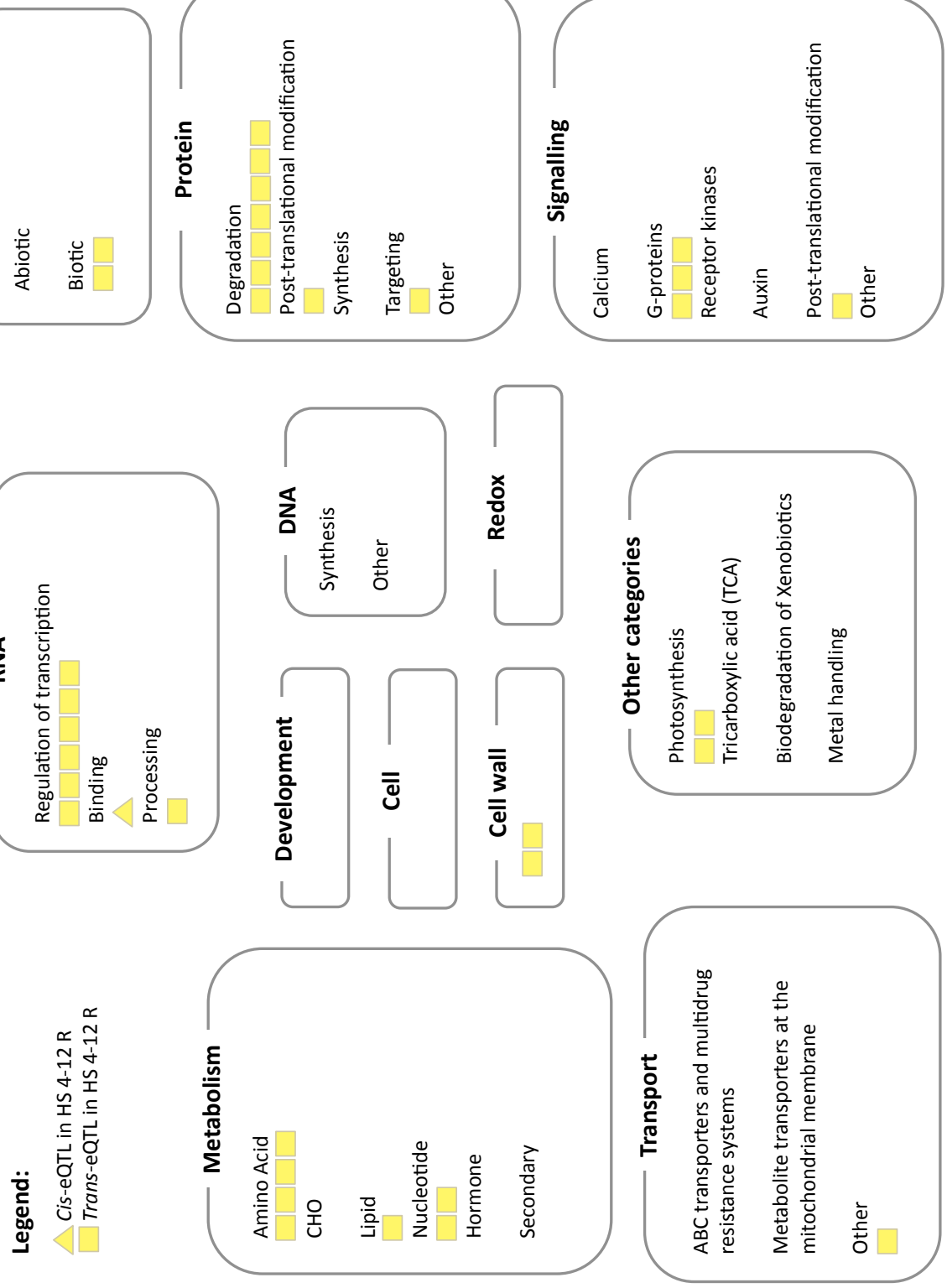


Figure 5.4: An overview of the functional categories of the genes in the yellow module with eQTLs in the HS 4-12 R. MapMan BINs were used as a basis to group the genes into categories. Each block represent a maize gene model.

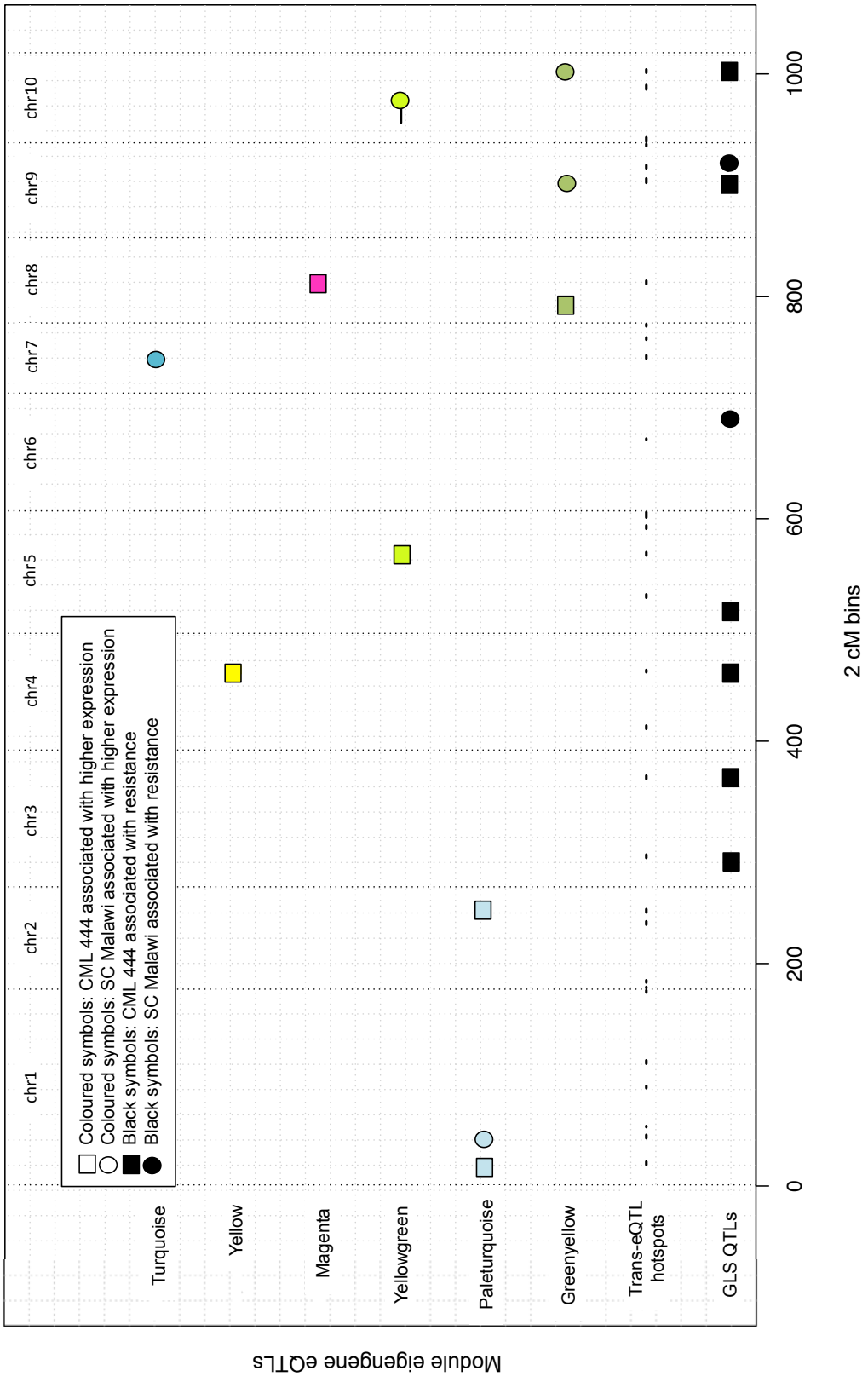


Figure 5.5: "*A priori*" network analysis (according to terminology used by Hansen *et al.*, 2008). The module eigengene profiles of the gene co-expression modules identified in Chapter 3, were used as the traits in a QTL analysis. Coloured blocks indicate module eQTLs where CML444 was the parental allele associated with higher expression and coloured circles indicate module eQTLs where SC Malawi was the parental allele associated with higher expression. Black blocks indicate GLS severity QTLs where CML444 was the parental allele associated with resistance and black circles indicate GLS severity QTLs where SC Malawi was the parental allele associated with resistance. The *trans*-eQTL hotspot regions (that were identified in Chapter 4) are also indicated.

Table 5.1: Gene models in the greenyellow co-expression module with eQTLs in HS 9-6 S and HS 10-10 S (part 1 of 2).

Agilent ID	Gene ID	Gene core bin[a]	GLS severity cor[b]		*Trans*-eQTL overlap[c]	Annotation (AT, Rice, B2G)[d]	Functional category[e]
A_92_P021132	GRMZM2G053206	2.09	0.52	*	HS 9-6 S + HS 10-10 S	Xylanase inhibitor	Cell wall
A_92_P017529	GRMZM2G058472	6.05	0.49	*	HS 9-6 S + HS 10-10 S	Secondary cell wall-related glycosyltransferase family 8	Cell wall - hemicellulose synthesis
A_92_P016791	GRMZM2G017337	3.05	0.46	*	HS 9-6 S	Chromatin assembly factor-1	DNA - synthesis/chromatin structure
A_92_P030660	GRMZM2G032977	2.02	0.53	*	HS 10-10 S	Nuclease I	DNA - synthesis/chromatin structure
A_92_P010667	GRMZM2G153536	1.03	0.48	*	HS 10-10 S	Branched-chain-amino-acid aminotransferase chloroplast	Metabolism - amino acid - synthesis
A_92_P035928	GRMZM2G033905	3.04	0.62	*	HS 9-6 S	Dihydrolipoyllysine-residue acetyltransferase	Metabolism - CHO
A_92_P022183	GRMZM2G173192	5.03	0.53	*	HS 9-6 S + HS 10-10 S	Lactate dehydrogenase	Metabolism - CHO
A_92_P013322	GRMZM2G131055	8.03	0.50	*	HS 10-10 S	Glycosyltransferase family 61	Metabolism - CHO
A_92_P018607	GRMZM2G086845	3.06	0.50	*	HS 10-10 S	Fructokinase 1	Metabolism - CHO - degradation
A_92_P032638	GRMZM2G051806	6.05	0.49	*	HS 10-10 S	Hexokinase 2	Metabolism - CHO - degradation
A_92_P039018	GRMZM2G164405	2.03	0.46	*	HS 9-6 S	ACC synthase	Metabolism - hormone - ethylene
A_92_P018154	GRMZM2G041699	4.08	0.43	*	HS 9-6 S + HS 10-10 S	Cytokinin-o-glucosyltransferase 2	Metabolism - hormone - cytokinin
A_92_P016737	GRMZM2G022679	3.07	0.37	*	HS 10-10 S	Gibberellin 2-oxidase	Metabolism - hormone - gibberelin
A_92_P012775	GRMZM5G848768	5.09	0.46	*	HS 9-6 S	3-ketoacyl-CoA thiolase	Metabolism - lipid - degradation
A_92_P024923	GRMZM2G169240	10.03	0.37	*	HS 9-6 S	Omega-6 fatty acid endoplasmic reticulum isozyme 2	Metabolism - lipid - FA desaturation
A_92_P032943	GRMZM2G149636	1.03	0.48	*	HS 9-6 S + HS 10-10 S	3-ketoacyl-CoA synthase 1	Metabolism - lipid - FA synthesis
A_92_P038520	GRMZM2G082924	7.02	0.54	*	HS 10-10 S	Cytidine deaminase	Metabolism - nucleotide - degradation
A_92_P038137	GRMZM2G057963	4.08	0.44	*	HS 9-6 S	Nudix hydrolase 13	Metabolism - nucleotide - salvage
A_92_P023090	GRMZM2G181236	1.07	0.56	*	HS 9-6 S	Cytochrome P450	Metabolism - secondary
A_92_P007140	GRMZM2G094375	10.02	0.48	*	HS 9-6 S	Laccase 7	Metabolism - secondary - simple phenols
A_92_P032474	GRMZM2G078500	3.06	0.49	*	HS 9-6 S + HS 10-10 S	Catalytic LigB subunit of aromatic ring-opening dioxygenase	Metabolism - secondary
A_92_P017679	GRMZM2G062396	2.01	0.46	*	HS 9-6 S + HS 10-10 S	Flavanone 3-hydroxylase	Metabolism - secondary - flavonoids
A_92_P009408	GRMZM2G124477	5.08	0.46	*	HS 9-6 S + HS 10-10 S	COP1-interacting protein 7	Photosynthesis
A_92_P002419	GRMZM2G024793	8.06	0.42	*	HS 10-10 S	Chloroplastic group IIA intron splicing facilitator chloroplastic-like	Photosynthesis
A_92_P030102	GRMZM2G064960	4.06	0.53	*	HS 9-6 S	Outer envelope membrane protein	Protein - chloroplast
A_92_P018938	GRMZM2G062724	1.01	0.55	*	HS 9-6 S	CHY zinc finger family	Protein - degradation - ubiquitin
A_92_P026115	GRMZM2G333756	1.02	0.52	*	HS 9-6 S	Kelch motif family protein	Protein - degradation
A_92_P022165	GRMZM2G143883	1.08	0.50	*	HS 9-6 S + HS 10-10 S	GMC oxidoreductase family	Redox
A_92_P018873	GRMZM2G063880	8.03	0.56	*	HS 9-6 S + HS 10-10 S	WRKY transcription factor	RNA - regulation of transcription

[a]Maize core bin where the reporter is located; [b]Pearson correlation coefficient of the reporter's expression profile with the GLS severity scores (* indicates significance; p-value <0.01); [c]The *trans*-eQTL hotspots that this gene belong to (or overlap when it is a *cis*-eQTL); [d]Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description; [e]MapMan BIN corresponding to Figure 5.2.

Table 5.2: Gene models in the greenyellow co-expression module with eQTLs in HS 9-6 S and HS 10-10 S (part 2 of 2).

Agilent ID	Gene ID	Gene core bin[a]	GLS severity cor[b]		*Trans*-eQTL overlap[c]	Annotation (AT, Rice, B2G)[d]	Functional category[e]
A_92_P013324	GRMZM2G068973	8.08	0.50	*	HS 10-10 S	NAC domain transcription factor	RNA - regulation of transcription
A_92_P015550	GRMZM2G106560	2.08	0.46	*	HS 10-10 S	WRKY transcription factor	RNA - regulation of transcription
A_92_P006123	GRMZM2G030673	8.04	0.39	*	HS 9-6 S	Calcium-dependent protein kinase	Signalling - calcium
A_92_P037035	GRMZM2G309327	9.05	0.40	*	*cis*-eQTL (HS 9-6 S)	Calmodulin-related calcium sensor protein / EF hand family	Signalling - calcium
A_92_P021227	GRMZM2G174315	6.04	0.55	*	HS 9-6 S + HS 10-10 S	Calmodulin-binding heat-shock	Signalling - calcium
A_92_P009091	GRMZM2G140231	2.07	0.49	*	HS 9-6 S	Cysteine-rich receptor-like kinase	Signalling - receptor kinases
A_92_P014596	GRMZM2G473266	10.04	0.58	*	HS 10-10 S	Two-component system sensor kinase	Signalling - postranslational modification
A_92_P019132	GRMZM2G002131	2.07	0.46	*	HS 10-10 S	Heat shock transcription factor	Stress - abiotic - heat
A_92_P022951	GRMZM2G052571	3.05	0.50	*	HS 9-6 S + HS 10-10 S	Glutathione S-transferase	Stress - abiotic
A_92_P007649	GRMZM2G125032	3.05	0.54	*	HS 9-6 S	Beta-1,3-glucanase	Stress - biotic
A_92_P034379	GRMZM2G028656	8.05	0.45	*	HS 9-6 S	Maize proteinase inhibitor	Stress - biotic
A_92_P001063	GRMZM2G447795	6.05	0.41	*	HS 9-6 S	Class III chitinase homologue	Stress - biotic
A_92_P009951	GRMZM2G139811	2.04	0.39	*	HS 9-6 S	C2 calcium/lipid-binding and GRAM domain containing	Stress - biotic
A_92_P006679	GRMZM2G116520	10.04	0.52	*	HS 9-6 S + HS 10-10 S	Bowman-birk type trypsin inhibitor	Stress - biotic
A_92_P022755	GRMZM2G005633	10.04	0.52	*	HS 9-6 S + HS 10-10 S	Chitinase	Stress - biotic
A_92_P030884	GRMZM2G113203	9.03	0.50	*	HS 9-6 S + HS 10-10 S	ABC transporter C family member	Transport - ABC transporters...
A_92_P019947	GRMZM2G099619	6.01	0.45	*	HS 10-10 S	White-brown-complex ABC transporter family	Transport - ABC transporters...
A_92_P002836	GRMZM2G126079	6.06	0.44	*	HS 10-10 S	NSP (nuclear shuttle protein)-interacting GTPase	Transport - other
A_92_P021490	GRMZM2G151418	2.04	0.50	*	HS 10-10 S	WPP domain-interacting protein	NA
A_92_P009914	GRMZM2G086714	7.02	0.35	*	HS 9-6 S	HD domain containing / RELA/SPOT homolog 3	NA
A_92_P013177	GRMZM2G420108	1.1	0.56	*	HS 9-6 S	Protein of unknown function	PUF
A_92_P001316	GRMZM2G094510	9.02	0.55	*	HS 9-6 S	Protein of unknown function	PUF
A_92_P038555	GRMZM2G321053	1.01	0.52	*	HS 10-10 S	Protein of unknown function	PUF
A_92_P008339	GRMZM2G087824	10.04	0.48	*	HS 10-10 S	Protein of unknown function	PUF
A_92_P013182	GRMZM2G083328	8.02	0.47	*	HS 9-6 S	Protein of unknown function	PUF
A_92_P016734	GRMZM5G883985	9.03	0.46	*	HS 10-10 S	Protein of unknown function	PUF
A_92_P013412	GRMZM2G435986	7.05	0.45	*	HS 9-6 S	Protein of unknown function	PUF
A_92_P008269	GRMZM2G113569	1.09	0.34	*	HS 9-6 S	Protein of unknown function	PUF
A_92_P007880	GRMZM2G542272	5.06	0.44	*	HS 9-6 S	Putative protein	PP

[a]Maize core bin where the reporter is located; [b]Pearson correlation coefficient of the reporter's expression profile with the GLS severity scores (* indicates significance; p-value <0.01); [c]The *trans*-eQTL hotspots that this gene belong to (or overlap when it is a *cis*-eQTL); [d]Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description; [e]MapMan BIN corresponding to Figure 5.2.

Table 5.3: Gene models in the turquoise co-expression module with eQTLs in HS 9-6 R and HS 9-7 R (part 1 of 3).

Agilent ID	Gene ID	Gene core bin[a]	GLS severity cor[b]		*Trans*-eQTL overlap[c]	Annotation (AT, Rice, B2G)[d]	Functional category[e]
A_92_P005314	GRMZM2G120300	7.03	-0.21		HS 9-6 R	Glyoxalase II	Biodegradation of Xenobiotics
A_92_P004713	GRMZM2G015861	5.03	-0.20		HS 9-7 R	Actin related protein	Cell - organisation
A_92_P009683	GRMZM2G071249	9.02	-0.24		HS 9-6 R	Proline-rich cell wall protein-like	Cell wall
A_92_P014185	GRMZM2G118467	1.08	-0.11		HS 9-6 R	Endonuclease exonuclease phosphatase	DNA - synthesis/chromatin structure
A_92_P010978	GRMZM2G322661	2.07	-0.25		HS 9-6 R	Exonuclease MUT-7 homolog	DNA - unspecified
A_92_P002976	GRMZM2G006507	4.04	-0.10		HS 9-6 R	Lysine ketoglutarate reductase *trans*-splicing related 1	Metabolism - amino acid - degradation
A_92_P011541	GRMZM2G107741	1.05	-0.18		HS 9-6 R + HS 9-7 R	3-hydroxyisobutyryl-CoA hydrolase-like protein	Metabolism - amino acid - degradation
A_92_P003562	GRMZM2G015906	8.04	-0.17		HS 9-6 R	Saposin-like type region 1 protein	Metabolism - lipid - degradation
A_92_P008295	GRMZM2G005886	9.03	-0.31	*	*cis*-eQTL (HS 9-6 R)	S-adenosyl-L-methionine-dependent methyltransferases	Metabolism - lipid - phospholipid synthesis
A_92_P007490	GRMZM2G158629	2.07	-0.05		HS 9-7 R	Enoyl-CoA hydratase peroxisomal-like	Metabolism - lipid
A_92_P013525	GRMZM2G093945	4.04	-0.22		HS 9-6 R	Galactose mutarotase-like	Metabolism - CHO
A_92_P013032	GRMZM2G174481	4.1	-0.20		HS 9-6 R	NADP-dependent D-sorbitol-6-phosphate dehydrogenase	Metabolism - CHO
A_92_P010785	GRMZM2G059212	5.03	-0.50	*	HS 9-6 R	Callose synthase 3-like	Metabolism - CHO - callose
A_92_P012007	GRMZM2G025528	3.05	-0.20		HS 9-6 R	Ureide permease	Metabolism - nucleotide - degradation
A_92_P006916	GRMZM5G880102	3.04	-0.39	*	HS 9-6 R	Protein translocase subunit chloroplastic-like	Photosynthesis
A_92_P008576	GRMZM2G001457	3.09	-0.33	*	HS 9-6 R	Phototropin 1	Photosynthesis
A_92_P005648	GRMZM2G019358	4.09	-0.27	*	HS 9-6 R	NAD(P)-binding Rossmann-fold superfamily protein	Photosynthesis
A_92_P007992	GRMZM2G011777	5.04	-0.36	*	HS 9-6 R + HS 9-7 R	ATP synthase mitochondrial F1 complex assembly factor 1	Protein
A_92_P010960	GRMZM2G087598	6.02	-0.22		HS 9-6 R	ATP-dependent zinc metalloprotease FtsH	Protein - degradation - metalloprotease
A_92_P007211	GRMZM2G053909	2.02	-0.24		HS 9-6 R + HS 9-7 R	Zinc C3HC4 type / RING/U-box	Protein - degradation - ubiquitin
A_92_P011044	GRMZM2G148213	7.05	-0.18		HS 9-7 R	BTB/POZ and math domain-containing protein 2-like	Protein - degradation - ubiquitin
A_92_P002436	GRMZM2G117746	9.05	-0.20		*cis*-eQTL (HS 9-6 R)	FKBP-type peptidyl-prolyl *cis-trans* isomerase	Protein - folding
A_92_P005283	GRMZM2G100107	7.02	-0.32		HS 9-7 R	Kinetochore-associated protein	Protein - postranslational modification
A_92_P013016	GRMZM2G098957	3.04	-0.33	*	HS 9-6 R	50S ribosomal protein L14	Protein - synthesis - ribosomal protein
A_92_P002189	GRMZM2G151285	3.05	-0.31	*	HS 9-6 R	Ribosomal protein S24 S35	Protein - synthesis - ribosomal protein
A_92_P006377	GRMZM2G163769	2.07	-0.27	*	HS 9-6 R	Ribosomal protein L17e	Protein - synthesis - ribosomal protein
A_92_P009003	GRMZM2G139407	10	-0.21		HS 9-6 R	Nuclear transport factor 2 and RNA motif domain-containing	Protein - targeting - nucleus
A_92_P006737	GRMZM2G136563	10	-0.20		HS 9-6 R	Vacuolar sorting-associated protein 26-like	Protein - targeting - secretory pathway
A_92_P007882	GRMZM2G314679	7.03	-0.21		HS 9-6 R + HS 9-7 R	Nonclathrin coat Zeta2-COP / SNARE-like	Protein - targeting - secretory pathway

[a]Maize core bin where the reporter is located; [b]Pearson correlation coefficient of the reporter's expression profile with the GLS severity scores (* indicates significance; p-value <0.01); [c]The *trans*-eQTL hotspots that this gene belong to (or overlap when it is a *cis*-eQTL); [d]Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description; [e]MapMan BIN corresponding to Figure 5.3.

Table 5.4: Gene models in the turquoise co-expression module with eQTLs in HS 9-6 R and HS 9-7 R (part 2 of 3).

Agilent ID	Gene ID	Gene core bin[a]	GLS severity cor[b]		*Trans*-eQTL overlap[c]	Annotation (AT, Rice, B2G)[d]	Functional category[e]
A_92_P001302	GRMZM2G081970	7.02	-0.28	*	HS 9-7 R	Metallo-hydrolase/oxidoreductase	Redox
A_92_P005627	GRMZM2G384327	7.04	-0.28	*	HS 9-6 R	Splicing factor 4-like protein	RNA - processing
A_92_P003645	GRMZM2G174938	4.08	-0.26		HS 9-6 R	Regulation of nuclear pre-mRNA domain-containing 1B	RNA - processing
A_92_P006669	GRMZM2G120167	4.03	-0.39	*	HS 9-6 R	bZIP transcription factor	RNA - Regulation of transcription
A_92_P005298	GRMZM5G890323	1.07	-0.31	*	HS 9-6 R + HS 9-7 R	AP2-domain DRE binding factor	RNA - Regulation of transcription
A_92_P005677	GRMZM2G119886	1.03	-0.17		HS 9-6 R + HS 9-7 R	ADP-ribosylation factor GTPase-activating protein	RNA - Regulation of transcription
A_92_P003738	GRMZM2G550865	9.06	-0.27	*	HS 9-7 R	TOPLESS-related protein / WD repeat-containing	RNA - Regulation of transcription
A_92_P006758	GRMZM2G120971	7.05	-0.14		HS 9-7 R	Transcription elongation factor	RNA - Regulation of transcription
A_92_P008762	GRMZM2G123796	5.06	-0.32	*	HS 9-6 R	RNA motif-containing protein	RNA - RNA binding
A_92_P011285	GRMZM2G051247	9.03	-0.31	*	HS 9-6 R + HS 9-7 R	RNA-binding family protein	RNA - RNA binding
A_92_P009517	GRMZM5G893343	7.05	-0.31	*	HS 9-7 R	Adenylate cyclase	Signalling
A_92_P014412	GRMZM2G114702	4.04	-0.27	*	HS 9-7 R	Sarcoplasmic reticulum histidine-rich calcium-binding	Signalling
A_92_P005875	GRMZM2G064193	4.04	-0.09		HS 9-7 R	Gibberellin receptor GID1L2	Signalling
A_92_P008905	GRMZM5G847466	6.07	-0.15		HS 9-6 R	Calmodulin-like /EF hand calcium-binding	Signalling - calcium
A_92_P005700	GRMZM2G157317	3.08	-0.33	*	HS 9-6 R	RanBP1 domain containing	Signalling - G-proteins
A_92_P003146	GRMZM2G071071	5	-0.24		HS 9-6 R	Mitochondrial Rho GTPase 1	Signalling - G-proteins
A_92_P010609	GRMZM2G070323	1.09	-0.28	*	HS 9-6 R	Serine threonine-protein phosphatase BSL2-like	Signalling - postranslational modification
A_92_P004896	GRMZM2G433433	9.02	-0.25		HS 9-6 R	Serine threonine kinase	Signalling - postranslational modification
A_92_P001248	GRMZM2G134389	6.04	-0.24		HS 9-6 R	Tyrosine specific phosphatase	Signalling - postranslational modification
A_92_P003478	GRMZM2G130943	9.06	-0.18		HS 9-6 R	Protein phosphatase	Signalling - postranslational modification
A_92_P010477	GRMZM2G160237	1.03	-0.31	*	HS 9-6 R + HS 9-7 R	Serine threonine phosphatase	Signalling - postranslational modification
A_92_P004672	GRMZM2G165383	4.01	-0.25		HS 9-7 R	Phosphoserine phosphatase	Signalling - postranslational modification
A_92_P012669	GRMZM2G176519	6.06	-0.15		HS 9-7 R	CBL-interacting serine threonine-protein kinase 11	Signalling - postranslational modification
A_92_P015294	GRMZM2G099598	1.01	-0.11		HS 9-7 R	ATP binding / BR-signaling kinase	Signalling - postranslational modification
A_92_P003478	GRMZM2G130943	9.06	-0.18		*cis*-eQTL (HS 9-7 R)	Protein phosphatase 2C family	Signalling - postranslational modification
A_92_P008364	GRMZM2G151567	10	-0.31	*	HS 9-6 R	Receptor-like serine threonine kinase	Signalling - receptor kinases
A_92_P014316	GRMZM2G049895	3.04	0.00		HS 9-6 R	B-type lectin receptor-like tyrosine-protein kinase	Signalling - receptor kinases
A_92_P005246	GRMZM5G809587	8.03	-0.37	*	HS 9-6 R	Cation proton exchanger 1A	Transport - calcium
A_92_P014444	GRMZM2G083677	1.1	-0.27	*	HS 9-6 R	Outward-rectifying potassium channel	Transport - potassium

[a]Maize core bin where the reporter is located; [b]Pearson correlation coefficient of the reporter's expression profile with the GLS severity scores (* indicates significance; p-value <0.01); [c]The *trans*-eQTL hotspots that this gene belong to (or overlap when it is a *cis*-eQTL); [d]Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description; [e]MapMan BIN corresponding to Figure 5.3.

Table 5.5: Gene models in the turquoise co-expression module with eQTLs in HS 9-6 R and HS 9-7 R (part 3 of 3).

Agilent ID	Gene ID	Gene core bin[a]	GLS severity cor[b]		*Trans*-eQTL overlap[c]	Annotation (AT, Rice, B2G)[d]	Functional category[e]
A_92_P007096	GRMZM2G034061	4.09	-0.24		*cis*-eQTL (HS 9-6 R)	Hexose carrier protein HEX6	Transport - sugars
A_92_P011761	GRMZM2G131785	9.05	-0.25		HS 9-7 R	ABC transporter family protein	Transport - ABC transporters...
A_92_P000588	GRMZM2G177912	3.09	-0.17		HS 9-7 R	Vacuolar ATP synthase subunit C	Transport - P and V-type ATPases
A_92_P012143	GRMZM2G017421	6.01	-0.26	*	HS 9-7 R	AP-1 complex subunit gamma-1	Transport - vesicle transport
A_92_P011520	GRMZM2G129513	1.07	-0.28	*	HS 9-6 R	Lactate/malate dehydrogenase	Tricarboxylic acid - transformation
A_92_P007798	GRMZM2G036099	3.06	-0.31	*	HS 9-7 R	Cupin, RmlC-type	NA
A_92_P014668	GRMZM2G074102	1.04	-0.22		HS 9-6 R + HS 9-7 R	SPFH domain-containing protein	NA
A_92_P015277	GRMZM2G115674	9.07	-0.32	*	HS 9-6 R + HS 9-7 R	Protein of unknown function	PUF
A_92_P003770	GRMZM2G099547	7.02	-0.30	*	HS 9-6 R	Protein of unknown function	PUF
A_92_P014224	GRMZM2G059753	3.07	-0.25		HS 9-6 R + HS 9-7 R	Protein of unknown function	PUF
A_92_P001487	GRMZM2G456564	8.03	-0.21		HS 9-6 R	Protein of unknown function	PUF
A_92_P002395	GRMZM2G068984	5.02	-0.20		HS 9-6 R	Protein of unknown function	PUF
A_92_P013448	GRMZM2G092256	9.03	-0.15		HS 9-6 R	Protein of unknown function	PUF
A_92_P010552	.	9.06	-0.27		HS 9-6 R	Putative protein	PP

[a]Maize core bin where the reporter is located; [b]Pearson correlation coefficient of the reporter's expression profile with the GLS severity scores (* indicates significance; p-value <0.01); [c]The *trans*-eQTL hotspots that this gene belong to (or overlap when it is a *cis*-eQTL); [d]Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description; [e]MapMan BIN corresponding to Figure 5.3.

Table 5.6: Gene models in the yellow co-expression module with eQTLs in HS 4-12 R (part 1 of 2).

Agilent ID	Gene ID	Gene core bin[a]	GLS severity cor[b]		*Trans*-eQTL overlap[c]	Annotation (AT, Rice, B2G)[d]	Functional category[e]
A_92_P004318	GRMZM2G008710	1.01	-0.30	*	HS 4-12 R	Post-GPI attachment to proteins / PER-1-like	Cell wall
A_92_P007232	GRMZM2G167872	8.06	-0.30	*	HS 4-12 R	Bifunctional polymyxin resistance arnA	Cell wall - precursor synthesis
A_92_P006115	GRMZM5G807211	1.1	-0.45	*	HS 4-12 R	Erythronate-4-phosphate dehydrogenase family	Metabolism - amino acid
A_92_P004954	GRMZM2G386802	3.09	-0.15		HS 4-12 R	Aminotransferase-like protein	Metabolism - amino acid
A_92_P005121	GRMZM2G022269	1.12	-0.20		HS 4-12 R	GTP binding mitochondrial elongation factor Tu	Metabolism - amino acid - elongation
A_92_P005951	GRMZM2G353874	7.04	-0.11		HS 4-12 R	Elongation factor 1	Metabolism - amino acid - elongation
A_92_P005900	GRMZM2G044337	8.01	-0.18		HS 4-12 R	Sphingosine-1-phosphate lyase	Metabolism - lipid - sphingolipids
A_92_P002244	GRMZM2G032163	4.07	-0.03		HS 4-12 R	Nudix hydrolase 3-like	Metabolism - nucleotide
A_92_P004620	GRMZM2G000622	1.1	-0.17		HS 4-12 R	Phosphoribosylformyl-glycinamidine synthase	Metabolism - nucleotide - synthesis
A_92_P008436	GRMZM2G121034	4.09	-0.36	*	*cis*-eQTL HS 4-12 R	Rossmann-fold NAD(P)-binding domain-containing	Photosynthesis
A_92_P006121	GRMZM2G039238	1.05	-0.21		HS 4-12 R	Serine hydroxymethyltransferase	Photosynthesis - photorespiration
A_92_P001834	GRMZM2G107116	6.04	-0.13		HS 4-12 R	Proteasome activator complex subunit 4 / HEAT repeat protein	Protein - degradation
A_92_P001246	GRMZM2G131245	1.06	-0.18		HS 4-12 R	RING U-box domain-containing	Protein - degradation - ubiquitin
A_92_P001368	GRMZM2G331368	10	-0.17		HS 4-12 R	Ubiquitin-protein ligase 1	Protein - degradation - ubiquitin
A_92_P001526	GRMZM2G171604	5.04	-0.09		HS 4-12 R	26S proteasome regulatory particle triple-A ATPase subunit4	Protein - degradation - ubiquitin
A_92_P001528	GRMZM2G089466	5.08	-0.22		HS 4-12 R	E3 ubiquitin-protein ligase / RING/U-box	Protein - degradation - ubiquitin
A_92_P005501	GRMZM2G071602	9.03	-0.26	*	HS 4-12 R	Zinc RING finger	Protein - degradation - ubiquitin
A_92_P008657	GRMZM2G072894	5	-0.27	*	HS 4-12 R	WWE protein-protein interaction domain protein family	Protein - degradation - ubiquitin
A_92_P000772	GRMZM2G019597	4.1	-0.26	*	HS 4-12 R	tRNA (guanine-N)-methyltransferase	Protein - postranslational modification
A_92_P011418	GRMZM2G073498	9.06	0.14		HS 4-12 R	Transport Sec61 subunit alpha	Protein - targeting - secretory pathway
A_92_P009358	GRMZM2G140339	1.01	0.09		HS 4-12 R	Winged-helix DNA-binding transcription factor	RNA - regulation of transcription
A_92_P000526	GRMZM2G014750	5.05	-0.28	*	HS 4-12 R	RNA recognition motif-containing	RNA - regulation of transcription
A_92_P001942	GRMZM2G129261	2.03	-0.21		HS 4-12 R	C2H2 zinc finger protein	RNA - regulation of transcription
A_92_P003065	GRMZM5G836167	2.07	-0.12		HS 4-12 R	Chromatin remodeling complex subunit	RNA - regulation of transcription
A_92_P004401	GRMZM2G088114	5.03	-0.21		HS 4-12 R	Heat-shock transcription factor	RNA - regulation of transcription
A_92_P004415	GRMZM2G370332	5.04	-0.22		HS 4-12 R	Homeobox transcription factor	RNA - regulation of transcription
A_92_P002134	GRMZM2G392798	5.05	-0.24		HS 4-12 R	Polyribonucleotide nucleotidyltransferase	RNA - processing
A_92_P005184	GRMZM2G016296	4.08	-0.31	*	*cis*-eQTL HS 4-12 R	RNA motif containing / splicing arginine serine-rich 2	RNA - RNA binding
A_92_P007145	GRMZM2G140602	1.08	-0.15		HS 4-12 R	Guanine nucleotide-binding 3 homolog / GTPase	Signalling - G-proteins
A_92_P001262	GRMZM2G114192	2.03	-0.14		HS 4-12 R	Rho GTPase activating protein 2	Signalling - G-proteins
A_92_P003297	GRMZM2G075719	7.03	-0.09		HS 4-12 R	Ras-related small GTP-binding	Signalling - G-proteins

[a]Maize core bin where the reporter is located; [b]Pearson correlation coefficient of the reporter's expression profile with the GLS severity scores (* indicates significance; p-value <0.01); [c]The *trans*-eQTL hotspot that this gene belong to (or overlap when it is a *cis*-eQTL); [d]Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description; [e]MapMan BIN corresponding to Figure 5.4.

Table 5.7: Gene models in the yellow co-expression module with eQTLs in HS 4-12 R (part 2 of 2).

Agilent ID	Gene ID	Gene core bin[a]	GLS severity cor[b]		*Trans*-eQTL overlap[c]	Annotation (AT, Rice, B2G)[d]	Functional category[e]
A_92_P001026	GRMZM5G895554	9.06	-0.31	*	HS 4-12 R	Adenylyl-sulfate kinase	Signalling - postranslational modification
A_92_P008456	GRMZM2G044724	3.04	-0.04		HS 4-12 R	NBS-LRR resistance protein	Stress - biotic
A_92_P011931	GRMZM5G803160	9.04	-0.40	*	HS 4-12 R	Zinc finger A20 and AN1 domains-containing protein	Stress - biotic
A_92_P005524	GRMZM2G477741	8.05	-0.32	*	HS 4-12 R	Metal tolerance protein A2 / cation efflux	Transport - metal
A_92_P004729	GRMZM2G171803	10	-0.15		HS 4-12 R	Transposon mutator	NA
A_92_P000691	GRMZM2G034096	1.05	-0.29	*	HS 4-12 R	Nuclear protein E3-3 isoform	NA
A_92_P001936	GRMZM2G126517	5.06	-0.06		HS 4-12 R	Pseudouridine-5'-phosphate	NA
A_92_P008901	GRMZM2G102815	4.07	-0.28	*	HS 4-12 R	Protein of unknown function	PUF
A_92_P003283	GRMZM2G323912	7.03	-0.05		HS 4-12 R	Protein of unknown function	PUF
A_92_P003289	.	4.08	-0.25		*cis*-eQTL HS 4-12 R	Putative protein	PP

[a]Maize core bin where the reporter is located; [b]Pearson correlation coefficient of the reporter's expression profile with the GLS severity scores (* indicates significance; p-value <0.01); [c]The *trans*-eQTL hotspot that this gene belong to (or overlap when it is a *cis*-eQTL); [d]Functional annotation derived from the best *Arabidopsis* and rice BLAST hits, as well as the Blast2GO description; [e]MapMan BIN corresponding to Figure 5.4.

Chapter 6

Concluding remarks

This study was designed to investigate the transcriptional variation underlying the quantitative genetic response to grey leaf spot (GLS) disease, caused by the fungus *Cercospora zeina*, in a recombinant inbred line (RIL) population derived from a CML444×SC Malawi cross. The major aims of the study were to: (i) annotate the reporter set of the Agilent-016047 microarray, using the maize B73 genome sequence; (ii) establish whether coordinated transcriptional responses to *C. zeina* infection under field conditions were evident in the RIL population; (iii) combine QTL mapping for GLS severity with eQTL analysis to investigate the molecular basis of the quantitative disease response to *C. zeina* infection; and finally (iv) elucidate the genetic basis for the coordinated expression responses to GLS disease. Agilent 44K microarrays were used for gene expression profiling across 100 RILs. A gene co-expression module analysis revealed 42 modules, of which eight were significantly correlated with GLS severity. Making use of 167 polymorphic DNA markers, eight QTLs for GLS severity were identified. In an eQTL analysis, 31,549 eQTLs for 30,280 e-traits were detected, comprising 4,866 *cis*-eQTLs (37.2%), 23,313 *trans*-eQTLs (74%) and 3,370 eQTLs (11%) that could not be classified. The analysis revealed 32 *trans*-eQTLs hotspots, five of which overlapped with GLS severity QTLs. The genes in three of the co-expression modules correlating with GLS severity were identified to have a significant number of *trans*-eQTLs in specific QTL-overlapping hotspots. This gave rise to a final list of 171 genes possibly involved in the mechanisms that might explain the resistance/susceptibility response for the respective GLS severity QTLs.

The QTL analysis revealed eight loci that affected GLS severity, implying underlying DNA polymorphisms in at least eight different genes. This was expected to result in eight

resistance/susceptibility "mechanisms", some of which may converge on the same resistance/susceptibility pathway. Assuming that some of these causal DNA polymorphisms also give rise to a change in gene expression and that such causal genes' expression profiles correlate with GLS severity, several mechanisms of defense were hypothesised. Although not all QTLs may be explained by gene expression differences, the study was able to highlight two hypotheses, both for QTL 9-5, based on potential networks where *cis* variation in regulatory factors gives rise to changes in expression levels for numerous genes in *trans*, potentially giving rise to phenotypic variation. For the susceptible response, it is hypothesised that a calmodulin-related protein with a *cis*-eQTL acts as a global regulator of various pathogenesis-related proteins that are activated in the susceptible RILs presumably in reaction to pathogen spreading in the leaf, but too late overall in the plant to result in resistance or tolerance. For the resistant response, it is hypothesised that a serine threonine-protein kinase with a *cis*-eQTL acts as a post-translational global regulator regulating phosphatases and kinases involved in activation of defense gene expression. The proposed hypotheses and identified QTLs may be transferable to other maize inbred lines, only if the underlying DNA polymorphism is found in those lines. Balint-Kurti *et al.* (2008) reported cases where GLS severity QTLs from different maize populations mapped to the same bin, which could indicate the same QTL with the same underlying DNA polymorphism, in which case the markers might be transferable. However, downstream responses are likely to be found also in other maize lines, since plants are thought to have a limited number of defense pathways, for example the salicylic acid and jasmonic acid signalling pathways (Dong, 1998) that lead to the production of antifungal proteins like pathogenesis-related proteins. Interestingly, in two cases, the genes in a co-expression module correlating with GLS severity were identified to have a significant number of *trans*-eQTLs in two different QTL-overlapping hotspots. This suggests that more than one regulatory gene likely participates in explaining the expression variation in the same endpoint genes/mechanisms.

This study further contributed towards the development of (i) a strategy to annotate the Agilent-016047 maize microarray and a publicly accessible annotation database (Coetzer *et al.*, 2011); and (ii) an eQTL data analysis pipeline in Galaxy, which can be used for global eQTL analyses in any species for which a draft reference genome sequence is available. Finally, a systems genetics strategy was developed to incorporate

the analysis of gene co-expression with phenotypic QTL and eQTL mapping, in order to prioritise candidate genes and identify putative regulatory gene networks associated with GLS resistance/susceptibility. This strategy can be replicated in other studies with similar research objectives.

The analysis of global gene expression profiling of genetically related plant populations has evolved over the past decade. A few authors studied the overall genetic architecture that explains expression polymorphisms in genome-wide eQTL studies (Kirst *et al.*, 2005; West *et al.*, 2007; Keurentjes *et al.*, 2007). A major step forward in plant QTL cloning has occurred via eQTLs, which can be utilised to search for associations between gene expression polymorphisms and phenotypic QTLs (Potokina *et al.*, 2008; Swanson-Wagner *et al.*, 2009; Chen *et al.*, 2010*b*; Claverie *et al.*, 2012). Correlation analysis between gene expression profiles and the phenotype values across the individuals of a population has proved useful for identification of candidate genes associated with phenotypic traits (Druka *et al.*, 2008). Furthermore, co-expression module-based network analysis has recently been used to discover specific genes and pathways relating to morphological phenotypic traits (Kugler *et al.*, 2013). Ultimately, combining the analysis of gene co-expression within genetic populations with genome-wide eQTL analysis is a powerful approach to dissecting quantitative phenotypic traits (Wang *et al.*, 2014). This study is close to the forefront of strategies for analysis of global gene expression data in the plant research community and implements a unique strategy by sequentially conducting a co-expression module-based network analysis identifying trait-related modules, followed by a global eQTL analysis identifying QTL-overlapping *trans*-eQTL hotspots. Finally these results are combined to identify a genetic basis for the coordinated expression responses to GLS disease.

A few limitations influenced the interpretation of the results from this study. This study was based on a small (100 RIL) bi-parental mapping population, implying limited mapping resolution and restricted allelic variation. Instead, using the maize IBM (Intermated B73 Mo17) RIL population would provide higher statistical power and genetic resolution than conventional RIL populations. Using the maize NAM population, a collection of 5,000 RILs, would capture the allelic variation between 25 diverse inbred lines and the reference line B73 (Yu *et al.*, 2008). Microarrays were used for gene expression profiling across the RILs. In microarray-based eQTL studies, spurious eQTLs can

be caused by (i) technical confounding factors (for example variations introduced during sample preparation or expression measurements); (ii) non-specific cross-hybridisation to highly similar sequences, gene families or alternatively spliced variants, i.e. cross-hybridisation of one probe to several targets; or (iii) hybridisation differences that are due to sequence polymorphisms rather than actual expression differences (for example when there are differences between the reference genome reporters and the subject genome). High-throughput sequencing technologies such as RNAseq or third-generation sequencing platforms (e.g. Pacific Biosciences) is an alternative that could overcome most of these limitations. This study was based on 100 RILs, which is less than what Ferreira *et al.* (2006) suggests. They proposed that that a total of 200 individuals is sufficient for the construction of reasonably accurate genetic maps. Increasing the size of a RIL population would result in (i) more measurements per allelic class, leading to an increase in statistical power to detect a QTL at a given location; and (ii) more recombination events within the population, providing greater genetic resolution (Hansen et al. 2008). Simple sequence repeat (SSR) and restriction fragment length polymorphism (RFLP) markers were used for QTL and eQTL mapping, with an average resolution of one marker every 11 cM. High-density single-nucleotide polymorphism (SNP) panels would have facilitated a higher density map (Ernst and Steibel, 2013), however this would only add value if a larger population was used. The genome sequences of the parental lines were lacking and thus candidate causative DNA polymorphisms could not be identified. Due to limited functional annotations in maize, BLAST alignments were employed to obtain annotations from better annotated plant species. The RIL population was sampled at only one late time point, during flowering when GLS lesions were evident. Thus, the initial defense response was not captured. A slight negative relationship between the GLS severity scores and the number of days after planting until flowering occurs (correlation coefficient of -0.21; p-value=0.01), indicating that late maturing lines appeared to be generally more resistant than early maturing lines (similar to the finding of Saghai Maroof *et al.*, 1996). A potential drawback in the strategy of using eQTLs to identify the polymorphism responsible for a phenotypic QTL, is that the approach will not be useful when the underlying DNA polymorphism does not give rise to change in the expression level of a gene. Furthermore, the strategy used in this study assumed that the causal gene's expression profile correlates (positively or negatively) with the phenotypic trait of

interest. A further assumption was that the allele associated with a higher expression was linked to the trait of interest (i.e. GLS resistance or susceptibility), depending on the underlying polymorphism. If this assumption is false, an eQTL will be identified as a candidate associated with the opposite effect, for example if GLS resistance is studied and the causal gene is negatively linked to this phenotype, the gene will be identified as a candidate associated with GLS susceptibility.

In future work, candidate genes that have been identified in this study as potentially important in GLS resistance (or susceptibility) need to be validated, typically in transformed maize or by generating near-isogenic lines (NILs), a longer approach, but one that does not suffer from possible extopic expression effects. The aim would be to show that over-expression of a specific gene (for example a gene proposed to be associated with resistance), results in increased resistance to GLS, or that silencing of such a gene results in increased susceptibility to GLS. Gene expression profiling of maize transformed for over-expression or silencing of a putative regulator can be used to validate whether the expression of selected genes with shared *trans*-eQTLs are explained by a causal gene at the position of these *trans*-eQTLs. It could be valuable to perform a genome-wide identification and expression profiling analysis for selected candidate gene families or pathways. Another form of validation would be to look at genome-wide association in independent maize pedigrees or populations like the NAM population (Yu *et al.*, 2008).

The future of understanding the genetic basis of phenotypic traits will be based on implementing similar approaches to the systems genetics strategy that was developed in this study. The incorporation of expression patterns of genes and gene modules with linkage mapping, contributes in this regard to elucidate the complex molecular networks underlying phenotypic traits. With technologies advancing and costs decreasing, it will be possible to conduct systems genetics analyses on larger samples, more environmental conditions, more developmental time points and more tissues. Metabolic QTL (mQTL) and protein QTL (pQTL) studies, which can be performed with data analysis strategies similar to those presented in this study for eQTLs, will provide a more complete picture of the effects of genetic perturbations on the physiology of whole organisms. Incorporating results from such studies with phenotypic QTL studies can be used to further validate and create new hypotheses. Furthermore, sizable and interactive databases to manage the different types of data and new statistical methodology to infer significant biological

networks will be needed. The eQTL data analysis pipeline that was developed in this study as a workflow in Galaxy, is a powerful and easily adaptable tool that can be used in future eQTL, mQTL and pQTL studies, for parallel QTL mapping and post-processing analysis of the resulting data, however interactive visualisation of such complex data types to allow data mining by users remain a challenge and will be the focus of future bioinformatics research.

In this study, a pioneering approach was developed to investigate the transcriptional variation underlying the quantitative genetic response of maize to GLS disease. Both the eQTL data analysis pipeline and the systems genetics strategy have broad application for systems genetics (Mizrachi *et al.*, 2012), not only in plant science, but also the medical and industrial biotechnology fields. This work led to candidate gene discovery in maize for fungal resistance, which can lead to novel control methods of GLS disease in the long term and provide fundamental new insight into the complex biology and genetics of plant-pathogen interactions.

Chapter 7

Bibliography

Abramovitch, R. B., Anderson, J. C. and Martin, G. B. (2006) Bacterial elicitation and evasion of plant innate immunity. *Nature Reviews* **7**, 601–611.

Adhikari, B. N., Savory, E. A., Vaillancourt, B., Childs, K. L., Hamilton, J. P., Day, B. and Buell, C. R. (2012) Expression profiling of *Cucumis sativus* in response to Infection by *Pseudoperonospora cubensis*. *PLoS One* **7**, 4.

Afzal, A. J., Wood, A. J. and Lightfoot, D. A. (2008) Plant receptor-like serine threonine kinases: roles in signaling and plant defense. *Molecular Plant-Microbe Interactions* **21**, 5, 507–517.

Agnihotri, G. and Liu, H. (2003) Enoyl-CoA hydratase: reaction, mechanism, and inhibition. *Bioorganic and Medicinal Chemistry* **11**, 1, 9–20.

Agrios, G. N. (2005) *Plant pathology.* Elsevier Academic Press 5th edition.

Albert, R. (2005) Scale-free networks in cell biology. *Journal of Cell Science* **118**, 4947–4957.

Albert, R. and Barabasi, A. (2000) Topology of evolving networks: local events and universality. *Physical Review Letters* **85**, 24, 5234–5237.

Albert, R. and Barabási, A.-L. (2002) Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**, 47–97.

Albert, R., Jeong, H. and Barabási, A.-L. (2000) Error and attack tolerance of layered complex networks. *Nature* **406**, 378–382.

Alberts, R., Terpstra, P., Li, Y., Breitling, R., Nap, J.-P. and Jansen, R. C. (2007) Sequence polymorphisms cause many false *cis* eQTLs. *PloS One* **2**, 7, e622.

Alexa, A. and Rahnenfuhrer, J. (2010) TopGO: Enrichment analysis for gene ontology. R package version 2.16.0.

Alonso-Blanco, C. and Koornneef, M. (2000) Naturally occurring variation in *Arabidopsis*: an underexploited resource for plant genetics. *Trends in Plant Science* **5**, 1, 22–29.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 3, 403–410.

Andorf, C. M., Lawrence, C. J., Harper, L. C., Schaeffer, M. L., Campbell, D. A. and Sen, T. Z. (2010) The Locus Lookup tool at MaizeGDB: identification of genomic regions in maize by integrating sequence information with physical and genetic maps. *Bioinformatics* **26**, 3, 434–436.

Angot, A., Peeters, N., Lechner, E., Vailleau, F., Baud, C., Gentzbittel, L., Sartorel, E., Genschik, P., Boucher, C. and Genin, S. (2006) *Ralstonia solanacearum* requires F-box-like domain-containing type III effectors to promote disease on several host plants. *Proceedings of the National Academy of Sciences* **103**, 39, 14620–14625.

Aoki, K., Ogata, Y. and Shibata, D. (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiology* **48**, 3, 381–390.

Aten, J., Fuller, T., Lusis, A. and Horvath, S. (2008) Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC Systems Biology* **2**, 34.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Chory, J., Dean, C., Debieu, M., Meaux, J. D., Joseph, R., Faure, N., Kniskern, J. M., Jones, J. D. G., Michael, T., Roux, F., Salt, D. E., Tang, C., Todesco, M. and Traw, M. B. (2010) Genome-wide association study of 107 phenotypes in a common set of *Arabidopsis thaliana* inbred lines. *Nature* **465**, 7298, 627–631.

Audenaert, K., Meyer, G. B. D. and Ho, M. M. (2002) Abscisic acid determines basal susceptibility of tomato to *Botrytis cinerea* and suppresses salicylic acid-dependent signaling mechanisms. *Plant Physiology* **128**, 491–501.

Ayroles, J. F., Carbone, M. A., Stone, E. A., Jordan, K. W., Lawrence, F., Anholt, R. R. H. and Mackay, T. F. C. (2009) Systems genetics of complex traits in *Drosophila melanogaster*. *Nature Genetics* **41**, 3, 299–307.

Azuaje, F., Zhang, L., Jeanty, C., Puhl, S.-L., Rodius, S. and Wagner, D. R. (2013) Analysis of a gene co-expression network establishes robust association between Col5a2 and ischemic heart disease. *BMC Medical Genomics* **6**, 13.

Balint-Kurti, P. J. and Johal, G. S. (2009) Maize disease resistance in Bennetzen, J. L. and Hake, S. C., editors, *Handbook of Maize: Its Biology* 229–250 Springer New York.

Balint-Kurti, P. J., Simmons, S. J., Blum, J. E., Ballaré, C. L. and Stapleton, A. E. (2010) Maize leaf epiphytic bacteria diversity patterns are genetically correlated with resistance to fungal pathogen infection. *Molecular Plant-Microbe Interactions* **23**, 4, 473–484.

Balint-Kurti, P. J., Wisser, R. and Zwonitzer, J. C. (2008) Use of an advanced intercross line population for precise mapping of quantitative trait loci for gray leaf spot resistance in maize. *Crop Science* **48**, 1696–1704.

Barabási, A.-L. and Bonabeau, E. (2003) Scale-free networks. *Scientific American* **288**, 60–69.

Barabási, A.-L. and Oltvai, Z. N. (2004) Network biology: understanding the cell's functional organization. *Nature Reviews* **5**, 2, 101–113.

Bari, R. and Jones, J. D. G. (2009) Role of plant hormones in plant defence responses. *Plant Molecular Biology* **69**, 4, 473–488.

Basten, C., Weir, B. and Zeng, Z.-B. (2001) *QTL Cartographer, Version 1.15.* Department of Statistics, North Carolina State University, Raleigh, NC.

Basten, C. J., Weir, B. S. and Zeng, Z.-B. (1994) Zmap–a QTL cartographer in Smith, C., Gavora, J. S., Benkel, B., Chesnais, J., Fairfull, W., Gibson, J. P., Kennedy, B. W. and

Burnside, E. B., editors, *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production: Computing Strategies and Software* volume 22 65–66 Organizing Committee, 5th World Congress on Genetics Applied to Livestock Production, Guelph, Ontario, Canada.

Beckman, P. M. and Payne, G. A. (1983) Cultural techniques and conditions influencing growth and sporulation of *Cercospora zeae-maydis* and lesion development in corn. *Phytopathology* **73**, 286–289.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57**, 1, 289–300.

Bergelson, J., Kreitman, M., Stahl, E. A. and Tian, D. (2001) Evolutionary dynamics of plant R-genes. *Science* **292**, 5525, 2281–2285.

Berger, D. K., Carstens, M., Korsman, J. N., Middleton, F., Kloppers, F. J., Tongoona, P. and Myburg, A. A. (2014) Mapping QTL conferring resistance in maize to gray leaf spot disease caused by Cercospora zeina. *BMC Genetics* **15**, 60.

Bhattacharjee, S. (2012) The language of reactive oxygen species signaling in plants. *Journal of Botany* **2012**, 1–22.

Bing, N. and Hoeschele, I. (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170**, 2, 533–542.

Boller, T. (1995) Chemoperception of microbial signals in plant cells. *Annual Review of Plant Physiology and Plant Molecular Biology* **46**, 189–214.

Bolwell, G. P. (1999) Role of active oxygen species and NO in plant defence responses. *Current Opinion in Plant Biology* **2**, 4, 287–294.

Bourdenx, B., Bernard, A., Domergue, F., Pascal, S., Léger, A., Roby, D., Pervent, M., Vile, D., Haslam, R. P., Napier, J. A., Lessire, R. and Joubès, J. (2011) Overexpression of *Arabidopsis* ECERIFERUM1 promotes wax very-long-chain alkane biosynthesis and influences plant response to biotic and abiotic stresses. *Plant Physiology* **156**, 1, 29–45.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y. and Buckler, E. S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 19, 2633–2635.

Breimans, A., Fawcettt, T. W., Ghirardill, M. L. and Mattool, A. K. (1992) Plant organelles contain distinct peptidylprolyl *cis*, *trans*-isomerases. *The Journal of Biological Chemistry* **267**, 30, 21293–21296.

Breitling, R., Li, Y., Tesson, B. M., Fu, J., Wu, C., Wiltshire, T., Gerrits, A., Bystrykh, L. V., de Haan, G., Su, A. I. and Jansen, R. C. (2008) Genetical genomics: spotlight on QTL hotspots. *PLoS Genetics* **4**, 10, e1000232.

Brem, R. B., Yvert, G., Clinton, R. and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 5568, 752–755.

Bu, Q., Jiang, H., Li, C.-B., Zhai, Q., Zhang, J., Wu, X., Sun, J., Xie, Q. and Li, C. (2008) Role of the *Arabidopsis thaliana* NAC transcription factors ANAC019 and ANAC055 in regulating jasmonic acid-signaled defense responses. *Cell Research* **18**, 7, 756–767.

Bubeck, D. M., Goodman, M. M., Beavis, W. D. and Grant, D. (1993) Quantitative trait loci controlling resistance to gray leaf spot in maize. *Crop Science* **33**, 4, 838–847.

Buckler, E. S., Gaut, B. S. and McMullen, M. D. (2006) Molecular and functional diversity of maize. *Current Opinion in Plant Biology* **9**, 2, 172–176.

Camera, S. L., Geoffroy, P., Samaha, H., Ndiaye, A., Rahim, G., Legrand, M. and Heitz, T. (2005) A pathogen-inducible patatin-like lipid acyl hydrolase facilitates fungal and bacterial host colonization in *Arabidopsis*. *The Plant Journal* **44**, 810–825.

Canonne, J., Froidure-Nicolas, S. and Rivas, S. (2011) Phospholipases in action during plant defense signaling. *Plant Signaling & Behavior* **6**, 1, 13–18.

Carter, S. L., Brechbühler, C. M., Griffin, M. and Bond, A. T. (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* **20**, 14, 2242–2250.

Casati, P. and Walbot, V. (2008) Maize lines expressing RNAi to chromatin remodeling factors are similarly hypersensitive to UV-B radiation but exhibit distinct transcriptome responses. *Epigenetics* **3**, 4, 216–229.

Cedara (1996) A profile of small to medium sized farmers in Natal and KwaZulu. Cedara Report No. N/A/96/14 - KwaZulu Department of Agriculture, Private Bag X9059, Pietermaritzburg, 3200, South Africa.

Chen, L.-Q., Hou, B.-H., Lalonde, S., Takanaga, H., Hartung, M. L., Xiao-Qu, Q., Guo, W.-J., Kim, J.-G., Underwood, W., Chaudhuri, B., Chermak, D., Antony, G., White, F. F., Somerville, S. C., Beth, M. and Frommer, W. B. (2010*a*) Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature* **468**, 7323, 527–532.

Chen, M., Wang, Q.-Y., Cheng, X.-G., Xu, Z.-S., Li, L.-C., Ye, X.-G., Xia, L.-Q. and Ma, Y.-Z. (2007) GmDREB2, a soybean DRE-binding transcription factor, conferred drought and high-salt tolerance in transgenic plants. *Biochemical and Biophysical Research Communications* **353**, 299–305.

Chen, X., Hackett, C. A., Niks, R. E., Hedley, P. E., Booth, C., Druka, A., Marcel, T. C., Vels, A., Bayer, M., Milne, I., Morris, J., Ramsay, L., Marshall, D., Cardle, L. and Waugh, R. (2010*b*) An eQTL analysis of partial resistance to *Puccinia hordei* in barley. *PloS One* **5**, 1, e8598.

Chilosi, G., Caruso, C., Caporale, C., Leonardi, L., Bertini, L., Buzi, A., Nobile, M., Magro, P. and Buonocore, V. (2000) Antifungal activity of a Bowman Birk-type trypsin inhibitor from wheat kernel. *Journal of Phytopathology* **148**, 477–481.

Choi, J., Huh, S. U., Kojima, M., Sakakibara, H., Paek, K.-H. and Hwang, I. (2010) The cytokinin-activated transcription factor ARR2 promotes plant immunity via TGA3/NPR1-dependent salicylic acid signaling in *Arabidopsis*. *Developmental Cell* **19**, 284–95.

Chou, H.-H., Hsia, A.-P., Mooney, D. L. and Schnable, P. S. (2004) Picky: oligo microarray design for large genomes. *Bioinformatics* **20**, 17, 2893–2902.

Claverie, M., Souquet, M., Jean, J., Forestier-Chiron, N., Lepitre, V., Pré, M., Jacobs,

J., Llewellyn, D. and Lacape, J.-M. (2012) cDNA-AFLP-based genetical genomics in cotton fibers. *Theoretical and Applied Genetics* **124**, 4, 665–683.

Clements, M. J., Dudley, J. W. and White, D. G. (2000) Quantitative trait loci associated with resistance to gray leaf spot of corn. *Phytopathology* **90**, 9, 1018–1025.

Cline, M., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A., Vailaya, A., Wang, P., Adler, A., Conklin, B., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G., Ideker, T. and Bader, G. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols* **2**, 10, 2366–2382.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. and de Hoon, M. J. L. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 11, 1422–1423.

Coetzer, N., Gazendam, I., Oelofse, D. and Berger, D. K. (2010) SSHscreen and SSHdb, generic software for microarray based gene discovery: application to the stress response in cowpea. *Plant Methods* **6**, 10.

Coetzer, N., Myburg, A. A. and Berger, D. K. (2011) Maize microarray annotation database. *Plant Methods* **7**, 31.

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

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 18, 3674–3676.

Copeland, R. (2008) *Essential SQLAalchemy.* O'Reilly.

Cordero, M. J., Raventos, D. and Segundo, B. S. (1994) Expression of a maize proteinase inhibitor gene is induced in response to wounding and fungal infection: systemic wound-response of a monocot gene. *The Plant Journal* **6**, 2, 141–150.

Cowan, M. M. (1999) Plant products as antimicrobial agents. *Clinical Microbiology Reviews* **12**, 4, 564–582.

Crampton, B., Hein, I. and Berger, D. K. (2009) Salicylic acid confers resistance to a biotrophic rust pathogen, *Puccinia substriata*, in pearl millet (*Pennisetum glaucum*) *Molecular Plant Pathology* **10**, 2, 291–304.

Crous, P. W., Groenewald, J. Z., Groenewald, M., Caldwell, P., Braun, U. and Harrington, T. C. (2006) Species of *Cercospora* associated with grey leaf spot of maize. *Studies in Mycology* **55**, 189–197.

Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695.

Cubillos, F. A., Coustham, V. and Loudet, O. (2012) Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. *Current Opinion in Plant Biology* **15**, 2, 192–198.

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., Watson, S. J. and Meng, F. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research* **33**, 20, e175.

Dangl, J. L. and Jones, J. D. G. (2001) Plant pathogens and integrated defence responses to infection. *Nature* **411**, 6839, 826–833.

de Jonge, R. (2010) Conserved fungal LysM effector Ecp6 prevents chitin-triggered immunity in plants. *Science* **329**, 953–955.

de Koning, D.-J. and Haley, C. S. (2005) Genetical genomics in humans and model organisms. *Trends in Genetics* **21**, 7, 377–381.

de Wit, P. J. G. M. (2002) Plant biology: on guard. *Nature* **416**, 6883, 801–803.

Dean, J. D., Goodwin, P. H. and Hsiang, T. (2005) Induction of glutathione S-transferase genes of *Nicotiana benthamiana* following infection by *Colletotrichum destructivum* and *C. orbiculare* and involvement of one in resistance. *Journal of Experimental Botany* **56**, 416, 1525–1533.

Delessert, C., Kazan, K., Wilson, I. W., Van Der Straeten, D., Manners, J., Dennis, E. S. and Dolferus, R. (2005) The transcription factor ATAF2 represses the expression of pathogenesis-related genes in *Arabidopsis*. *The Plant Journal* **43**, 5, 745–757.

Delker, C. and Quint, M. (2011) Expression level polymorphisms: heritable traits shaping natural variation. *Trends in Plant Science* **16**, 9, 481–488.

Dill, A., Thomas, S. G., Hu, J., Steber, C. M. and Sun, T.-P. (2004) The *Arabidopsis* F-box protein SLEEPY1 targets gibberellin signaling repressors for gibberellin-induced degradation. *The Plant Cell* **16**, 1392–1405.

Dimosthenis, K. and Montserrat, P. (2002) Maize DRE-binding proteins DBF1 and DBF2 are involved in rab17 regulation through the drought-responsive element in an ABA-dependent pathway. *The Plant Journal* **30**, 6, 679–689.

Dixon, R. A., Achnine, L., Kota, P., Liu, C.-J., Reddy, M. S. S. and Wang, L. (2002) The phenylpropanoid pathway and plant defence-a genomics perspective. *Molecular Plant Pathology* **3**, 5, 371–390.

Doebley, J., Stec, A. and Hubbard, L. (1997) The evolution of apical dominance in maize. *Nature* **386**, 6624, 485–488.

Doerge, R. W. and Churchill, G. A. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285–294.

Dong, X. (1998) SA, JA, ethylene, and disease resistance in plants. *Current Opinion in Plant Biology* **1**, 4, 316–323.

Donmez, N., Bazykin, G. A., Brudno, M. and Kondrashov, A. S. (2009) Polymorphism due to multiple amino acid substitutions at a codon site within *Ciona savignyi*. *Genetics* **181**, 2, 685–690.

Doxey, A. C., Yaish, M. W. F., Moffatt, B. A., Griffith, M. and Mcconkey, B. J. (2007) Functional divergence in the *Arabidopsis* beta-1,3-glucanase gene family inferred by phylogenetic reconstruction of expression states. *Molecular Biology and Evolution* **24**, 4, 1045–1055.

Drost, D. R., Benedict, C. I., Berg, A., Novaes, E., Novaes, C. R. D. B., Yu, Q., Dervinis, C., Maia, J. M., Yap, J., Miles, B. and Kirst, M. (2010) Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of Populus. *Proceedings of the National Academy of Sciences* **107**, 18, 8492–8497.

Druka, A., Potokina, E., Luo, Z., Bonar, N., Druka, I., Zhang, L., Marshall, D. F., Steffenson, B. J., Close, T. J., Wise, R. P., Kleinhofs, A., Williams, R. W., Kearsey, M. J. and Waugh, R. (2008) Exploiting regulatory variation to identify genes underlying quantitative resistance to the wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici* in barley. *Theoretical and Applied Genetics* **117**, 2, 261–272.

Druka, A., Potokina, E., Luo, Z., Jiang, N., Chen, X., Kearsey, M. and Waugh, R. (2010) Expression quantitative trait loci analysis in plants. *Plant Biotechnology Journal* **8**, 1, 10–27.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**, 14863–14868.

Ellinger, D., Naumann, M., Falter, C., Zwikowics, C., Jamrow, T., Manisseri, C., Somerville, S. C. and Voigt, C. A. (2013) Elevated early callose deposition results in complete penetration resistance to powdery mildew in *Arabidopsis*. *Plant Physiology* **161**, 3, 1433–1444.

Emrich, S. J., Barbazuk, W. B., Li, L. and Schnable, P. S. (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research* **17**, 1, 69–73.

Erdös, P. and Rényi, A. (1960) On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**, 17–61.

Ernst, C. W. and Steibel, J. P. (2013) Molecular advances in QTL discovery and application in pig breeding. *Trends in Genetics* **29**, 4, 215–224.

Ferreira, A., Flores, M. and Cruz, C. D. (2006) Estimating the effects of population size and type on the accuracy of genetic maps. *Genetics and Molecular Biology* **29**, 1, 187–192.

Flatman, R., McLauchlan, W. R., Juge, N., Furniss, C., Berrin, J.-G., Hughes, R. K., Manzanares, P., Ladbury, J. E., O'Brien, R. and Williamson, G. (2002) Interactions defining the specificity between fungal xylanases and the xylanase-inhibiting protein XIP-I from wheat. *The Biochemical Journal* **365**, 773–781.

Flint-Garcia, S. A., Thornsberry, J. M. and Buckler, E. S. (2003) Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* **54**, 357-374.

Francia, D., Chiltz, A., Lo Schiavo, F., Pugin, A., Bonfante, P. and Cardinale, F. (2011) AM fungal exudates activate MAP kinases in plant cells in dependence from cytosolic Ca^{2+} increase. *Plant Physiology and Biochemistry Journal* **49**, 9, 963–969.

Freeman, B. and Beattie, G. (2008) An overview of plant defenses against pathogens and herbivores. *The Plant Health Instructor*.

Fu, J. and Jansen, R. C. (2006) Optimal design and analysis of genetic studies on gene expression. *Genetics* **172**, 3, 1993–1999.

Fukao, T., Xu, K., Ronald, P. C. and Bailey-Serres, J. (2006) A variable cluster of ethylene response factor-like genes regulates metabolic and developmental acclimation responses to submergence in rice. *The Plant Cell* **18**, 2021–2034.

Gaj, S., van Erk, A., van Haaften, R. I. M. and Evelo, C. T. A. (2007) Linking microarray reporters with protein functions. *BMC Bioinformatics* **8**, 360.

Galbraith, D. W. and Edwards, J. (2010) Applications of microarrays for crop improvement: here, there, and everywhere. *BioScience* **60**, 5, 337–348.

García-Muniz, N., Martínez-Izquierdo, J. A. and Puigdomènech, P. (1998) Induction of mRNA accumulation corresponding to a gene encoding a cell wall hydroxyproline-rich glycoprotein by fungal elicitors. *Plant Molecular Biology* **38**, 623–632.

Ge, X., Li, G.-J., Wang, S.-B., Zhu, H., Zhu, T., Wang, X. and Xia, Y. (2007) AtNUDT7, a negative regulator of basal immunity in *Arabidopsis*, modulates two distinct defense response pathways and is involved in maintaining redox homeostasis. *Plant Physiology* **145**, 1, 204–215.

Gertz, E. M., Sengupta, K., Difilippantonio, M. J., Ried, T. and Schäffer, A. A. (2009) Evaluating annotations of an Agilent expression chip suggests that many features cannot be interpreted. *BMC Genomics* **10**, 566.

Ghazalpour, A., Doss, S., Kang, H., Farber, C., Wen, P.-Z., Brozell, A., Castellanos, R., Eskin, E., Smith, D. J., Drake, T. A. and Lusis, A. J. (2008) High-resolution mapping of gene expression using association in an outbred mouse stock. *PLoS Genetics* **4**, 8, e1000149.

Gilad, Y., Rifkin, S. A. and Pritchard, J. K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics* **24**, 8, 408–415.

Glazebrook, J. (2005) Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annual Review of Phytopathology* **43**, 205–227.

Glombitza, S., Dubuis, P.-H., Thulke, O., Welzl, G., Bovet, L., Götz, M., Affenzeller, M., Geist, B., Hehn, A., Asnaghi, C., Ernst, D., Seidlitz, H. K., Gundlach, H., Mayer, K. F., Martinoia, E., Werck-Reichhart, D., Mauch, F. and Schäffner, A. R. (2004) Crosstalk and differential response to abiotic and biotic stressors reflected at the transcriptional level of effector genes from secondary metabolism. *Plant Molecular Biology* **54**, 6, 817–835.

Gomi, K., Sasaki, A., Itoh, H., Ueguchi-tanaka, M., Ashikari, M., Kitano, H. and Matsuoka, M. (2004) GID2, an F-box subunit of the SCF E3 complex, specically interacts with phosphorylated SLR1 protein and regulates the gibberellin-dependent degradation of SLR1 in rice. *The Plant Journal* **37**, 626–634.

Gordon, S. G., Bartsch, M., Matthies, I., Gevers, H. O., Lipps, P. E. and Pratt, R. C. (2004) Linkage of molecular markers to *Cercospora zeae-maydis* resistance in maize. *Crop Science* **44**, 628–636.

Grennan, A. K. (2006) Gibberellin metabolism enzymes in rice. *Plant Physiology* **141**, 524–526.

Gutterson, N. and Reuber, T. L. (2004) Regulation of disease resistance pathways by AP2/ERF transcription factors. *Current Opinion in Plant Biology* **7**, 4, 465–471.

Hagenblad, J. and Nordborg, M. (2002) Sequence variation and haplotype structure surrounding the flowering time locus FRI in *Arabidopsis thaliana*. *Genetics* **161**, 1, 289–298.

Hansen, B. G., Halkier, B. A. and Kliebenstein, D. J. (2008) Identifying the molecular basis of QTLs: eQTLs add a new dimension. *Trends in Plant Science* **13**, 2, 72–77.

Harjes, C. E., Rocheford, T. R., Bai, L., Brutnell, T. P., Kandianis, C. B., Sowinski, S. G., Stapleton, A. E., Vallabhaneni, R., Williams, M., Wurtzel, E. T., Yan, J. and Buckler, E. S. (2008) Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. *Science* **319**, 5861, 330–333.

He, Y., Chung, E.-H., Hubert, D. A., Tornero, P. and Dangl, J. L. (2012) Specific missense alleles of the *Arabidopsis* jasmonic acid co-receptor COI1 regulate innate immune receptor accumulation and function. *PLoS Genetics* **8**, 10, e1003018.

Heath, M. C. (2000) Hypersensitive response-related death. *Plant Molecular Biology* **44**, 3, 321–334.

Hedden, P. and Kamiya, Y. (1997) Gibberellin biosynthesis: enzymes, genes and their regulation. *Annual Review of Plant Physiology and Plant Molecular Biology* **48**, 431–460.

Hinch, J. and Clarke, A. E. (1982) Callose formation in *Zea mays* as a response to infection with *Phytophthora cinnamomi*. *Physiologial Plant Pathology* **21**, 113–124.

Hiruma, K., Nishiuchi, T., Kato, T., Bednarek, P., Okuno, T., Schulze-Lefert, P. and Takano, Y. (2011) *Arabidopsis* ENHANCED DISEASE RESISTANCE 1 is required for pathogen-induced expression of plant defensins in nonhost resistance, and acts through interference of MYC2-mediated repressor function. *The Plant Journal* **67**, 6, 980–92.

Hoffmann, L., Besseau, S., Geoffroy, P., Ritzenthaler, C., Meyer, D., Lapierre, C., Pollet, B. and Legrand, M. (2004) Silencing of hydroxycinnamoyl-coenzyme a shikimate/quinate hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. *The Plant Cell* **16**, 1446–1465.

Holloway, B. and Li, B. (2010) Expression QTLs: applications for crop improvement. *Molecular Breeding* **26**, 3, 381–391.

Holloway, B., Luck, S., Beatty, M., Rafalski, J.-A. and Li, B. (2011) Genome-wide expression quantitative trait loci (eQTL) analysis in maize. *BMC Genomics* **12**, 336.

Hsieh, M.-H. and Goodman, H. M. (2005) A novel gene family in *Arabidopsis* encoding putative heptahelical transmembrane proteins homologous to human adiponectin receptors and progestin receptors. *Journal of Experimental Botany* **56**, 422, 3137–3147.

Hu, J., Baker, A., Bartel, B., Linka, N., Mullen, R. T. and Reumann, S. (2012) Plant Peroxisomes: Biogenesis and Function. *The Plant Cell* **24**, 2279–2303.

Hubner, N., Wallace, C. A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., Musilova, A., Kren, V., Causton, H., Game, L., Born, G., Schmidt, S., Müller, A., Cook, S. A., Kurtz, T. W., Whittaker, J., Pravenec, M. and Aitman, T. J. (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics* **37**, 3, 243–253.

Hulbert, S. H., Webb, C. A., Smith, S. M. and Sun, Q. (2001) Resistance gene complexes: evolution and utilization. *Annual Review of Phytopathology* **39**, 285–312.

Igawa, T., Tokai, T., Kudo, T., Yamaguchi, I. and Kimura, M. (2005) A wheat xylanase inhibitor gene, Xip-I, but not Taxi-I, is significantly induced by biotic and abiotic signals that trigger plant defense. *Bioscience Biotechnology and Biochemistry* **69**, 5, 1058–1063.

Iida, S. and Terada, R. (2004) A tale of two integrations, transgene and T-DNA: gene targeting by homologous recombination in rice. *Current Opinion in Biotechnology* **15**, 2, 132–138.

Ingvarsson, P. K. and Street, N. R. (2011) Association genetics of complex traits in plants. *The New Phytologist* **189**, 4, 909–922.

Iyer-Pascuzzi, A. S., Symonova, O., Mileyko, Y., Hao, Y., Belcher, H., Harer, J., Weitz, J. S. and Benfey, P. N. (2010) Imaging and analysis platform for automatic phenotyping and trait ranking of plant root systems. *Plant Physiology* **152**, 1148–1157.

Jacobs, A. K., Lipka, V., Burton, R. A., Panstruga, R., Strizhov, N., Schulze-Lefert, P. and Fincher, G. B. (2003) An *Arabidopsis* callose synthase, GSL5, is required for wound and papillary callose formation. *The Plant Cell* **15**, 2503–2513.

Jansen, R. C. and Nap, J. P. (2001) Genetical genomics: the added value from segregation. *Trends in Genetics* **17**, 7, 388–91.

Jen, C.-H., Manfield, I. W., Michalopoulos, I., Pinney, J. W., Willats, W. G. T., Gilmartin, P. M. and Westhead, D. R. (2006) The *Arabidopsis* co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis. *The Plant Journal* **46**, 336–348.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabási, A. L. (2000) The large-scale organization of metabolic networks. *Nature* **407**, 6804, 651–654.

Jin, H., Vacic, V., Girke, T., Lonardi, S. and Zhu, J.-K. (2008) Small RNAs and the regulation of *cis*-natural antisense transcripts in *Arabidopsis*. *BMC Molecular Biology* **9**, 6.

Jones, J. D. G. and Dangl, J. L. (2006) The plant immune system. *Nature* **444**, 7117, 323–329.

Joosen, R. V. L., Ligterink, W., Hilhorst, H. W. M. and Keurentjes, J. J. B. (2009) Advances in genetical genomics of plants. *Current Genomics* **10**, 8, 540–549.

Jordan, M. C., Somers, D. J. and Banks, T. W. (2007) Identifying regions of the wheat genome controlling seed development by mapping expression quantitative trait loci. *Plant Biotechnology Journal* **5**, 3, 442–453.

Juliatti, F., Pedrosa, M., Silva, H. and da Silva, J. (2009) Genetic mapping for resistance to gray leaf spot in maize. *Euphytica* **169**, 227–238.

Kang, H. M., Ye, C. and Eskin, E. (2008*a*) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* **180**, 4, 1909–1925.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J. and Eskin, E. (2008*b*) Efficient control of population structure in model organism association mapping. *Genetics* **178**, 3, 1709–1723.

Kao, C.-H., Zeng, Z.-B. and Teasdale, R. D. (1999) Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 203–1216.

Kavanagha, K. L., Jçrnvallb, H., Perssonc, B. and Oppermann, U. (2008) The SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes. *Cell and Molecular Life Sciences* **65**, 3895–3906.

Kerr, M. K. and Churchill, G. A. (2001) Experimental design for gene expression microarrays. *Biostatistics* **2**, 2, 183–201.

Keurentjes, J. J. B., Fu, J., Terpstra, I. R., Garcia, J. M., van den Ackerveken, G., Snoek, L. B., Peeters, A. J. M., Vreugdenhil, D., Koornneef, M. and Jansen, R. C. (2007) Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences* **104**, 5, 1708–1713.

Keurentjes, J. J. B., Sulpice, R., Gibon, Y., Steinhauser, M.-C., Fu, J., Koornneef, M., Stitt, M. and Vreugdenhil, D. (2008) Integrative analyses of genetic variation in enzyme activities of primary carbohydrate metabolism reveal distinct modes of regulation in *Arabidopsis thaliana*. *Genome Biology* **9**, 8, R129.

Kim, H., Ridenour, J. B., Dunkle, L. D. and Bluhm, B. H. (2011) Regulation of pathogenesis by light in *Cercospora zeae-maydis*: an updated perspective. *The Plant Pathology Journal* **27**, 2, 103–109.

Kim, H. S. (2002) *Arabidopsis* SON1 is an F-box protein that regulates a novel induced defense response independent of both salicylic acid and systemic acquired resistance. *The Plant Cell Online* **14**, 7, 1469–1482.

Kim, M. C., Lee, S. H., Kim, J. K., Chun, H. J., Choi, M. S., Chung, W. S., Moon, B. C., Kang, C. H., Park, C. Y., Yoo, J. H., Kang, Y. H., Koo, S. C., Koo, Y. D., Jung, J. C., Kim, S. T., Schulze-Lefert, P., Lee, S. Y. and Cho, M. J. (2002) Mlo, a modulator of plant defense and cell death, is a novel calmodulin-binding protein. Isolation and characterization of a rice Mlo homologue. *The Journal of Biological Chemistry* **277**, 22, 19304–19314.

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D. and Nordborg, M. (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* **39**, 9, 1151–1155.

Kirst, M., Basten, C. J., Myburg, A. A., Zeng, Z.-B. and Sederoff, R. R. (2005) Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid. *Genetics* **169**, 4, 2295–2303.

Kirst, M., Caldo, R., Casati, P., Tanimoto, G., Walbot, V., Wise, R. P. and Buckler, E. S. (2006) Genetic diversity contribution to errors in short oligonucleotide microarray analysis. *Plant Biotechnology Journal* **4**, 5, 489–498.

Kliebenstein, D. J. (2009) Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annual Review of Plant Biology* **60**, 93–114.

Kliebenstein, D. J., West, M. A. L., van Leeuwen, H., Loudet, O., Doerge, R. W. and St Clair, D. A. (2006) Identification of QTLs controlling gene expression networks defined *a priori*. *BMC Bioinformatics* **7**, 308.

Kloosterman, B., Kumari, A. M., Chibon, P.-Y., Oortwijn, M., van der Linden, G. C., Visser, R. G. and Bachem, C. W. (2012) Organ specificity and transcriptional control of metabolic routes revealed by expression QTL profiling of source–sink tissues in a segregating potato population. *BMC Plant Biology* **12**, 17.

Krattinger, S. G., Lagudah, E. S., Spielmeyer, W., Singh, R. P., Huerta-espino, J., Mcfadden, H., Bossolini, E., Selter, L. L. and Keller, B. (2009) A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. *Science* **323**, 1360–1363.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. and Marra, M. A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Research* **19**, 9, 1639–1645.

Kugler, K. G., Siegwart, G., Nussbaumer, T., Ametz, C., Spannagl, M., Steiner, B., Lemmens, M., Mayer, K. F., Buerstmayr, H. and Schweiger, W. (2013) Quantitative trait loci-dependent analysis of a gene co-expression network associated with Fusarium head blight resistance in bread wheat (*Triticum aestivum* L.). *BMC Genomics* **14**, 728.

Kump, K. L., Bradbury, P. J., Wisser, R. J., Buckler, E. S., Belcher, A. R., Oropeza-Rosas, M. A., Zwonitzer, J. C., Kresovich, S., McMullen, M. D., Ware, D., Balint-

Kurti, P. J. and Holland, J. B. (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nature Genetics* **43**, 2, 163–168.

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., Zhang, M., Jiao, Y., Ni, P., Zhang, J., Li, D., Guo, X., Ye, K., Jian, M., Wang, B., Zheng, H., Liang, H., Zhang, X., Wang, S., Chen, S., Li, J., Fu, Y., Springer, N. M., Yang, H., Wang, J., Dai, J., Schnable, P. S. and Wang, J. (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genetics* **42**, 11, 1027–1030.

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T.-K., Flowers, M. T., Schueler, K. L., Manly, K. F., Williams, R. W., Kendziorski, C. and Attie, A. D. (2006) Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics* **2**, 1, e6.

Lander, E. S. and Botstein, D. (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 1, 185–199.

Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., Lincoln, S. E., Newberg, L. A. and Newburg, L. (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**, 174–181.

Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 599.

Langfelder, P. and Horvath, S. (2012) Tutorial for the WGCNA package for R: network analysis of liver expression data in female mice.

Latchman, D. S. (2005) *Gene regulation.* Taylor & Francis Group 5th edition.

Lattanzio, V., Lattanzio, V. M. T. and Cardinali, A. (2006) Role of phenolics in the resistance mechanisms of plants against fungal pathogens and insects in Imperato, F., editor, *Phytochemistry: Advances in Research* volume 661 23–67 India: Research Signpost.

Latterell, F. M. and Rossi, A. E. (1983) Gray leaf spot of corn: a disease on the move. *Plant Disease* **67**, 842–847.

Law, P. J., Claudel-Renard, C., Joubert, F., Louw, A. I. and Berger, D. K. (2008) MADIBA: a web server toolkit for biological interpretation of Plasmodium and plant gene clusters. *BMC Genomics* **9**, 105.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M. and Rhee, S. Y. (2010) Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nature Biotechnology* **28**, 2, 149–156.

Lehmensiek, A., Esterhuizen, A., van Staden, D., Nelson, S. W. and Retief, A. E. (2001) Genetic mapping of gray leaf spot (GLS) resistance genes in maize. *Theoretical and Applied Genetics* **103**, 5, 797–803.

Lemos, B., Araripe, L. O., Fontanillas, P. and Hartl, D. L. (2008) Dominance and the evolutionary accumulation of *cis*- and *trans*-effects on gene expression. *Proceedings of the National Academy of Sciences* **105**, 38, 14471–14476.

Li, H., Bradbury, P., Ersoz, E., Buckler, E. S. and Wang, J. (2011) Joint QTL linkage mapping for multiple-cross mating design sharing one common parent. *PloS One* **6**, 3, e17573.

Li, J., Lease, K. A., Tax, F. E. and Walker, J. C. (2001) BRS1, a serine carboxypeptidase, regulates BRI1 signaling in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences* **98**, 10, 5916–5921.

Li, L., Petsch, K., Shimizu, R., Liu, S., Xu, W. W., Ying, K., Yu, J., Scanlon, M. J., Schnable, P. S., Timmermans, M. C. P., Springer, N. M. and Muehlbauer, G. J. (2013) Mendelian and non-Mendelian regulation of gene expression in maize. *PLoS Genetics* **9**, 1, e1003202.

Li, X., Zhang, J. B., Song, B., Li, H. P., Xu, H. Q., Qu, B., Dang, F. J. and Liao, Y. C. (2010) Resistance to Fusarium head blight and seedling blight in wheat is associated with activation of a cytochrome P450 gene. *Phytopathology* **100**, 2, 183–191.

Lincoln, J. E., Richael, C., Overduin, B., Smith, K., Bostock, R. and Gilchrist, D. G. (2002) Expression of the antiapoptotic baculovirus p35 gene in tomato blocks programmed cell death and provides broad-spectrum resistance to disease. *Proceedings of the National Academy of Sciences* **99**, 23, 15217–15221.

Liu, H., Wang, X., Zhang, H., Yang, Y., Ge, X. and Song, F. (2008) A rice serine carboxypeptidase-like gene OsBISCPL1 is involved in regulation of defense responses against biotic and oxidative stress *Gene* **420**, 57–65.

Liu, K. and Xu, X. (2013) First report of gray leaf spot of maize caused by *Cercospora zeina* in China. *Plant Disease* **97**, 1656.

Longhi, S., Hamblin, M. T., Trainotti, L., Peace, C. P., Velasco, R. and Costa, F. (2013) A candidate gene based approach validates *Md-PG1* as the main responsible for a QTL impacting fruit texture in apple (*Malus* x *domestica* Borkh). *BMC Plant Biology* **13**, 37.

López-García, B., Hernández, M. and Segundo, B. S. (2012) Bromelain, a cysteine protease from pineapple (*Ananas comosus*) stem, is an inhibitor of fungal plant pathogens. *Letters in Applied Microbiology* **55**, 62–67.

Lorenz, W. W., Alba, R., Yu, Y.-S., Bordeaux, J. M., Simões, M. and Dean, J. F. (2011) Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (*P. taeda* L.). *BMC Genomics* **12**, 264.

Lorrain, S., Lin, B., Auriac, M. C., Kroj, T., Saindrenan, P., Nicole, M., Roby, D. and Balague, C. (2004) VASCULAR ASSOCIATED DEATH1, a novel GRAM domain-containing protein, is a regulator of cell death and defense responses in vascular tissues. *The Plant Cell* **16**, 2217–2232.

Lu, Y., Shah, T., Hao, Z., Taba, S., Zhang, S., Gao, S., Liu, J., Cao, M., Wang, J., Prakash, A. B., Rong, T. and Xu, Y. (2011) Comparative SNP and haplotype analysis reveals a higher genetic diversity and rapider LD decay in tropical than temperate germplasm in maize. *PloS One* **6**, 9, e24861.

Lyimo, H. J. F., Pratt, R. C. and Mnyuku, R. S. O. W. (2013) Infection process in resistant and susceptible maize (*Zea mays* L.) genotypes to *Cercospora zeae-maydis* (type II). *Plant Protection Science* **49**, 1, 11–18.

Ma, B., Luo, Y., Jia, L., Qi, X., Zeng, Q., Xiang, Z. and He, N. (2013) Genome-wide identification and expression analyses of cytochrome P450 genes in mulberry (*Morus notabilis*). *Journal of Integrative Plant Biology* **0**, 1–15.

Ma, J., Duncan, D., Morrow, D. J., Fernandes, J. and Walbot, V. (2007) Transcriptome profiling of maize anthers using genetic ablation to analyze pre-meiotic and tapetal cell types. *Plant Journal* **50**, 4, 637–48.

Ma, J., Morrow, D. J., Fernandes, J. and Walbot, V. (2006) Comparative profiling of the sense and antisense transcriptome of maize lines. *Genome Biology* **7**, 3, R22.

Ma, J., Skibbe, D. S., Fernandes, J. and Walbot, V. (2008) Male reproductive development: gene expression profiling of maize anther and pollen ontogeny. *Genome Biology* **9**, 12, R181.

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

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

Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 16, 3448–3449.

Maes, T., Keukeleire, P. D. and Gerats, T. (1999) Plant tagnology. *Trends in Plant Science* **4**, 3, 90–96.

Majello, B., De Luca, P. and Lania, L. (1997) Sp3 is a bifunctional transcription regulator with modular independent activation and repression comains. *Journal of Biological Chemistry* **272**, 7, 4021–4026.

Manfield, I. W., Jen, C.-H., Pinney, J. W., Michalopoulos, I., Bradford, J. R., Gilmartin, P. M. and Westhead, D. R. (2006) *Arabidopsis* Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. *Nucleic Acids Research* **34**, Web Server issue, W504–W509.

Manly, K., Cudmore, R. and Meer, J. (2001) Map Manager QTX, cross-platform software for genetic mapping. *Mammalian Genome* **12**, 930–932.

Marathe, R. and Dinesh-Kumar, S. P. (2003) Plant defense: one post, multiple guards?! *Molecular Cell* **11**, 2, 284–286.

Marone, D., Russo, M. A., Laidò, G., Leonardis, A. M. D. and Mastrangelo, A. M. (2013) Plant nucleotide binding site-leucine-rich repeat (NBS-LRR) genes: active guardians in host defense responses. *International Journal of Molecular Sciences* **14**, 7302–7326.

Marrs, K. A. (1996) The functions and regulation of glutathione S-transferases in plants. *Annual Review of Plant Physiology and Plant Molecular Biology* **47**, 127–158.

Martinoia, E., Klein, M., Geisler, M., Bovet, L., Forestier, C., Kolukisaoglu, U., Müller-Röber, B. and Schulz, B. (2002) Multifunctionality of plant ABC transporters - more than just detoxifiers. *Planta* **214**, 345–355.

Mauch-Mani, B. and Mauch, F. (2005) The role of abscisic acid in plant-pathogen interactions. *Current Opinion in Plant Biology* **8**, 4, 409–414.

Mccallum, C. M., Comai, L., Greene, E. A. and Henikoff, S. (2000) Targeting Induced Local Lesions IN Genomes (TILLING) for plant functional genomics. *Plant Physiology* **123**, June, 439–442.

Meisel, B., Korsman, J., Kloppers, F. J. and Berger, D. K. (2009) *Cercospora zeina* is the causal agent of grey leaf spot disease of maize in southern Africa. *European Journal of Plant Pathology* **124**, 4, 577–583.

Melotto, M., Underwood, W., Koczan, J., Nomura, K. and He, S. Y. (2006) Plant stomata function in innate immunity against bacterial invasion. *Cell* **126**, 969–980.

Menkir, A. and Ayodele, M. (2005) Genetic analysis of resistance to gray leaf spot of midaltitude maize inbred lines. *Crop Science* **45**, 163–170.

Messmer, R., Fracheboud, Y., Bänziger, M., Vargas, M., Stamp, P. and Ribaut, J.-M. (2009) Drought stress and tropical maize: QTL-by-environment interactions and stability of QTLs across environments for yield components and secondary traits. *Theoretical and Applied Genetics* **119**, 913–30.

Mészáros, T., Helfer, A., Hatzimasoura, E., Magyar, Z., Serazetdinova, L., Rios, G., Bardóczy, V., Teige, M., Koncz, C., Peck, S. and Bögre, L. (2006) The *Arabidopsis* MAP kinase kinase MKK1 participates in defence responses to the bacterial elicitor flagellin. *The Plant Journal* **48**, 4, 485–98.

Michaelson, J. J., Loguercio, S. and Beyer, A. (2009) Detection and interpretation of expression quantitative trait loci (eQTL). *Methods* **48**, 3, 265–276.

Mignery, G. A., Pikaard, C. S. and Park, W. D. (1988) Molecular characterization of the patatin multigene family of potato. *Gene* **62**, 1, 27–44.

Miller, J., Oldham, M. C. and Geschwind, D. H. (2008) A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. *The Journal of Neuroscience* **28**, 6, 1410–1420.

Mizrachi, E., Mansfield, S. D. and Myburg, A. A. (2012) Cellulose factories: advancing bioenergy production from forest trees. *The New Phytologist* **194**, 54–62.

Morant, A. V., Jø rgensen, K., Jø rgensen, C., Paquette, S. M., Sánchez-pérez, R., Mø ller, B. L. and Bak, S. (2008) Beta-Glucosidases as detonators of plant chemical defense. *Phytochemistry* **69**, 1795–1813.

Moscou, M. J., Lauter, N., Steffenson, B. and Wise, R. P. (2011) Quantitative and qualitative stem rust resistance factors in barley are associated with transcriptional suppression of defense regulons. *PLoS Genetics* **7**, 7, e1002208.

Mueller, M., Goel, A., Thimma, M., Dickens, N. J., Aitman, T. J. and Mangion, J. (2006) eQTL Explorer: integrated mining of combined genetic linkage and expression experiments. *Bioinformatics* **22**, 4, 509–511.

Munkvold, G. P., Martinson, C. A., Shriver, J. M. and Dixon, P. M. (2001) Probabilities for profitable fungicide use against gray leaf spot in hybrid maize. *Phytopathology* **91**, 5, 477–484.

Nadeau, J. H. and Dudley, A. M. (2011) Systems genetics. *Science* **311**, 6020, 1015–1016.

Nakashita, H., Yasuda, M., Nitta, T., Asami, T., Fujioka, S., Arai, Y., Sekimata, K., Takatsuto, S., Yamaguchi, I. and Yoshida, S. (2003) Brassinosteroid functions in a broad range of disease resistance in tobacco and rice. *The Plant Journal* **33**, 5, 887–898.

Naoumkina, M., Zhao, Q., Gallego-giraldo, L., Dai, X., Zhao, P. X. and Dixon, R. A. (2010) Genome-wide analysis of phenylpropanoid defence pathways. *Molecular Plant Pathology* **11**, 6, 829–846.

Navarro, L., Bari, R., Achard, P., Lison, P., Nemri, A., Harberd, N. P. and Jones, J. D. G. (2008) DELLAs control plant immune responses by modulating the balance of jasmonic acid and salicylic acid signaling. *Current Biology* **18**, 650–655.

Navarro, L., Dunoyer, P., Jay, F., Arnold, B., Dharmasiri, N., Estelle, M., Voinnet, O. and Jones, J. D. G. (2006) A plant miRNA contributes to antibacterial resistance by repressing auxin signaling. *Science* **312**, 5772, 436–439.

Nordborg, M., Borevitz, J. O., Bergelson, J., Berry, C. C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J. N., Noyes, T., Oefner, P. J., Stahl, E. A. and Weigel, D. (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* **30**, 2, 190–193.

Nordborg, M. and Tavaré, S. (2002) Linkage disequilibrium: what history has to tell us. *Trends in Genetics* **18**, 2, 83–90.

Obayashi, T., Hayashi, S., Saeki, M., Ohta, H. and Kinoshita, K. (2009) ATTED-II provides coexpressed gene networks for *Arabidopsis*. *Nucleic Acids Research* **37**, 987–991.

Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K. and Ohta, H. (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in *Arabidopsis*. *Nucleic Acids Research* **35**, 863–869.

Ohtsubo, N., Mitsuhara, I., Koga, M., Seo, S. and Ohashi, Y. (1999) Ethylene promotes the necrotic lesion formation and basic PR gene expression in TMV-infected tobacco. *Plant and Cell Physiology* **40**, 8, 808–817.

Pandey, S. P. and Somssich, I. E. (2009) The role of WRKY transcription factors in plant immunity. *Plant Physiology* **150**, August, 1648–1655.

Panstruga, R. (2005) Serpentine plant MLO proteins as entry portals for powdery mildew fungi. *Biochemical Society Transactions* **33**, Pt 2, 389–392.

Paponov, I. A., Paponov, M., Teale, W., Menges, M., Chakrabortee, S., Murray, J. A. H. and Palme, K. (2008) Comprehensive transcriptome analysis of auxin responses in *Arabidopsis*. *Molecular Plant* **1**, 2, 321–337.

Park, C. C., Gale, G. D., de Jong, S., Ghazalpour, A., Bennett, B. J., Farber, C. R., Langfelder, P., Lin, A., Khan, A. H., Eskin, E., Horvath, S., Lusis, A. J., Ophoff, R. A. and Smith, D. J. (2011) Gene networks associated with conditional fear in mice identified using a systems genetics approach. *BMC Systems Biology* **5**, 43.

Peidis, P., Giannakouros, T., Burow, M. E., Williams, R. W. and Scott, R. E. (2010) Systems genetics analyses predict a transcription role for P2P-R: molecular confirmation that P2P-R is a transcriptional co-repressor. *BMC Systems Biology* **4**, 14.

Pickel, B. and Schaller, A. (2013) Dirigent proteins: molecular characteristics and potential biotechnological applications. *Applied Microbiology and Biotechnology* **97**, 19, 8427–8438.

Poland, J. A., Balint-Kurti, P. J., Wisser, R. J., Pratt, R. C. and Nelson, R. J. (2009) Shades of gray: the world of quantitative disease resistance. *Trends in Plant Science* **14**, 1, 21–29.

Potokina, E., Druka, A., Luo, Z., Wise, R., Waugh, R. and Kearsey, M. (2008) Gene expression quantitative trait locus analysis of 16,000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant Journal* **53**, 1, 90–101.

Poulsen, L., Sø e, M. J., Snakenborg, D., Mø ller, L. B. and Dufva, M. (2008) Multi-stringency wash of partially hybridized 60-mer probes reveals that the stringency along the probe decreases with distance from the microarray surface. *Nucleic Acids Research* **36**, 20, e132.

Pré, M., Atallah, M., Champion, A., De Vos, M., Pieterse, C. M. J. and Memelink, J. (2008) The AP2/ERF domain transcription factor ORA59 integrates jasmonic acid and ethylene signals in plant defense. *Plant Physiology* **147**, 3, 1347–1357.

Punja, Z. K. and Zhang, Y. Y. (1993) Plant chitinases and their roles in resistance to fungal diseases. *Journal of Nematology* **25**, 4, 526–540.

Putilina, T., Wong, P. and Gentleman, S. (1999) The DHHC domain: a new highly conserved cysteine-rich motif. *Molecular and Cellular Biochemistry* **195**, 219–226.

Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* **5**, 2, 94–100.

Rajhi, I., Yamauchi, T., Takahashi, H., Nishiuchi, S., Shiono, K., Watanabe, R., Mliki, A., Nagamura, Y., Tsutsumi, N., Nishizawa, N. K. and Nakazono, M. (2011) Identification of genes expressed in maize root cortical cells during lysigenous aerenchyma formation using laser microdissection and microarray analyses. *New Phytologist* **190**, 351–368.

Ramm, M., Dangoor, K. and Sayfan, G. (2006) *Rapid web applications with TurboGears: Using Python to create Ajax-powered sites.* Prentice Hall PTR Upper Saddle River, NJ, USA.

Ranty, B., Aldon, D. and Galaud, J.-P. (2006) Plant calmodulins and calmodulin-related proteins. *Plant Signaling & Behavior* **1**, 3, 96–104.

Ravasz, E., Somera, A., Mongru, D., Oltvai, Z. and Barabási, A.-L. (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297**, 5586, 1551–1555.

Reddy, A. S. N., Ali, G. S., Celesnik, H. and Day, I. S. (2011) Coping with stresses: roles of calcium- and calcium/calmodulin-regulated gene expression. *The Plant Cell* **23**, 6, 2010–2032.

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., Kresovich, S., Goodman, M. M. and Buckler, E. S. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences* **98**, 20, 11479–11484.

Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273**, 5281, 1516–1517.

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A. and Smyth, G. K. (2007) A comparison of background correction methods for two-colour microarrays. *Bioinformatics* **23**, 20, 2700–2707.

Robert-seilaniantz, A., Grant, M. and Jones, J. D. G. (2011) Hormone crosstalk in plant disease and defense: more than just JASMONATE-SALICYLATE antagonism. *Annual Review of Phytopathology* **49**, 317–343.

Robert-seilaniantz, A., Navarro, L., Bari, R. and Jones, J. D. G. (2007) Pathological hormone imbalances. *Current Opinion in Plant Biology* **10**, 372–379.

Ronceret, A., Gadea-Vacas, J., Guilleminot, J. and Devic, M. (2008) The alpha-N-acetylglucosaminidase gene is transcriptionally activated in male and female gametes prior to fertilization and is essential for seed development in *Arabidopsis*. *Journal of Experimental Botany* **59**, 13, 3649–3659.

Rouhier, N., Lemaire, S. D. and Jacquot, J.-P. (2008) The role of glutathione in photosynthetic organisms: emerging functions for glutaredoxins and glutathionylation. *Annual Review of Plant Biology* **59**, 143–166.

Ryals, J. A., Neuenschwander, U. H., Willits, M. G., Molina, A., Steiner, H. Y. and Hunt, M. D. (1996) Systemic acquired resistance. *The Plant Cell* **8**, 10, 1809–1819.

Saghai Maroof, M. A., Yue, Y. G., Xiang, Z. X., Stromberg, E. L. and Rufener, G. K. (1996) Identification of quantitative trait loci controlling resistance to gray leaf spot disease in maize. *Theoretical and Applied Genetics* **93**, 539–546.

Salvi, S., Sponza, G., Morgante, M., Tomes, D., Niu, X., Fengler, K., Meeley, R., Ananiev, E. V., Svitashev, S., Bruggemann, E., Li, B., Hainey, C. F., Radovic, S., Zaina, G., Rafalski, J.-A., Tingey, S. V., Miao, G.-H., Phillips, R. L. and Tuberosa, R. (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proceedings of the National Academy of Sciences* **104**, 27, 11376–11381.

Salvi, S. and Tuberosa, R. (2005) To clone or not to clone plant QTLs: present and future challenges. *Trends in Plant Science* **10**, 6, 297–304.

Schadt, E. E., Monks, S. A., Drake, T. A., Lusisk, A. J., Chek, N., Colinayok, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B. and Friend, S. H. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 279–302.

Schapire, A. L., Valpuesta, V. and Botella, M. A. (2006) TPR Proteins in Plant Hormone Signaling Arnaldo. *Plant Signaling & Behavior* **1**, 5, 229–230.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J., Deragon, J., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A. and Wilson, R. K. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 5956, 1112–1115.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M. and Ragg, T. (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology* **7**, 3.

Sekhon, R. S., Lin, H., Childs, K. L., Hansey, C. N., Robin Buell, C., de Leon, N. and Kaeppler, S. M. (2011) Genome-wide atlas of transcription through maize development. *Plant Journal* **66**, 4, 553–563.

Sen, T. Z., Andorf, C. M., Schaeffer, M. L., Harper, L. C., Sparks, M. E., Duvick, J., Brendel, V. P., Cannon, E., Campbell, D. A. and Lawrence, C. J. (2009) MaizeGDB becomes 'sequence-centric'. *Database* bap020.

Sen, T. Z., Harper, L. C., Schaeffer, M. L., Andorf, C. M., Seigfried, T. E., Campbell, D. a. and Lawrence, C. J. (2010) Choosing a genome browser for a model organism database: surveying the maize community. *Database* baq007.

Sera, T. and Wolffe, A. P. (1998) Role of histone H1 as an architectural determinant of chromatin structure and as a specific repressor of transcription on *Xenopus* oocyte 5S rRNA genes. *Molecular and Cellular Biology* **18**, 7, 3668–3680.

Sessa, G. and Martin, G. B. (2000) Protein kinases in the plant defense response. *Plant Pathology* **32**, 379–404.

Shaik, R. and Ramakrishna, W. (2013) Genes and co-expression modules common to drought and bacterial stress responses in *Arabidopsis* and rice. *PloS One* **8**, 10, e77261.

Shi, C., Uzarowska, A., Ouzunova, M., Landbeck, M., Wenzel, G. and Lübberstedt, T. (2007) Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint x Flint maize recombinant inbred line population. *BMC Genomics* **8**, 22.

Shi, H., Liu, Z., Zhu, L., Zhang, C., Chen, Y., Zhou, Y., Li, F. and Li, X. (2012) Overexpression of cotton (*Gossypium hirsutum*) *dirigent1* gene enhances lignification that blocks the spread of *Verticillium dahliae*. *Acta Biochimica et Biophysica Sinica* **44**, 7, 555–564.

Sieberts, S. K. and Schadt, E. E. (2007) Moving toward a system genetics view of disease. *Mammalian Genome* **18**, 389–401.

Singh, S., Tan, H. Q. and Singh, J. (2012) Mutagenesis of barley malting quality QTLs with Ds transposons. *Functional & Integrative Genomics* **12**, 1, 131–141.

Skibbe, D. S., Fernandes, J. F., Medzihradszky, K. F., Burlingame, A. L. and Walbot, V. (2009) Mutator transposon activity reprograms the transcriptomes and proteomes of developing maize anthers. *Plant Journal* **59**, 4, 622–633.

Slater, G. S. C. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. and Ideker, T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 3, 431–432.

Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, 3.

Soller, M. and Beckmann, J. S. (1987) Cloning quantitative trait loci by insertional mutagenesis. *Theoretical and Applied Genetics* **74**, 369–378.

Solomon, M., Belenghi, B., Delledonne, M., Menachem, E. and Levine, A. (1999) The involvement of cysteine proteases and protease inhibitor genes in the regulation of programmed cell death in plants. *The Plant Cell* **11**, 431–444.

Song, W.-Y., Hörtensteiner, S., Tomioka, R., Lee, Y. and Martinoia, E. (2011) Common functions or only phylogenetically related? The large family of PLAC8 motif-containing/PCR genes. *Molecules and Cells* **31**, 1–7.

Song, W.-Y., Martinoia, E., Lee, J., Kim, D., Kim, D.-Y., Vogt, E., Shim, D., Choi, K. S., Hwang, I. and Lee, Y. (2004) A Novel Family of Cys-Rich Membrane Proteins Mediates Cadmium Resistance in *Arabidopsis*. *Plant Physiology* **135**, 1027–1039.

Spassieva, S. D., Markham, J. E. and Hille, J. (2002) The plant disease resistance gene Asc-1 prevents disruption of sphingolipid metabolism during AAL-toxin-induced programmed cell death. *The Plant Journal* **32**, 561–572.

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., Iniguez, A. L., Barbazuk, W. B., Jeddeloh, J. a., Nettleton, D. and Schnable, P. S. (2009) Maize inbreds exhibit high levels of copy

number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genetics* **5**, 11, e1000734.

Steinbrenner, A., Goritschnig, S., Krasileva, K., Schreiber, K. and Staskawicz, B. (2012) Effector recognition and activation of the *Arabidopsis thaliana* NLR innate immune receptors. *Cold Spring Harbor Symposia on Quantitative Biology* **77**, 249–257.

Steinhauser, D., Usadel, B., Luedemann, A., Thimm, O. and Kopka, J. (2004) CSB.DB: a comprehensive systems-biology database. *Bioinformatics* **20**, 18, 3647–3651.

Stone, E. A. and Ayroles, J. F. (2009) Modulated modularity clustering as an exploratory tool for functional genomic inference. *PLoS Genetics* **5**, 5, e1000479.

Stukkens, Y., Bultreys, A., Trombik, T., Vanham, D. and Boutry, M. (2005) NpPDR1, a pleiotropic drug resistance-type ATP-binding cassette transporter from *Nicotiana plumbaginifolia*, plays a major role in plant pathogen defense. *Plant Physiology* **139**, 341–352.

Sun, W. and Hu, Y. (2012) eQTL mapping using RNA-seq data. *Statistics in Biosciences* **5**, 1, 198–219.

Sun, W., Yu, T. and Li, K.-C. (2007) Detection of eQTL modules mediated by activity levels of transcription factors. *Bioinformatics* **23**, 17, 2290–2297.

Swanson-Wagner, R. A., DeCook, R., Jia, Y., Bancroft, T., Ji, T., Zhao, X., Nettleton, D. and Schnable, P. S. (2009) Paternal dominance of *trans*-eQTL influences gene expression patterns in maize hybrids. *Science* **326**, 5956, 1118–1120.

Sweetman, C., Drew, D. P. and Ford, C. M. (2013) VTCdb: a gene co-expression database for the crop species *Vitis vinifera* (grapevine). *BMC Genomics* **14**, 882.

Takahara, M., Magori, S., Soyano, T., Okamoto, S., Yoshida, C., Yano, K., Sato, S., Tabata, S., Yamaguchi, K., Shigenobu, S., Takeda, N., Suzaki, T. and Kawaguchi, M. (2013) Too much love, a novel Kelch repeat-containing F-box protein, functions in the long-distance regulation of the legume-Rhizobium symbiosis. *Plant & Cell Physiology* **54**, 4, 433–447.

Tan, L., Eberhard, S., Pattathil, S., Warder, C., Glushka, J., Yuan, C., Hao, Z., Zhu, X., Avci, U., Miller, J. S., Baldwin, D., Pham, C., Orlando, R., Darvill, A., Hahn, M. G., Kieliszewski, M. J. and Mohnen, D. (2013) An *Arabidopsis* cell wall proteoglycan consists of pectin and arabinoxylan covalently linked to an arabinogalactan protein. *The Plant Cell* **25**, 1, 270–287.

Tan, S. and Wu, S. (2012) Genome wide analysis of nucleotide-binding site disease resistance genes in *Brachypodium distachyon Comparative and Functional Genomics*.

Tavares, B., Domingos, P., Dias, P. N., Feijó, J. A. and Bicho, A. (2011) The essential role of anionic transport in plant cells: the pollen tube as a case study. *Journal of Experimental Botany* **62**, 7, 2273–2798.

Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F. and Gaut, B. S. (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proceedings of the National Academy of Sciences* **98**, 16, 9161–9166.

Terpstra, I. R., Snoek, L. B., Keurentjes, J. J. B., Peeters, A. J. M. and van den Ackerveken, G. (2010) Regulatory network identification by genetical genomics: signaling downstream of the *Arabidopsis* receptor-like kinase ERECTA. *Plant Physiology* **154**, 3, 1067–1078.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L. a., Rhee, S. Y. and Stitt, M. (2004) Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal* **37**, 6, 914–939.

Thomma, B. P., Eggermont, K., Penninckx, I. A., Mauch-Mani, B., Vogelsang, R., Cammue, B. P. and Broekaert, W. F. (1998) Separate jasmonate-dependent and salicylate-dependent defense-response pathways in *Arabidopsis* are essential for resistance to distinct microbial pathogens. *Proceedings of the National Academy of Sciences* **95**, 25, 15107–15111.

Thorson, P. R. and Martinson, C. A. (1993) Development and survival of *Cercospora zeae-maydis* germlings in different relative-humidity environments. *Phytopathology* **83**, 153–157.

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T. R., McMullen, M. D., Holland, J. B. and Buckler, E. S. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetics* **43**, 2, 159–162.

Tian, W., Zhang, L. V., Ta, M., Gibbons, F. D., King, O. D., Park, J., Wunderlich, Z., Cherry, J. M. and Roth, F. P. (2008) Combining guilt-by-association and guilt-by-profiling to predict *Saccharomyces cerevisiae* gene function. *Genome Biology* **9**, 1, S7.

Tohge, T., Watanabe, M., Hoefgen, R. and Fernie, A. R. (2013) Shikimate and phenylalanine biosynthesis in the green lineage. *Frontiers in Plant Science* **4**, 1–13.

Torii, K. U., Mitsukawa, N., Oosumi, T., Matsuura, Y., Yokoyama, R., Whittier, R. F. and Komeda, Y. (1996) The *Arabidopsis* ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. *The Plant Cell* **8**, 735–746.

Toufighi, K., Brady, S. M., Austin, R., Ly, E. and Provart, N. J. (2005) The Botany Array Resource: e-northerns, expression angling, and promoter analyses. *The Plant Journal* **43**, 153–163.

Trognitz, F., Manosalva, P., Gysin, R., Niñio Liu, D., Simon, R., del Herrera, M. R., Trognitz, B., Ghislain, M. and Nelson, R. (2002) Plant defense genes associated with quantitative resistance to potato late blight in *Solanum phureja* x dihaploid *S. tuberosum* hybrids. *Molecular Plant-Microbe Interactions* **15**, 6, 587–597.

Trujillo, M. and Shirasu, K. (2010) Ubiquitination in plant immunity. *Current Opinion in Plant Biology* **13**, 4, 402–408.

van den Burg, H. A., Harrison, S. J., Joosten, M. H. A. J., Vervoort, J. and de Wit, P. J. G. M. (2006) *Cladosporium fulvum* Avr4 protects fungal cell walls against hydrolysis by plant chitinases accumulating during infection. *Molecular Plant-Microbe Interactions* **19**, 12, 1420–1430.

van den Burg, H. A., Spronk, C. A. E. M., Boeren, S., Kennedy, M. A., Vissers, J. P. C., Vuister, G. W., de Wit, P. J. G. M. and Vervoort, J. (2004) Binding of the AVR4 elicitor of *Cladosporium fulvum* to chitotriose units is facilitated by positive allosteric

protein-protein interactions: the chitin-binding site of AVR4 represents a novel binding site on the folding scaffold shared between the invertebrate and the plant chitin-binding domain. *The Journal of Biological Chemistry* **279**, 16, 16786–16796.

van den Burg, H. A., Tsitsigiannis, D. I., Rowland, O., Lo, J., Rallapalli, G., Maclean, D., Takken, F. L. W. and Jones, J. D. G. (2008) The F-Box protein ACRE189/ACIF1 regulates cell death and defense responses activated during pathogen recognition in tobacco and tomato. *The Plant Cell* **20**, 697–719.

van der Biezen, E. A. and Jones, J. D. (1998) Plant disease-resistance proteins and the gene-for-gene concept. *Trends in Biochemical Sciences* **23**, 12, 454–456.

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. A., Bink, M. C., 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 Loon, L. C., Rep, M. and Pieterse, C. M. J. (2006) Significance of inducible defense-related proteins in infected plants. *Annual Review of Phytopathology* **44**, 135–162.

van Ooijen, J. W. (2006) *JoinMap 4.0, Software for the calculation of genetic linkage maps in experimental populations.* Kyazma B.V., Wageningen, Netherlands.

van Ooijen, J. W. (2009) *MAPQTL 6.0, Software for the mapping of quantitative trait loci in experimental populations of diploid species.* Kyazma B.V., Wageningen, Netherlands.

Villa-Vialaneix, N., Liaubet, L., Laurent, T., Cherel, P., Gamot, A. and SanCristobal, M. (2013) The structure of a gene co-expression network reveals biological functions underlying eQTLs. *PloS One* **8**, 4, e60045.

Voitsik, A.-M., Muench, S., Deising, H. B. and Voll, L. M. (2013) Two recently duplicated maize NAC transcription factor paralogs are induced in response to *Colletotrichum graminicola* infection. *BMC Plant Biology* **13**, 85.

Voorrips, R. E. (2001) MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. *Journal of Heredity* **93**, 1, 77–78.

Wang, D., Oses-Prieto, J. A., Li, K. H., Fernandes, J. F., Burlingame, A. L. and Walbot, V. (2010*a*) The male sterile 8 mutation of maize disrupts the temporal progression of the transcriptome and results in the mis-regulation of metabolic functions. *Plant Journal* **63**, 6, 939–951.

Wang, H., Nussbaum-Wagler, T., Li, B., Zhao, Q., Vigouroux, Y., Faller, M., Bomblies, K., Lukens, L. and Doebley, J. F. (2005) The origin of the naked grains of maize. *Nature* **436**, 714–719.

Wang, J., Levy, M. and Dunkle, L. D. (1998) Sibling species of *cercospora* associated with gray leaf spot of maize. *Phytopathology* **88**, 12, 1269–1275.

Wang, J., Williams, R. W. and Manly, K. F. (2003) WebQTL: Web-based complex trait analysis. *Neuroinformatics* **1**, 4, 299–308.

Wang, J., Yu, H., Weng, X., Xie, W., Xu, C., Li, X., Xiao, J. and Zhang, Q. (2014) An expression quantitative trait loci-guided co-expression analysis for constructing regulatory network using a rice recombinant inbred line population. *Journal of Experimental Botany* **65**, 4, 1069–1079.

Wang, J., Yu, H., Xie, W., Xing, Y., Yu, S., Xu, C., Li, X., Xiao, J. and Zhang, Q. (2010*b*) A global analysis of QTLs for expression variations in rice shoots at the early seedling stage. *The Plant Journal* **63**, 6, 1063–1074.

Wang, S., Basten, C. J. and Zeng, Z.-B. (2012*a*) *Windows QTL Cartographer 2.5.* Department of Statistics, North Carolina State University, Raleigh, NC.

Wang, S., Yin, Y., Ma, Q., Tang, X., Hao, D. and Xu, Y. (2012*b*) Genome-scale identification of cell-wall related genes in *Arabidopsis* based on co-expression network analysis. *BMC Plant Biology* **12**, 138.

Wang, S., Zheng, T. and Wang, Y. (2007) Transcription activity hot spot, is it real or an artifact? *BMC Proceedings* **1**, Suppl 1, S94.

Wang, X., Basnayake, B. M. V. S., Zhang, H., Li, G., Li, W., Virk, N., Mengiste, T. and Song, F. (2009) The *Arabidopsis* ATAF1, a NAC transcription factor, is a negative regulator of defense responses against necrotrophic fungal and bacterial pathogens. *Molecular Plant-Microbe Interaction* **22**, 10, 1227–1238.

Ward, J. and Nowell, D. (1998) Integrated management practices for the control of maize grey leaf spot. *Integrated Pest Management Reviews* **3**, 177–188.

Ward, J. M. J., Stromberg, E. L., Nowell, D. C. and Nutter, F. W. (1999) Gray leaf spot: A disease of global importance in maize production. *Plant Disease* **83**, 884–895.

Waterhouse, P. M. and Helliwell, C. A. (2003) Exploring plant genomes by RNA-induced gene silencing. *Nature Reviews Genetics* **4**, 1, 29–38.

Wei, F., Stein, J. C., Liang, C., Zhang, J., Fulton, R. S., Baucom, R. S., De Paoli, E., Zhou, S., Yang, L., Han, Y., Pasternak, S., Narechania, A., Zhang, L., Yeh, C.-T., Ying, K., Nagel, D. H., Collura, K., Kudrna, D., Currie, J., Lin, J., Kim, H., Angelova, A., Scara, G., Wissotski, M., Golser, W., Courtney, L., Kruchowski, S., Graves, T. A., Rock, S. M., Adams, S., Fulton, L. a., Fronick, C., Courtney, W., Kramer, M., Spiegel, L., Nascimento, L., Kalyanaraman, A., Chaparro, C., Deragon, J.-M., Miguel, P. S., Jiang, N., Wessler, S. R., Green, P. J., Yu, Y., Schwartz, D. C., Meyers, B. C., Bennetzen, J. L., Martienssen, R. A., McCombie, W. R., Aluru, S., Clifton, S. W., Schnable, P. S., Ware, D., Wilson, R. K. and Wing, R. A. (2009*a*) Detailed analysis of a contiguous 22-Mb region of the maize genome. *PLoS Genetics* **5**, 11, e1000728.

Wei, F., Zhang, J., Zhou, S., He, R., Schaeffer, M., Collura, K., Kudrna, D., Faga, B. P., Wissotski, M., Golser, W., Rock, S. M., Graves, T. A., Fulton, R. S., Coe, E., Schnable, P. S., Schwartz, D. C., Ware, D., Clifton, S. W., Wilson, R. K. and Wing, R. A. (2009*b*) The physical and genetic framework of the maize B73 genome. *PLoS Genetics* **5**, 11, e1000715.

West, M. A. L., Kim, K., Kliebenstein, D. J., van Leeuwen, H., Michelmore, R. W., Doerge, R. W. and St Clair, D. A. (2007) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* **175**, 3, 1441–1450.

Winkel-Shirley, B. (2002) Biosynthesis of flavonoids and effects of stress. *Current Opinion in Plant Biology* **5**, 218–223.

Wisser, R. J., Balint-Kurti, P. J. and Nelson, R. J. (2006) The genetic architecture of disease resistance in maize: a synthesis of published studies. *Phytopathology* **96**, 2, 120–129.

Wisser, R. J., Kolkman, J. M., Patzoldt, M. E., Holland, J. B., Yu, J., Krakowsky, M., Nelson, R. J. and Balint-Kurti, P. J. (2011) Multivariate analysis of maize disease resistances suggests a pleiotropic genetic basis and implicates a GST gene. *Proceedings of the National Academy of Sciences* **108**, 18, 7339–7344.

Wright, F. A., Shabalin, A. A. and Rusyn, I. (2012) Computational tools for discovery and interpretation of expression quantitative trait loci. *Pharmacogenomics* **13**, 3, 343–352.

Xie, D.-X., Feys, B. F., James, S., Nieto-Rostro, M. and Turner, J. G. (1998) COI1: an *Arabidopsis* gene required for jasmonate-regulated defense and fertility. *Science* **280**, 5366, 1091–1094.

Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R., Heuer, S., Ismail, A. M., Bailey-Serres, J., Ronald, P. C. and Mackill, D. J. (2006) Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* **442**, 705–708.

Yabuta, T. and Sumiki, Y. (1983) On the crystal of gibberellin, a substance to promote plant growth. *Journal of the Agricultural Chemical Society of Japan* **14**, 1526.

Yamaguchi, K., Yamada, K., Ishikawa, K., Yoshimura, S., Hayashi, N., Uchihashi, K., Ishihama, N., Kishi-kaboshi, M., Takahashi, A., Tsuge, S., Ochiai, H. and Tada, Y. (2013) A receptor-like cytoplasmic kinase targeted by a plant pathogen effector Is directly phosphorylated by the chitin receptor and mediates rice immunity. *Cell Host and Microbe* **13**, 3, 347–357.

Yan, J., Zhang, C., Gu, M., Bai, Z., Zhang, W., Qi, T., Cheng, Z., Peng, W., Luo, H., Nan, F., Wang, Z. and Xie, D. (2009) The *Arabidopsis* CORONATINE INSENSITIVE1 protein is a jasmonate receptor. *The Plant Cell* **21**, 8, 2220–2236.

Yang, D.-L., Li, Q., Deng, Y.-W., Lou, Y.-G., Wang, M.-Y., Zhou, G.-X., Zhang, Y.-Y. and He, Z.-H. (2008) Altered disease development in the *eui* mutants and *Eui* overexpressors indicates that gibberellins negatively regulate rice basal disease resistance. *Molecular Plant* **1**, 3, 528–537.

Yang, J., Kong, L., Chen, X., Wang, D., Qi, L., Zhao, W., Zhang, Y., Liu, X. and Peng, Y.-L. (2012) A carnitine-acylcarnitine carrier protein, MoCrc1, is essential for pathogenicity in *Magnaporthe oryzae*. *Current Genetics* **58**, 3, 139–148.

Yang, Y. H. and Thome, N. P. (2003) Normalization for two-color cDNA microarray data in Goldstein, D. R., editor, *Statistics and science: a Festschrift for Terry Speed* 403–418 Institute of Mathematical Statistics, Beachwood, OH.

Young, N. D. (2000) The genetic architecture of resistance. *Current Opinion in Plant Biology* **3**, 285–290.

Yu, J., Holland, J. B., McMullen, M. D. and Buckler, E. S. (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**, 1, 539–551.

Zeng, Z. B. (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences* **90**, 23, 10972–10976.

Zhang, B. and Horvath, S. (2005) A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* **4**, 17.

Zhang, G., Sun, Y. F., Li, Y. M., Dong, Y. L., Huang, X. L., Yu, Y. T., Wang, J. M., Wang, X. M., Wang, X. J. and Kang, Z. S. (2013) Characterization of a wheat C2 domain protein encoding gene regulated by stripe rust and abiotic stresses. *Biologia Plantarum* **57**, 4, 701–710.

Zhang, H., Gao, Z., Zheng, X. and Zhang, Z. (2012*a*) The role of G-proteins in plant immunity. *Plant Signaling & Behavior* **7**, 10, 1284–1288.

Zhang, J., Xiang, Y., Ding, L., Keen-Circle, K., Borlawsky, T. B., Ozer, H. G., Jin, R., Payne, P. and Huang, K. (2010) Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia. *BMC Bioinformatics* **11**, Suppl 9, S5.

Zhang, X., Cal, A. J. and Borevitz, J. O. (2011) Genetic architecture of regulatory variation in *Arabidopsis thaliana*. *Genome Research* **21**, 5, 725–733.

Zhang, Y., Xu, L., Fan, X., Tan, J., Chen, W. and Xu, M. (2012*b*) QTL mapping of resistance to gray leaf spot in maize. *Theoretical and Applied Genetics*.

Zhu, C., Gore, M., Buckler, E. S. and Yu, J. (2008*a*) Status and prospects of association mapping in plants. *The Plant Genome Journal* **1**, 1, 5–20.

Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., Bumgarner, R. E. and Schadt, E. E. (2008*b*) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics* **40**, 7, 854–861.

Zhu, Z., Xu, F., Zhang, Y., Ti, Y., Wiermer, M., Li, X. and Zhang, Y. (2010) Arabidopsis resistance protein SNC1 activates immune responses through association with a transcriptional corepressor. *Proceedings of the National Academy of Sciences* **107**, 31, 13960–13965.

Zimmermann, P., Hirsch-hoffmann, M., Hennig, L. and Gruissem, W. (2004) GENEVESTIGATOR. *Arabidopsis* microarray database and analysis toolbox. *Bioinformatics* **136**, 2621–2632.

Zou, W., Aylor, D. L. and Zeng, Z.-B. (2007) eQTL Viewer: visualizing how sequence variation affects genome-wide transcription. *BMC Bioinformatics* **8**, 7.

Zou, W. and Zeng, Z.-B. (2008) Statistical methods for mapping multiple QTL. *International Journal of Plant Genomics* **108**, 36, 14992–14997.