High-throughput computational methods and software for quantitative trait locus (QTL) mapping

Arends, Danny

High-throughput computational methods and software for quantitative trait locus (QTL) mapping

The work described in this thesis was carried out at the Groningen Bioinformatics Centre, University of Groningen, The Netherlands. This research was financially supported by the Centre for BioSystems Genomics (CBSG) and the Netherlands Consortium of Systems Biology (NCSB), both of which are part of the Netherlands Genomics Initiative / Netherlands Organisation for Scientific Research. Printing of this thesis was financially supported by the University of Groningen.

Artwork by: Anna Mulder
Printed by: RCG Grafimedia - Groningen

ISBN (Print): 978-90-367-7210-5
ISBN (Digital): 978-90-367-7209-9

rijksuniversiteit
groningen

High-throughput computational methods and software for quantitative trait locus (QTL) mapping

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

vrijdag 17 oktober 2014 om 16.15 uur

door

Derk Arends

geboren op 15 juli 1983
te Zwolle

Promotor
Prof. dr. R. C. Jansen

Copromotor
Dr. M. A. Swertz

Beoordelingscommissie
Prof. dr. E. O. de Brock
Prof. dr. G. A. Brockmann
Prof. dr. M. H. Hofker

voor mama

Wie legt me uit hoe alles werkt, hoe groot het gat is tussen nu en nooit
En hoe het komt dat ik nu merk, jij bent weg maar dichterbij dan ooit

Bløf – "Dichterbij dan ooit"

Contents

Summary **10**

1 Introduction **12**
1.1 From pea plants to 'Big Data' 13
1.2 Approaches 15
1.3 The first genetic theories (1865-1930) 17
1.4 DNA and QTLs (1930-1990) 19
1.5 From phenotypes to genetical omics and GWAS (1990-2010) 21
1.6 Cluster and cloud computing (since 1980) 24

2 High-throughput generation of genetic markers **26**
2.1 Pheno2Geno 27
2.2 Background 28
2.3 Features 28
2.4 Results 30
2.5 Conclusions and discussion 31

3 High-throughput (Multiple) QTL mapping **37**
3.1 Single marker QTL mapping (R/qtl) 38
3.2 Multiple QTL mapping 38
3.2.1 Features 39
3.2.2 Conclusions and discussion 42
3.3 Mapping classical phenotypes 43
3.3.1 Background 43
3.3.2 Results 48
3.3.3 Conclusions and discussion 54
3.4 Metabolites in a designGG experiment 60
3.4.1 Background 60
3.4.2 Results 63

4 High-throughput QTL mapping using correlated traits 79
4.1 CTL mapping 80
4.1.1 Introduction 80
4.1.2 Calculating a CTL 81
4.1.3 Inference of a hierarchical relationship between traits 84
4.1.4 Power analysis 87
4.1.5 CTL power calculation applies to QTL and GWA 87
4.2 Cell type specific eQTL mapping in human GWAS data 88
4.2.1 Background 88
4.2.2 Method 89
4.2.3 Results 90
4.2.4 Discussion 94

5 High-throughput infrastructure for systems genetics 98
5.1 High-throughput data analysis (xQTL workbench) 100
5.1.1 Features 101
5.1.2 Reusable software (MOLGENIS) 102
5.1.3 Extensible genotype and phenotype data (XGAP) 107
5.2 A worm database (WormQTL) 111
5.2.1 Results 112
5.2.2 Extensions into WormQTL-HD 114
5.2.3 Conclusion 115

6 Conclusion, discussion and future perspectives 119
6.1 General discussion 120
6.2 Highlight of the results 120
6.3 Pheno2Geno performance 121
6.4 MQM revisited 123
6.5 Visualising the output 125
6.6 Limitations of R as an analysis platform for R/qtl 126
6.7 CTL & cell type specific eQTL 127
6.8 xQTL workbench for infrastructural issues 129
6.9 Applications in biology 132
6.10 Future perspectives 133

Contents

7 Additional for dissertation **136**

7.1 Dutch summary / Nederlandse samenvatting 137

7.2 Abbreviations and acronyms 139

7.3 Acknowledgements 141

7.4 About the author 144

7.4.1 Curriculum Vitae 144

7.4.2 List of publications 145

7.4.3 List of presentations 148

7.4.4 List of posters 149

7.4.5 Awards 150

Bibliography **151**

Summary

Systems genetics is the interdisciplinary field which deals with the consequences of genetic variation on all biomolecular levels of a biological system. The aim of systems genetics is to understand biological systems by partitioning variation into three major categories: genetic, environment and error variation, and explain how complex phenotypes arise from a combination of these three major factors across biomolecular levels.

Currently, naturally occurring genetic variation (or genetic pertubation) can be used to interrogate the genetic basis of phenotype variation on many biomolecular levels such as: genetics, transcriptomics, proteomics and metabolomics. Combined with environmental pertubation we can investigate the influence of different environments and the interaction between genetics and environment. Experimental design and advanced statistics are used to minimize and estimate error variance. To investigate all these factors influencing biological systems it is necessary to collect huge data sets on many individuals, many tissues, at all known biomolecular levels.

Modern high-throughput technologies generate large amounts of genomic, transcriptomic, proteomic and metabolomic data, creating a major challenge in bioinformatics because of the size of data collected and the multitude of technologies used. This thesis will highlight our solutions to the 'Big Data' challenges in systems genetics. We propose to develop smarter more optimized algorithms such as Pheno2Geno or Multiple QTL mapping, and to use a collaborative approach such as xQTL workbench to store and analyse high-throughput systems genetics data.

Chapter 1 contains an introduction to systems genetics, and highlights the challenges that inspired this thesis. These challenges, such as the massive increase in data production, cause an increasing complexity when diagnosing patients, or selecting crops for optimal yield.

Chapter 2 shows Pheno2Geno, an R package that deals with the creation of genetic maps from large scale omics data. The theory behind genetic map construction is around 100 years old. Most software was written in the 1980s, and software available for genetic map construction has not been adapted yet to make use of new technologies such as

multithreading or cluster computing. Pheno2Geno aims to provide analysis of data from tiling arrays and/or RNA-Seq to generate gene based expression markers (GEMs) and create high density genetic maps.

Chapter 3 describes the implementation of the Multiple QTL mapping (MQM) routine into R/qtl, adding a 'new' algorithm to the R/qtl toolset to provide a wider range of QTL mapping tools for inbred crosses. R/qtl is the basis of a toolset built around a unified data structure allowing easy adaptation and extension of the software. R/qtl allows researchers to analyse data from different sources, and to quickly compare different approaches. This chapter showcases our contributions to the R/qtl package such as: MQM, visualizations, parallel computation of QTLs and an improved permutation scheme.

Chapter 4 describes our current work on using differences in correlation to generate interaction networks and detect cell type specific QTL effects. Correlated Traits Locus analysis (or CTL mapping) enables researchers to find genetic loci controlling correlation differences in segregating phenotypes. A variation on this method has proven valuable in discovering cell type specific eQTL effects. Using these effects it is possible to untangle cell mixtures seen in whole blood.

Chapter 5 details our work to provide computational infrastructure for the Life Sciences. Our system xQTL workbench is currently being used as a back end to the WormQTL and WormQTL-HD database. xQTL workbench allows users to store and share their data in a local or web environment, and run analysis across data sets using the power of distributed computing. It comes standard with QTL mapping tools such as: R/qtl, PLINK and qtlbim but also provides a web inferface, data importers, APIs and visualizations.

I trust you enjoy reading this thesis as much as I enjoyed creating it during the last four years,

Danny Arends (August 2014)

1 Introduction

This first chapter introduces the reader to concepts such as gene, chromosome and DNA all the way up to Multiple QTL mapping (MQM) and Genome Wide Association Studies (GWAS).

1.1 From pea plants to 'Big Data'

Since the first genetic theories on the inheritance of phenotypes in pea plants were published by Gregor Mendel in 1865, the ongoing interest in the field has now resulted in an unprecedented stream of data to study biology. The knowledge gain from genetics research is huge as it has revealed how complex biology actually is.

Genetics research has shown how the activity of DNA is regulated by signals from all molecular levels, i.e. the genome, transcriptome, proteome, and metabolome. All these levels interact, and are modified by signals from the environment. Adding to the complexity, every level in the genetic paradigm has its own chemistry, functionality, and research technologies (Table 1.1). Systems genetics is the research field in which we consider all these levels and interpret them together, within the context of DNA. Systems genetics aims to find the DNA variants underlying the variation that is seen on the higher molecular levels, all the way up to the phenotypes, for example seed quality in plants, and Crohn's disease in humans.

The research fields of genetics and bioinformatics have been intertwined for several decades. Since the 1980s it has become impossible to analyse data without using some form of computational assistance. This forces geneticists either to become 'computer savvy' or to collaborate with bioinformaticians. In practice, genetics research has

Molecular level	**Molecule**	**Technology**
Genome	DNA	RFLP [1]
Genome	DNA	SNP chips [2]
Genome	DNA	DNA sequencing [3]
EpiGenome	DNA methylation	Bisulfite sequencing [4]
EpiGenome	DNA methylation	ChIP-on-chip [5]
EpiGenome	DNA methylation	ChIP-Seq [6]
Transcriptome	RNA	Microarray [7]
Transcriptome	RNA	Tiling array [8]
Transcriptome	RNA	RNA-Seq [9]
Proteome	Proteins	2D gel electrophoresis [10]
Proteome	Proteins	Mass Spectometry [11]
Proteome	Proteins	Antibody protein chip [12]
Metabolome	Metabolites	Mass Spectometry [13]
Metabolome	Metabolites	Nuclear magnetic resonance [14]

***Table 1.1** – Overview of current technologies to interrogate the molecular basis for life.*

evolved into an interdisciplinary research domain. This thesis enters the field from the bioinformatics perspective.

The ongoing stream of data that is being produced in the field of systems genetics poses several challenges to bioinformaticians. Data storage, computation on the data and visualising the results become increasingly harder when data sizes increase. For example, a clinician who has to deal with gigabytes of data while diagnosing a patient, requires access to good software. On a research level, understanding how a system works and reacts to different environments allows modification of the system to better suit our needs and environmental requirements. The expected impact of such knowledge about DNA variation is to optimise the yield of livestock and crops, explore the genetic basis of human diseases, improve the production of chemicals, and so forth. The use of computational tools in understanding how genetics and the environment intertwine is thus central to the field of systems genetics and forms the basis how to deal with the large amounts of phenotype and genotype data. How we transform such data into usable knowledge is the main theme of this thesis.

Currently new data types are coming in more often than before. Previously geneticists dealt with mostly classical phenotypes and a limited number of genetic markers. Now we are able to map thousands of traits onto millions of markers, creating not only more data, but also a more diverse kind of biological data. These new data also need to be considered when analysing and interpreting the outcome in terms of clinical and societal relevance. To merge all these inputs and provide algorithms with a unified view on such diverse data, generic data modeling and generators that use these models to create *de novo* (web)-infrastructure are required.

The phenomenon of larger sample sizes, more data collection and more intense computations is a natural consequence of the data driven research which has increasingly become the focus of systems genetics. Data driven research has caused the issue that we now call ‘Big Data’. Making sense of this ‘Big Data’ is increasingly difficult as it becomes harder to find clinical and societal relevant conclusions, while requiring ever larger computer facilities. This leads to increased societal costs, most notable in healthcare and industry. In the context of all these developments, it is important to consider more efficient usage of available resources, and avoid duplicate work effort. To contribute to this challenge we formed the following central question in this thesis:

How can computational bioinformatics help solve current ‘Big Data’ challenges in systems genetics?

The ‘Big Data’ issues that will be discussed in this thesis are:

- The increasing amount of time needed for computation on ‘Big Data’.
- Increased memory requirements for several algorithms with increased complexity needed for ‘Big Data’.
- Numerous big data sets collected on multiple species.
- Complex software for distributed computing which increases the number of steps needed for computation.

Furthermore we also focus on the next generation of statistical analysis such as Multiple QTL mapping, celltype specific eQTLs and detecting differences in correlation using a genome wide approach comparable to QTL mapping.

1.2 Approaches

Recent developments in computational infrastructure have created many new possibilities to deal with complex data in a powerful way. The impact of these developments is that all kinds of genetic concepts can now be tested. For example, 20 years ago the concept of Multiple QTL Mapping (MQM) was developed. This technique aims to compare all genetic markers to each other. At the time when MQM was conceptualised, it was computational very demanding. Because of the current increase of computational power, we are able to revisit MQM’s methodological challenges and demonstrate that the use of high-throughput computational methodologies will allow us to solve research questions and do this in reasonable time.

QTL mapping associates phenotype variation with genetic variation using appropriate statistical methods, such as *t*-tests, ANOVA, and generalised linear models. MQM is a two-step procedure. First, it uses generalised linear models and backward selection of genetic cofactors to model the number of genetic factors underlying the phenotypes. In the second step, it uses the optimal model for QTL detection conditional to the selected cofactors. MQM is a superior technique compared to other methods for QTL mapping, as it allows detection of QTL in repulsion phase and performs with higher statistical power to detect phenotype to genotype associations. These advantageous characteristics are key to our decision to reintroduce the MQM algorithms into the bioinformatics toolbox.

R/qtl [15, 16] is an open source toolbox for mapping QTLs in inbred animals. It provides many QTL mapping routines and allows researchers to easily switch between statistical approaches, or test new approaches for QTL detection. Furthermore, it comes with a large user base and is de facto standard in the mouse QTL mapping community.

High-throughput phenotyping provides us with data on thousands of phenotypes. Some of these phenotypes can be used as genetic markers because they show a major QTL effect. This leads to a situation in which we have a genetic map and an abundance of phenotypes, which could be used to improve the map quality by pinpointing all informative recombinations. This knowledge, combined with a lack of good genetic map creation software in the R/qtl package, has fuelled the development of the sensitive algorithm Pheno2Geno. It uses mixture modeling combined with a sensitive pre-selection approach. This results in faster genetic map creation as compared to other methods, and is the main reason for our ambition to add Pheno2Geno to the R/qtl toolbox.

QTL interactions are an important area of interest in systems genetics. For example, in recent years several papers were published on cell type or sex specificity of QTLs. Detection of these cell type specific QTLs has proven to be extremely difficult because it requires large sample sizes. Meta-analysis, where we combine the samples from separate studies to improve power, is a prerequisite when studying cell type specific expression-QTLs (eQTLs). We noticed a lack of good software to meta-analyse these cell type specific QTLs in a genome wide setting. Furthermore, since cell type specific eQTL analysis is comparable to analysis that uses multiple environments, we propose to conceptualise cell abundance similar to an environment in which a QTL can be present or absent. Therefore, to consider such interactions, our toolbox also needs to be adjusted to support meta-analysis on genetical genomics experiments in which these QTL × environment interactions are included.

All these advances are impossible without good infrastructure to funnel all the data streams into a network or an image that a biologist can understand. We noticed many separate tools for handling different facets of this ‘Big Data’ issue. Software such as database systems, web frameworks, computational tools, and APIs are available, but take considerable time to setup and master. Furthermore we noticed a lack of comprehensive software which combines all these tools into a comprehensible system, which is easy to use for a biologist.

New developments are common in computational biology, and only the application of these new methodologies on biological data sets can show the value of the developed theories and software. Only when applications to real data demonstrate significant

reductions in the amount of time needed to perform such analyses, or reveal new biological insights, we can conclude that a method is useful in the context of systems biology. The value comes from the application of these methodologies and tools to real data.

Considering these methodological 'Big Data' issues in systems genetics, the following sub-questions were formulated to guide the research:

1. Whether and how can QTL be used in high-throughput trait data to computationally generate high (or higher) numbers of genetic markers to more accurately fingerprint samples?

2. Whether and how can software for QTL analysis, in particular MQM for complex traits, be scaled up to cope with dense fingerprint/marker information and high-throughput trait data?

3. Can we use and/or redesign existing software infrastructure for storage of and computation on high-throughput fingerprint/marker and trait data?

4. Which benefits of the first three research questions can be demonstrated on real high-throughput data, and which (systems) biological insights does this reveal?

1.3 The first genetic theories (1865-1930)

As this research project approaches 'old' genetic theories with new bioinformatics tools, it is relevant to place them in their historic context. The research field of genetics is relatively old, as it is founded when Gregor Mendel published his work on the inheritance of phenotypes in pea plants. His work, described in 'Versuche über Pflanzen-Hybriden' in 1865-1866 [17], gives us Mendel's Laws of Inheritance. He observed that crossing two plants with different flower colours leads to an offspring population with predictable colour ratios. He postulated the idea of heredity units and stated that each individual carries two of these, which results in a single phenotype. Each heritable unit is received from one of the parents, but which one is passed on to the offspring is random. Mendel formulated this principle as First Law on Segregation: if two units are not equal, a dominant unit (colour 1) overrules another unit (colour 2 - recessive).

Additional to studying colour, Mendel also studied other phenotypes in the pea plants. Now we know that Mendel was fortunate to select phenotypes which are caused by a single gene and are inherited independently. Mendel also observed this phenomenon of

independent inheritance when comparing the inheritance of multiple phenotypes. He concluded that a unit of inheritance is passed from parent to offspring in an independent fashion, that is, independent from other phenotypes that are inherited. This principle is known as Mendel's Second Law on Independent Assortment.

The relevance of Mendel's inferences was not fully recognised until 30 years later, when his work was rediscovered by Hugo de Vries, Carl Correns, and Erich von Tschermak. They (re)defined the rules for Mendelian Genetics [18]. In doing so, they provided biologists for the first time with a mathematical framework to study heritability of phenotypes caused by a single genetic unit.

In 1913, twenty years after the rediscovery of Mendel's Laws, careful observations of Thomas Hunt Morgan introduced an addition to Mendel's theoretical work. He observed that some phenotypes in his *Drosophila melanogaster* mutant flies are inherited together. This is in conflict with Mendel's Laws, stating independent inheritance of genetic units. Morgan called this phenomenon genetic linkage. A specific situation arises when phenotypes are linked to the sex of the animal. This is called sex based inheritance [19, 20].

Mendel worked on plants, while Morgan worked on fruit flies. Although the genetics are very similar, sex linked inheritance is not observed in most plants as they are usually male and female at the same time. Currently four plant families (asparagus, hops, papaya and silene) have been discovered that developed sex chromosomes acting similar to the sex chromosomes in animals [21]. Interestingly, these phenotypes do not follow Mendel's Second Law, as the traits are not inherited independently, but are linked to the sex phenotype.

The concept of genetic linkage is commonly explained as beads on a string, in which multiple genetic units are linked together. Two beads close together are tightly linked and almost always inherited together. In contrast, two beads that are far apart are less linked and can be separated by meiosis. Different strands are inherited independently, and thus allow Mendelian inheritance by two beads on two different strings.

Sturtevant, one of Morgan's students, used his mentor's linkage theory between phenotypes as a measure of distance between units of inheritance. He created the first genetic map of *Drosophila melanogaster* in 1916, 40 years before the discovery of the DNA molecule [19, 20], and at a time when molecular mechanisms were still unknown. Sturtevant's genetic map, based on phenotypes, was used to study genetics based on careful observation of the segregation of phenotypes in large populations of flies, and then determining the distance between the units of inheritance relative to other

phenotypes. Observations showed that sometimes the link between two phenotypes is broken. The amount of times this is observed in a population, is taken as a measurement of distance between the two phenotypes. For example, when we have measured 200 phenotypes in crosses, we can estimate which belong together on the same chromosome, as these are the ones which are found together more often than randomly expected. Further exploration of the data can indicate the order of the phenotypes on the map.

Intermezzo - The same method to measure distances is nowadays used to construct genetic maps on phenotypes in Pheno2Geno. It differs from methods using genotypes, i.e. SNP chips, as they require DNA sequencing data. Such methods are more expensive because first the phenotypes are measured, followed by the genotypes. Pheno2Geno applies data of measurements of distance between phenotypes to subsequently define the genotypes, without a second measure.

In 1918 Ronald A. Fisher published a paper entitled 'The Correlation Between Relatives on the Supposition of Mendelian Inheritance'. The paper demonstrates that a large number of small effects from discrete genetic units eventually add up to the continuous phenotypes observed [22]. Fisher proposed to unify all phenotypes, discrete and continuous, in one model. Mendel's theories could only explain dichotomous phenotypes, e.g. yellow versus green, but the phenotypes that were most interesting to researchers, are continuous. As Mendel's theory did not apply to these, his work was ignored in favour of other theories. Fisher developed mathematical approaches to explain continuous phenotypes, such as human stature, and showed that they follow Mendel's Laws. He proposed that geneticists had to search for a number of small effect loci. Summation of these small dichotomous effects from many genetic units on a population scale would add up to the observed normal distribution. (This theory was only proven after the discovery of DNA [24, 25]).

With this observation we move to the next section, in which a book by Fischer again plays a major role in the development of population genetics.

1.4 DNA and QTLs (1930-1990)

In 1930, Fisher published the book 'The Genetical Theory of Natural Selection' [23]. Fisher describes how Mendelian genetics is consistent with the idea of evolution by natural selection. His reasoning turned out to remove one of the last barriers for large scale adoption of Mendel's theories in biology. The subsequent discovery of DNA as the carrier of heritability by Avery, MacLeod and McCarty in 1944 [24], and the discovery of its structure by Watson and Crick in 1953 [25], revealed the long sought-after molecular basis for genetics.

After the discovery of restriction enzymes by Luria and Human [26] it was possible to cut and paste pieces of DNA. This allowed to experimentally work with DNA and investigate the different (lengths of) fragments produced, create new combinations of known DNA, and even novel DNA sequences to study the heritability of phenotypes. In 1970 the first techniques were developed to sequence DNA. The first approach used two-dimensional chromatography, and was followed by fluorescence based sequencing methods [27].

By the end of the 1980s everything was there for the next big step in genetics. Three papers were published which detailed the use of Restriction Fragment Length Polymorphism (RFLP) linkage maps to localise genes responsible for variation in quantitative phenotypes. These three papers by David Botstein and Eric Lander [1, 28, 29] form the basis for modern day linkage analysis and genome wide association studies. This allowed, for the first time in history, the direct association of sequences of DNA with its effect on a phenotype.

These tools are used by geneticists and bioinformaticians to identify the region of DNA underlying phenotypes or diseases (if any). These regions are by themselves already interesting because they provide targets to optimise the phenotype (e.g. when improving plant yield), or to develop therapies/drugs for human diseases. Two methods of QTL analysis are applied in searching DNA regions: linkage analysis and association analysis.

Linkage analysis is the DNA technology to trace parents and offspring/children. It is used to build so called genetic family trees or pedigrees (Fig. 1.1) in which segregation of a phenotype (for example a disease) is displayed. When we find a genetic region that cosegregates with the disease, we can assume the DNA and the disease are linked, allowing assessment of the strength of the observed linkage [30]. When experimental crosses derived from inbred parents are used, the need to build up complex pedigrees is avoided, while retaining the advantage of being able to trace back the DNA to the founder strains. Inbred populations also provide high statistical power at the cost of reduced resolution for QTL mapping [31]. This reduced resolution is caused by the lack of recombination in small populations (20 to 1,000 individuals).

An example of a basic effect scan using linkage analysis is shown in Figure 1.2. At each marker the mean of the phenotype for AA and BB is calculated, and the difference is plotted per marker. This gives an initial indication which markers may be relevant biomarkers for the phenotype of interest.

Association analysis can be applied in outbred or natural populations, when tracking the origin of the DNA is difficult or impossible. Molecular markers still allow to find information about the underlying DNA. We can use e.g. single nucleotide polymorphisms

(SNPs) to associate a phenotype with the genetic marker [32]. The advantage of this approach is that the resolution is very high (around 200 - 300 kb in humans [33]) because outbred populations have a high number of recombinations compared to experimental populations such as inbred lines. The downside is that a large number of markers reduces the statistical power to detect effects, because corrections for multiple testing have to be applied. Heritable phenotypes can be mapped to genomic locations by using a combination of DNA restriction enzymes, Mendel's inheritance laws, and Morgan's linkage theory. Together these methodologies and theories provide the experimental and statistical background to analyse heritability in any population.

Linkage analysis and association analysis, both developed in the 1980s, are still the foundation of population genetics research. More sophisticated tools and algorithms have been developed and implemented, but the basic theoretical concepts are the same. Some of these more advanced methods are discussed in the next section to sketch a background for the work presented in this thesis.

1.5 From phenotypes to genetical omics and GWAS (1990-2010)

The basis of an observed phenotype expression is often more complex than a single causative gene [34, 35]. To model these more complex situations where multiple regulators are influencing gene expression levels, extension of the basic model for QTL mapping is required. Extending the model is done by incorporating sources of variation as cofactors. This allows us to associate/partition the observed variance to either environmental factors (E) or genetic loci (G). The basic principle of Multiple QTL Mapping (MQM) is to include additional genetic components into the model, and then scan for QTLs conditional on these other genetic effects. MQM belongs to a family of QTL mapping methods, that includes Haley-Knott regression [36] and composite interval mapping (CIM) [37]. MQM combines the strengths of generalised linear model regression with those of interval mapping [38, 39].

During the first years of QTL mapping, the genetic maps were of poor quality, with a lot of missing marker data and large gaps between markers. As a solution to this problem interval mapping was developed. Interval mapping uses prior knowledge about linkage in inbred populations to map QTLs between two genetic markers. Hidden Markov Models (HMMs) are used to estimate the unobserved genotypes in the population based on the surrounding genotype observations, thereby improving the QTL mapping resolution [37, 38].

Genetical genomics is the concept that deals with endophenotypes (RNA, protein and metabolite abundance) as if they were common phenotypes. They are mapped in bulk to the genome, in a way similar to classical phenotypes [31]. Natural occurring variation and new omics tools allow us to track variation from the genotype all the way up to the classical phenotypes (see Table 1.1 for a short overview of different methodologies). Tracing the effects of variation from genome to phenotype allows genetics to go from individual QTLs to a system-wide approach of analysing QTLs at all known biomolecular levels. This approach is called systems genetics [40, 41].

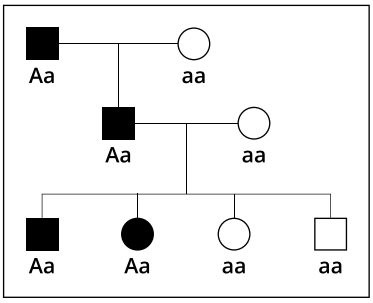


Figure 1.1 *– Example of a hypothetical pedigree on which linkage analysis can be performed. Here we show a hypothetical pedigree of individuals within a family (3 generations). Squares represent male individuals while circles are females. The phenotype is shown encoded by the fill color of the shape (black = affected, white = not affected). In this example pedigree, the "A" allele segregates with the disease. It is shared identical-by-descent in all affected individuals.*

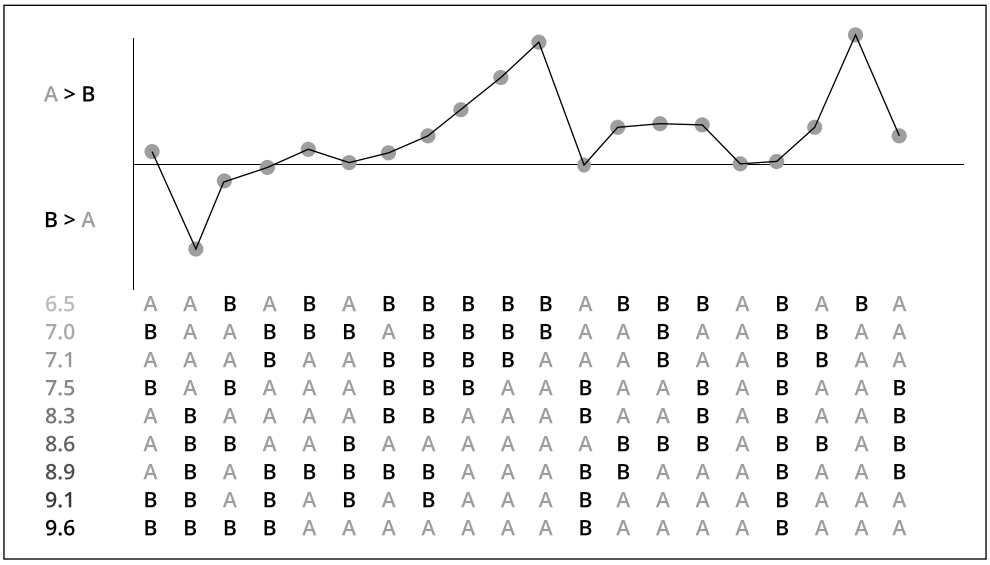


Figure 1.2 *– An example of how effect scans are performed using a single phenotype measured on 9 individuals (numbers) and genotypes represented by rows in the matrix (A or B). 20 markers (columns) were assessed: A and B are genetic variants obtained from either parental or maternal line.*

Studies in model organisms have shown high heritabilities for endophenotypes, such as gene expression, making these phenotypes ideal targets for QTL mapping [42, 43, 44]. In 2003 R/qtl was developed to provide reference implementations for QTL mapping. R/qtl is an extensible, interactive R package to map QTL in experimental crosses. It is implemented as an add-on package for the freely available and widely used statistical language/software R [45]. The main focus of R/qtl is to provide the mouse community with different QTL mapping methodologies, and allow them to deal with the aberrant segregation of the X chromosome. Furthermore it supports different types of inbred populations such as backcross (BC), F2, Recombinant Inbred Lines (RILs) and 4-way RILs [15].

When mapping gene expression or protein abundance, current knowledge of protein and DNA sequence allows us to locate their template on the genome. If QTL mapping resolution is high enough, we can even distinguish between traits mapping in the proximity of their respective gene (*cis*-eQTL) or to other regions in the genome (*trans*-eQTL). This information can be summarised into so called *cis-trans* plots, where the x-axis is the location of the eQTL and the y-axis the genetic location of the trait. Often so called *trans*-bands are observed; hotspots of many *trans*-eQTL mapping to a common region in the genome [46]. These are used to infer biological meaning and reconstruct co-expression and/or co-regulatory networks.

In the last decade Genome Wide Association Studies (GWAS) have identified thousands of genetic variants that are associated with human disease [47]. For reliable results GWAS needs a large cohort of genotyped and phenotyped individuals. Large consortia are working together to gather large amounts of human expression data from many different tissues. These data are then used in meta-analysis leading to eQTL GWA studies with even larger sizes (5,000+ individuals) [48], leading to more reliable results and enabling discovery of new modifiers of human gene expression. It is now recognised that many factors, such as effects on intermediate molecular phenotypes, influence the relationship between genotype and the eventual development of disease. It has also been observed that many of the disease predisposing variants are non-coding, which suggests that these variants have a regulatory function. Furthermore it has been shown that many disease predisposing variants (e.g. single nucleotide polymorphisms (SNPs)) affect the expression of nearby genes (i.e. *cis*-eQTLs) [48, 49, 50].

Currently data are being produced on all these biomolecular levels on an unprecedented scale. Advances in sequencing technology, transcriptomics, proteomics, and metabolomics allow data collection on a scale much larger than before [51, 52]. With this increasing scale of experimental data being produced in the lab, it will not be sufficient to have analysis software as simple downloads, because the researcher will also need

sizable compute and storage power [53]. How this increasing demand for storage and computational power can be satisfied in the future remains an ongoing question.

Federate computing providers may help researchers in their need for big computing solutions in the near future. However, this is not a universal solution, as the speed of data acquisition is currently higher than the speed of computing improvements [51, 52, 54]. Other solutions are sought by corporations and universities, who have started to combine their efforts and set up shared infrastructure. They attempt to deal with the necessary increase in compute power and to save overhead costs such as maintenance and electricity. The next section takes a closer look at the two main infrastructural developments for large scale computation: cluster and cloud computing.

1.6 Cluster and cloud computing (since 1980)

A cluster is a collection of computers dedicated to solve a computational task by divide and conquer [55, 56]. It basically means that a large supply of relatively homogeneous software is available on demand, including suitable compute and storage hardware. The hardware resources are divided by an internal scheduling system to facilitate efficient conduct of different tasks from various users. These users generally have little control over the computational environment, as the cluster administration and the scheduling system constrain its use. Computing clusters are fashioned to serve various types of hardware, such as:

- Ad-hoc networks are composed of many heterogeneous and relatively cheap systems such as: FPGAs, CPUs and/or ARM cores.

- Dedicated computer clusters, such as TARGET or national GRIDs, are usually a homogenous system of Linux machines used for computational tasks.

- Video card clusters for linear algebra, utilizing the power of many simple GPU cores for dedicated tasks.

Compute clusters do not have to be homogeneous in nature, but in many cases homogeneity is required to provide users with a stable computational environment to perform their tasks. Additionally the hardware (such as GPUs or dedicated ARM cores) can limit the usage of a cluster to a certain range of computational tasks.

The term ‘cloud’ is used in many different ways, and there is little consensus on what a cloud exactly is [57]. Essentially it means a user has software available on demand,

on suitable compute and storage hardware, and is charged for the time that it is being used. The leading cloud provider is currently Amazon, who has introduced the concept of a virtual Linux machine. This is a virtual Linux pre-initialised compute server, that is hosted within some large compute infrastructure, providing the user customised freedom for a reasonable price [58].

Many commercial, national and local compute centres are now also developing cloud compute server capacity. These virtual machines are therefore easy accessible provisions to distribute software and computation without the need for all participating computational nodes to install software. Thus, the infrastructure is specified jointly [57] but for the actual computation every partner can be private with their own data. This approach grants enormous computational power to everyone with minimal preparation - once a shared image is finalised [59].

The setup of such a compute cloud is not trivial and several initiatives are underway to ease this process. An example is the Debian Med initiative that organised workshops where bioinformatics tested a packaging system as a method to create a cloud. The initiative is considered a seed for an image which can be publicly shared.

Using this method the cloud infrastructure can be transferred to local computer clusters when necessary. Every participant has access to the server and can grant access to collaborators without having to pay the hosting fees. When complete, the server image can be ported to cloud providers, such as Amazon or Rackspace, to be reused by other researchers.

2 High-throughput generation of genetic markers

Using prior knowledge about phenotypes and how they are transferred during reproduction allows us to create phenotype based genetic markers. We can use these newly derived markers to saturate existing genetic maps or create them de novo when enough data is available. Pheno2Geno was developed to use a sensitive pre-selection and mixture modeling approach to improve performance. The package is designed to be fast enough to make use of data from diverse sources such as: microarrays, tiling arrays and/or RNA-Seq experiments. Using data from tiling arrays we were able to improve the genetic map resolution of an Arabidopsis thaliana Recombinant Inbred Line (RIL) derived from a Bayreuth (Bay-0) × Shahdara (Sha) cross.

Originally published as:

Konrad Zych, K. Joeri van der Velde, Ronny V. L. Joosen, Wilco Ligterink, Ritsert C. Jansen and **Danny Arends**
Pheno2Geno - High-throughput generation of genetic markers and maps from molecular phenotypes
Submitted

2.1 Pheno2Geno

Genetic markers and maps are instrumental for Quantitative Trait Locus (QTL) mapping in segregating populations. The resolution of QTL localization depends on the number of informative recombinations in the population and how well these recombinations are tagged by markers. Thus larger populations and denser marker maps perform better at detecting and locating QTLs. In practice, marker maps are often still too sparse. However, maps can be saturated or even be derived *de novo* from high-throughput omics data, such as gene expression, protein or metabolite abundance data. This is because molecular phenotypes are influenced by genetic variation and they will show a clear multimodal distribution due to major QTL effects, such information can therefore be converted into useful genetic markers.

The Pheno2Geno R package is developed for high-throughput generation of genetic markers and maps from molecular phenotypes. Pheno2Geno selects suitable phenotypes that show clear differential expression in the founders. Pheno2Geno uses mixture modeling to select phenotypes showing segregation ratios close to the expected Mendelian segregation ratios and transforms them into genetic markers suitable for map construction and/or saturation. Pheno2Geno analyzes the candidate genetic markers and excludes those showing multiple QTLs, epistatically interacting QTLs, and QTL by environment interactions to provide a set of robust markers for QTL mapping, protecting against genetic markers from a non-genetic origin.

We demonstrate our tool using gene expression data of 370,000 transcripts in 164 *Arabidopsis thaliana* Recombinant Inbred Lines (RILs). Pheno2Geno is able to saturate the existing genetic map decreasing the average distance between markers from 7.1 cM to 0.89 cM, close to the theoretical limit of 0.6 cM, pinpointing almost all of the informative recombinations in the population. Pheno2Geno is also able to create a *de novo* map from the gene expression data that is twice as dense as the original genetic map.

The Pheno2Geno package offers high-throughput *de novo* map construction and saturation of existing genetic maps. Processing of the showcase data set takes less than 30 minutes on an average desktop PC. Pheno2Geno improves QTL mapping results at no additional laboratory cost and with minimum computational effort. Pheno2Geno results are formatted for direct use in R/qtl, the leading R package for QTL studies. Pheno2Geno is freely available on CRAN under the GNU GPL version 3 licence.

2.2 Background

QTL mapping [29] is a powerful approach used in population analysis to link genetic variation to phenotype variation. It requires polymorphic genetic markers positioned on a genetic map. Phenotypes showing a dichotome 0/1 distribution with approximate equal proportions in, say, a RIL population can be used as genetic markers: genotypes can be called by connecting the 0/1 to the parental strains A/B. Such markers can then be used for *de novo* construction of the genetic map or for saturation of a known genetic map [60, 61].

Continuous (non-dichotome) phenotypes can also be used as markers if they show a major QTL: a major QTL will cause the phenotype to show a clear multimodal distribution to which a mixture model can be fitted [31, 38]. Posterior probabilities derived from mixture modeling are used for genotype calling. Such approaches have been used for up to 1,200 molecular phenotypes [62].

Here, we scale up the mixture model approach for non dichotome phenotypes in order to make analysis of hundreds of thousands of molecular phenotypes feasible, such as gene expression data.

Genetic maps created by Pheno2Geno can easily be used for QTL mapping: the package provides output structures compatible with R/qtl, the leading R package for QTL analysis in experimental crosses [15, 16]. Pheno2Geno allows users to explore and compare resulting maps with their favorite genome browser. Maps can be saved as a GFF (General Feature Format) file that is supported by most genome browsers.

2.3 Features

Pheno2Geno provides the following functionality to saturate and create genetic maps:

1. **Data preprocessing:** Pheno2Geno offers a selection of data transformation functions (including: *log*, *sqrt*, *reciprocal*, *probit* and *logit*). Gene expression data measured using microarrays are for example generally *log* [63] or *square root* [31, 62] transformed before further analysis.

2. **Analysis of parents of the segregating population:** When parental data are available Pheno2Geno uses a *t*-test to select molecular phenotypes showing significant differences between parental strains of the segregating population. This reduces the computational load of the analysis.

3. **Analysis of the segregating population:** Phenotypes with a major QTL will show clear multimodal distributions in a segregating population. Pheno2Geno fits a mixture model to the phenotype distribution [31, 38, 64]. Phenotypes are selected as candidate markers when significant multimodality is observed together with mixing proportions close to the expected segregation frequency, e.g. 1:1 for a bimodal distribution of two homozygous classes in a RIL, and 1:2:1 for a trimodal distribution of two homozygous and one heterozygous class in an F2 cross.

4. **Assigning genotypes:** The posterior probabilities of belonging to an underlying component distribution in the mixture is calculated for each component [64, 65]. Using these posterior probabilities, the continuous phenotype values are converted into discrete data (e.g. 0 or 1 for RILs; 0, 1 or 2 for F2). If the posterior probability for a specific marker-individual combination is lower than a user-specified threshold, a missing value (*) or partly informative value (e.g. not 0, but homozygous 2 or heterozygous 1) is assigned to avoid introducing genotyping errors. If parental data are available these can be given a parental origin label (A or B for RILs, A, H or B for F2). When parental data are not available, mixture-model based scores cannot be converted into parental origin labels. Pheno2Geno is able to solve that in case of RILs by forming twice as many linkage groups compared to the expected number of chromosomes and then merge anticorrelated pairs of linkage groups into a single chromosome.

5. ***De novo* construction of genetic maps:** When no initial map is available, Pheno2Geno can be used to create an initial 'skeleton' map. This skeleton map is produced using very strict settings in the mixture model analysis to obtain a limited number of highly trustworthy markers. These candidate markers are assigned to linkage groups using the R/qtl function *formLinkageGroups*. Additional information provided by the user is used in this step, e.g. known physical and genetic positions will be used by Pheno2Geno to assign physical chromosome IDs to linkage groups and to determine the correct orientation of chromosomes. The package then orders all the markers inside a linkage group by using the R/qtl *orderMarkers* function. Finally the skeleton map is saturated to improve resolution as described in the next section.

6. **Environment and epistasis:** West *et al.* [35] emphasized that creation of genetic markers from gene expression data is seriously hampered by the presence of environmental variation and multiple possibly interacting QTLs (epistasis) using R/qtl. Pheno2Geno tests if candidate markers are affected by multiple QTLs or pairwise interactions. When data are collected in multiple environments, potential environmental interactions are tested. The user decides whether affected candidate markers are flagged or removed from further analysis.

7. **Saturation of a known map:** Pheno2Geno performs interval mapping (using the R/qtl *scanone* function) of candidate markers on the original map. The candidate markers are placed on the position of the QTL peak. The map is re-estimated (using the R/qtl *est.map* function). Followed by removal of duplicate candidate markers and markers located at the position of a known marker.

8. **Detection of errors:** After saturation or *de novo* construction of a genetic map, Pheno2Geno can detect and correct genotyping errors (e.g. double recombinations, missing data, semi-informative markers) using the R/qtl function *fill.geno*. Furthermore, when saturating a known map with available genotype data, Pheno2Geno can detect sample mix-ups in the original data using R/lineup (which is part of the R/qtl toolset). Users can also use external tools such as MixupMapper [66] beforehand to detect and correct the original genotype data.

2.4 Results

In order to test our package, we performed an analysis of a population with a sparse map. The original AFLP map was created using a population of 420 RILs derived from a cross between A. *thaliana* Bayreuth (Bay-0) × Shahdara (Sha) [67]. Our data set consists of 164 RILs from the core population, which were assigned to 4 conditions using the designGG package [68]. Parents were measured in duplicate per condition. Gene expression was measured on 370,000 transcripts. During Quality Control 16 arrays (all RILs) were discarded, leaving 148 RILs and 16 parental arrays for further analysis.

The original map contained 69 AFLP markers at an average map distance of 7.1 cM [67]. The resolution of a genetic map is limited by the size of the population from which the map is derived. A distance of 1 cM is equal to 1 recombination per 100 individuals. Our sample size of 148 individuals implies that we can obtain at best a resolution of 0.68 cM between markers.

10,801 phenotypes are detected as being differentially expressed between parents (`P < 0.01`). Mixture modeling identified 1,230 potential markers showing approximately 1:1 segregation ratio. Pheno2Geno removed 267 markers which show QTL by environment interaction (`LOD >= 7.5`), and 286 candidate markers which show none (`LOD < 15`) (279 markers) or multiple QTL (7 markers). Scanning for epistasis showed 77 candidate markers which appear to show pairwise epistatic interactions (`LOD >= 7.5`).

Using the remaining 600 candidate markers, the original map was saturated, and 103 co-localizing markers were removed. This resulted in 497 new gene expression based

markers (720% increase). Map distances were re-estimated (*est.map*) using the Kosambi map function. Map expansion was observed on chromosomes 4 and 5 increasing total map length from 480.7 to 501.5 cM (Fig. 2.1). Nonetheless, the saturation resulted in decreasing the average map distance from 7.1 cM to 0.89 cM. This is close to the theoretical resolution limit (0.68 cM) given the size of the population. Saturation of the A. *thaliana* Bay-0 × Sha map led to a more than sevenfold improvement in marker density at no additional lab costs.

A *de novo* reconstruction using gene expression data only (ignoring the original markers and map) would have led to a skeleton map containing 227 markers with average distance of 2.2 cM.

Additionally, we performed QTL mapping of our previously published classical phenotype data set [69] onto the saturated map. This showed an increase in QTL likelihood for 56% of previously detected QTLs. Additionally, 29 new QTLs were detected on the saturated map. These QTLs showed LOD scores close to the threshold when mapped onto the original map (`3.4 <= LOD <= 5`, Fig. 2.2).

Finally, QTL mapping of all the gene expression probes showing differential expression between parents (10,801 probes) was performed. 5,837 probes had a significant (`LOD > 5`) QTL on the original map. Out of these, 3,943 probes (66%) showed an increase in QTL likelihood on the saturated map (Fig. 2.3) and additional 210 new significant QTLs were detected.

2.5 Conclusions and discussion

Pheno2Geno is a generic software package for generating genetic markers and maps from high-throughput molecular phenotypes for any inbred diploid population (e.g: backcross, F2 intercross and recombinant inbred lines). Pheno2Geno has four important features which we will discuss one by one:

1. **'Big Data' computation.** Pheno2geno can process large volumes of different kinds of molecular phenotypes [58]. Memory requirements of the algorithm are decreased by reading in and processing files in chunks rather than at once. Complete analysis of the showcase data (370,000 transcripts) is performed in less than an hour on an average desktop PC (Intel Core i5, 4 GB of RAM). For even larger data sets, the Pheno2Geno package is embedded in xQTL workbench [70, 71] allowing for easy parallelization, use of cluster and cloud computing.

2. **Integration with R/qtl.** The package is employing well optimized methods and functions of R/qtl for all the mapping steps, filling and (re-)estimation of maps. Moreover, genetic maps created by Pheno2Geno can be directly used in R/qtl, providing a smooth transition from genetic map creation to QTL mapping.

3. **Strict selection of candidate markers.** The analysis of Pheno2Geno contains multiple selection steps filtering out candidate markers of low quality. E.g. candidate markers affected by multiple QTLs and/or environment are flagged and can easily be excluded from the analysis.

4. **Gene expression phenotypes.** We have illustrated Pheno2Geno on arraybased gene expression data. If a gene expression phenotype shows a significant QTL (eQTL) and if this eQTL co-localizes with the probe (a local eQTL), then the derived marker will be mapped at the location of the probe. The eQTL may actually be caused by polymorphisms in the region targeted by the probe [72, 73]. If the QTL does not lo-localize with the probe (a distant eQTL), the derived marker will not be located at the region targeted by the original probe but correctly at the position of the distant eQTL.

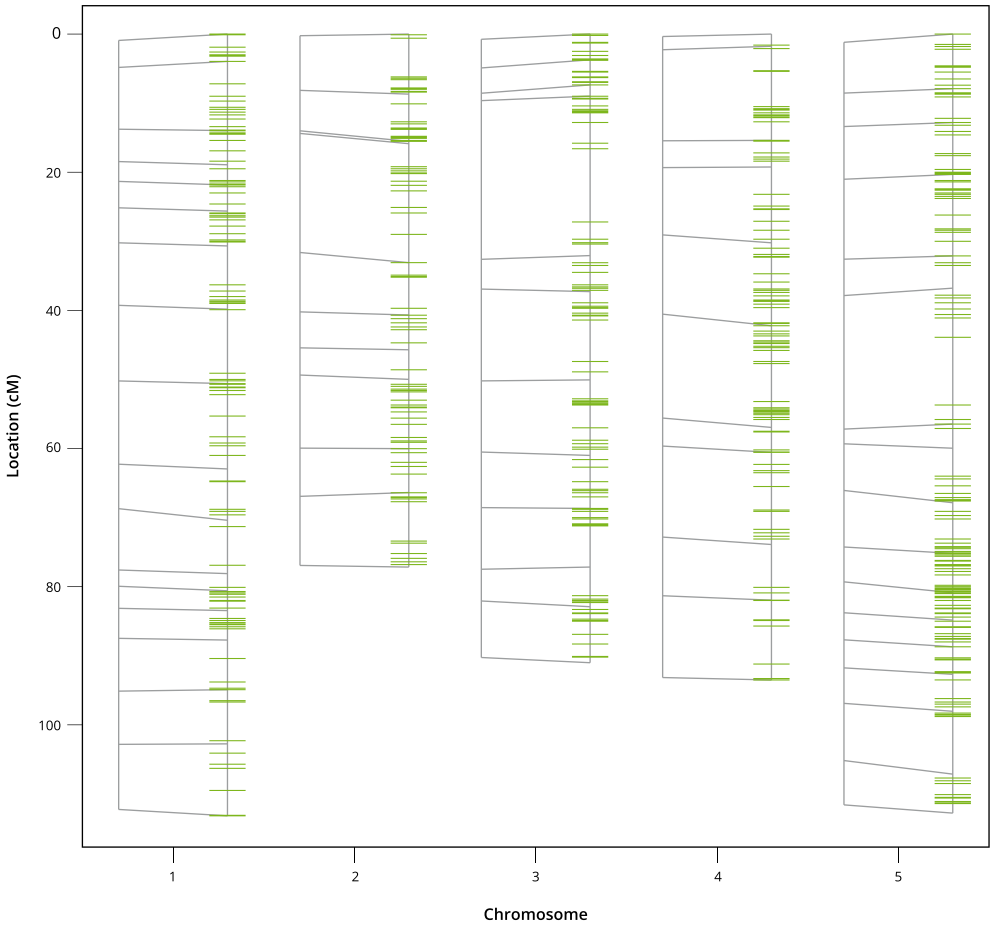


Figure 2.1 *– Map comparison plot drawn using R/qtl function plot.map [15, 16]. For each of the chromosomes the original map (gray lines) and the saturated map (green lines) are plotted. Lines are drawn to connect markers. Markers that exist in just one map and not the other are indicated by short line segments, on one side or the other, that are not connected across. Before plotting, both maps were re-estimated using the R/qtl function est.map. The original map consists of 5 chromosomes and 69 markers at an average distance of 7.1 cM. The saturated map consists of the original 69 markers plus 497 expression-based markers at an average marker distance of 0.89 cM.*

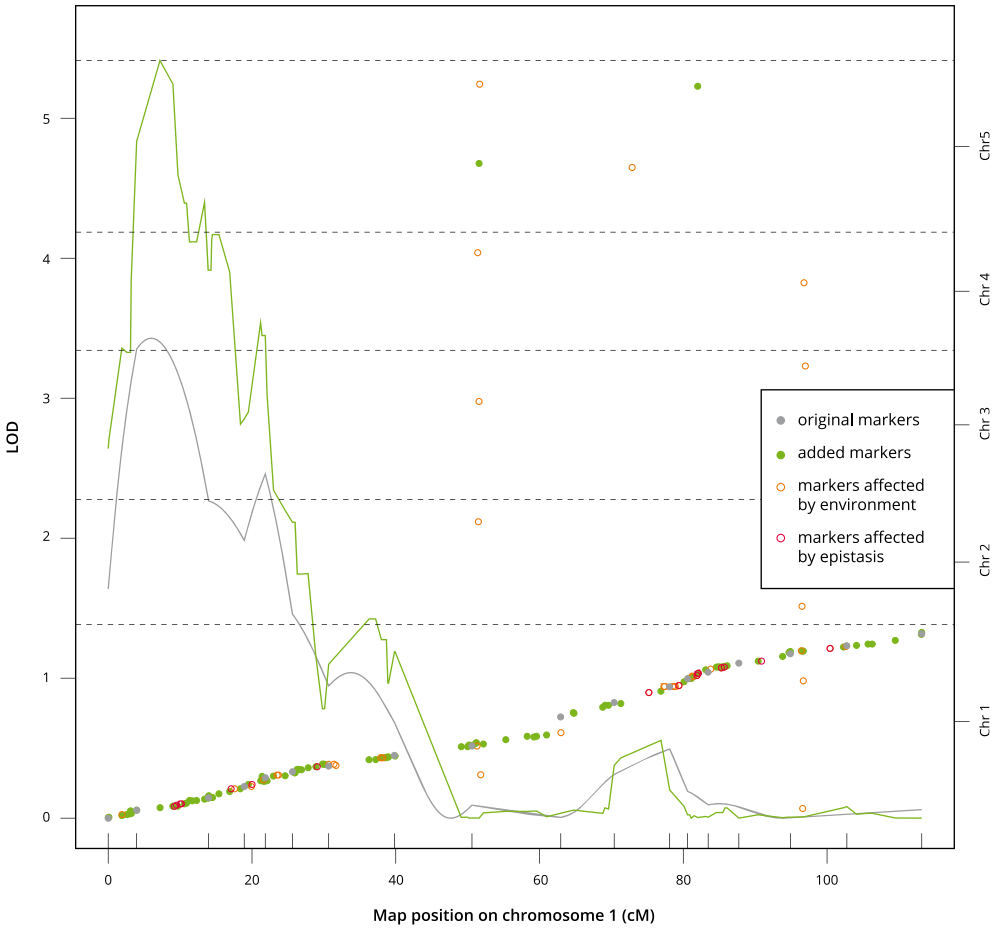


Figure 2.2 *– Results of single-marker QTL mapping of a classical phenotype on original (gray line) and saturated map (green line) (left axis). X-axis: only chromosome 1 is shown. Ticks on the x-axis: positions of the original markers on the genetic map. Gray dots: positions of the original markers on the physical map. Colored dots and circles: candidate markers detected by Pheno2Geno. Orange circles: candidate markers removed because they showed significant environmental influence. Red circles: candidate markers removed because they showed an epistatic interaction with other genetic markers. Green dots: markers used for saturation of the original map. The final saturated map consists of all the green and gray dots. Shown here are locations of the new markers on the old map. In this way maps align for better clarity.*

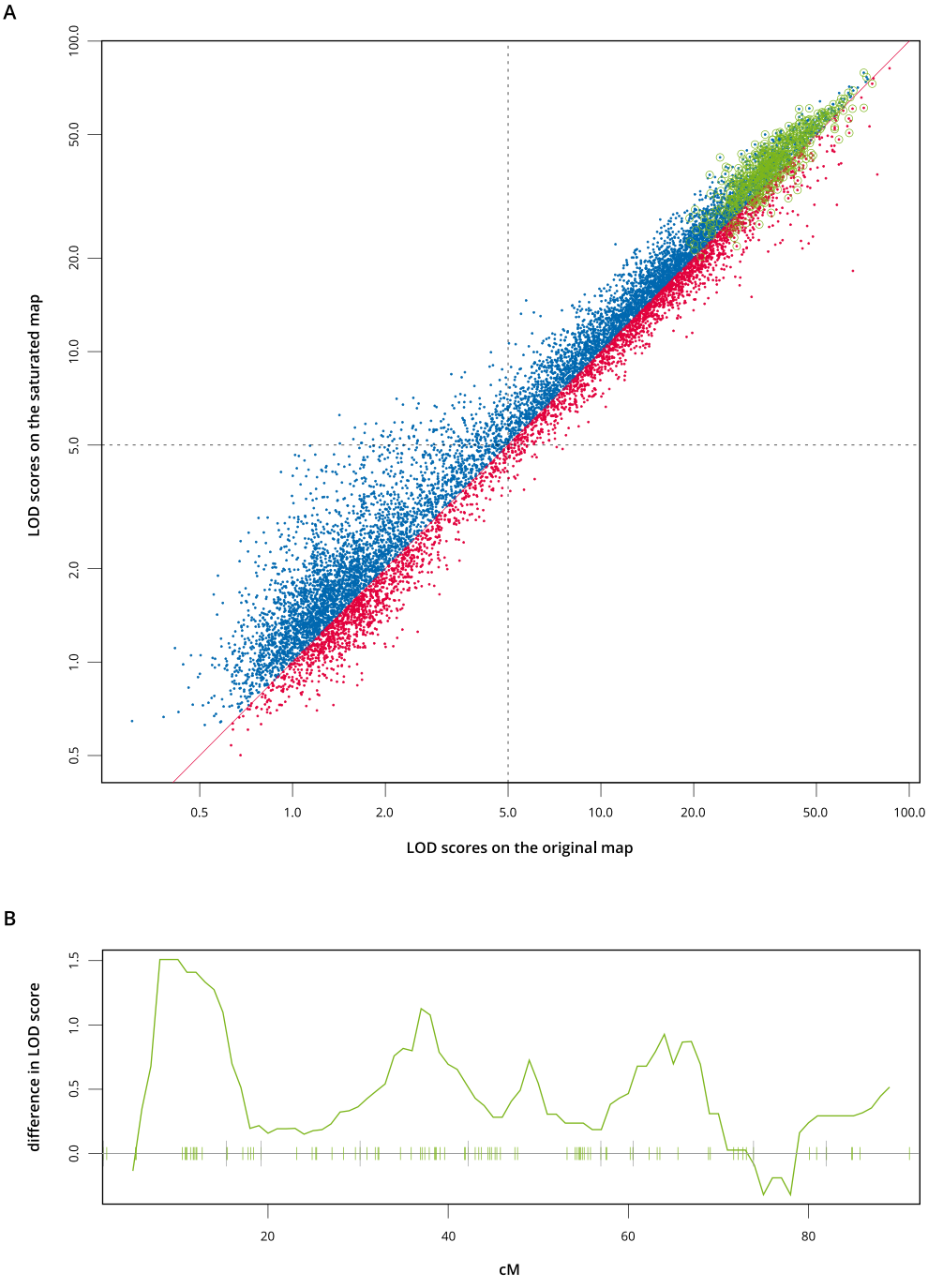

A
LOD scores on the saturated map
LOD scores on the original map
0.5
1.0
2.0
5.0
10.0
20.0
50.0
100.0
B
difference in LOD score
0.0
0.5
1.0
1.5
20
40
60
80
cM

Figure 2.3

A. *LOD scores on original and saturated map.* *QTL mapping was performed on all* 10,801 SNP *probes showing differential expression between parents (*`P < 0.01` *Student's t-test) using original and saturated map.* 5,837 *probes show a QTL with a* `LOD > 5` *on the original map. Blue dots:* 3,943 *probes* (66%) *that show an increased LOD score on the new saturated map. Moreover,* 210 *new QTLs were detected on the saturated map. Red dots: probes showing a decrease in LOD score on the saturated map. Green circles: probes used to saturate the map.*

B. *Changing LOD scores.* *For each of the phenotypes the top QTL peak was selected. If the peaks measured on original and saturated map shared locations, the differences between LOD scores was calculated. Solid green line: median of differences between peaks from chromosome 4, calculated inside a sliding window of* 10 cM, *moved across chromosome with a step of* 1 cM. *For each of the windows the value was plotted in the middle of the compartment (hence no value for the first and the last* 5 cM). *Ticks on the x-axis show the position of the markers: gray tall ticks: original markers; green short ticks: markers selected by Pheno2Geno. Only one region, where no new markers were added* (75–80 cM) *does not show increase in power.*

3 High-throughput (Multiple) QTL mapping

This chapter describes the implementation of Multiple QTL mapping (MQM) into R/qtl, adding a 'new' algorithm to the R/qtl toolset to provide a wider range of QTL mapping tools for inbred crosses. R/qtl allows researchers to analyse data from different sources or to quickly compare different approaches. This chapter showcases our contributions to the R/qtl package such as: MQM, visualizations, parallel computation of QTLs and an improved permutation scheme. We evaluated the performance of MQM on classical phenotypes and metabolite abundance in Arabidopsis thaliana, showing improvement in the number and significance of detected QTL, compared to mapping when using a single marker model.

Originally published as:

Danny Arends*, Pjotr Prins*, Ritsert C. Jansen and Karl W. Broman
R/qtl: high-throughput Multiple QTL mapping
Bioinformatics 26(23):2990-2 (2010)

Danny Arends*, Ronny V. L. Joosen*, Leo Willems, Wilco Ligterink, Henk Hilhorst, Ritsert C. Jansen
Visualizing the genetic landscape of Arabidopsis seed performance
Plant Physiology 158(2):570-89 (2011)

Danny Arends*, Ronny V. L. Joosen*, Yang Li*, Leo Willems, Joost J.B. Keurentjes, Wilco Ligterink, Ritsert C. Jansen, Henk Hilhorst
Identifying genotype-by-environment interactions in the metabolism of germinating Arabidopsis seeds using Generalized Genetical Genomics
Plant Physiology 162(2):553-66 (2013)

Here we show the contributions made to R/qtl and the application of our newly developed toolset to map genetic variation underlying classical phenotypes in *Arabidopsis thaliana* seed development. We show that by using a reduced number of individuals using the designGG strategy we can map main effects and interaction effects with more statistical power compared to other designs. After this we use a limited number of individuals and continue our analysis by zooming in on the metabolic level. We show that there are shared genetic loci between phenotypes at different biomolecular levels. Additionally, the use of the new optimized MQM routine allows us to detect more loci and/or give more confidence in the loci detected.

3.1 Single marker QTL mapping (R/qtl)

R/qtl is an extensible, interactive environment for the mapping of Quantitative Trait Loci (QTL) in experimental crosses. It is implemented as an add-on package for the freely available and widely used statistical language/software R [45]. Since its introduction, R/qtl [15] has become a reference implementation with an extensive guide on QTL mapping [74]. R/qtl development is continuous, with input from multiple collaborators and users. We have introduced a full testing environment with regression testing, updated the license to the GPL version 3, and hosted the source code repository on Github, which gives R/qtl software development high visibility and transparency.

The development of R/qtl reflects trends in quantitative genetics, in particular the use of larger data sets, larger calculations and requirements for controlling the false discovery rate. These developments are partly driven by high-throughput genetical genomics, the name coined for the study of gene expression QTL (eQTL) [31], metabolite QTL (mQTL), and protein QTL (pQTL).

3.2 Multiple QTL mapping

Multiple QTL Mapping (MQM) belongs to a family of QTL mapping methods that include Haley-Knott regression [36] and composite interval mapping (CIM) [37]. MQM combines the strengths of generalized linear model regression with those of interval mapping [38, 39].

Recent developments in QTL mapping include Bayesian modeling of multiple QTL e.g. R/qtlbim package [75, 76]. Bayesian modeling, however, is computationally expensive, and arguably has little additional power when applied to high density maps, and (nearly) complete genotype data [77]. Still, we intend to combine the strengths of the different methods in future versions of R/qtl.

These days, with most experimental setups and high density maps, improving precision may be achieved by increasing the population size first. For more information on QTL mapping and Bayesian analysis we refer to the 'Handbook of Statistical Genetics' [78]. MQM makes use of generalized linear models, thereby potentially providing unified analysis of non-normal traits.

MQM provides a practical, relevant and sensitive approach for mapping QTL in experimental populations. The theoretical framework of MQM was introduced and explored by Ritsert C. Jansen [79] and is explained in the 'Handbook of Statistical Genetics' chapter 21 [77]. MQM has one known commercial implementation [80], which has been used effectively in practical research, resulting in hundreds of papers, e.g., in mouse, plant, and fish, respectively [81, 82, 83]. Now, with MQM for R/qtl, we present the first free and open source implementation of MQM, that is multi-platform, scalable and suitable for automated procedures and large data sets.

3.2.1 Features

MQM for R/qtl is an automated three-stage procedure in which, in the first stage, missing genotype data is 'augmented'. In other words, rather than guessing one likely genotype, multiple genotypes are modelled with their estimated probabilities. In the second stage, important marker cofactors are selected by multiple regression and backward elimination. The third stage, a QTL is moved along the chromosomes using these pre-selected markers as cofactors (Fig. 3.1). QTL are interval-mapped using the most informative model selected by either maximum likelihood or restricted maximum likelihood. A refined and automated procedure for cases with large numbers of marker cofactors is included.

The method lets users test different QTL models by elimination of non-significant cofactors. MQM for R/qtl brings the following advantages to QTL mapping:

1. Higher power, as long as the QTL explain a reasonable amount of variation.

2. Protection against over-fitting, because MQM fixes the residual variance from the full model, which allows the use of more cofactors than may be used in, for example, composite interval mapping (CIM) [37].

3. Prevention of ghost QTL detection (between two QTL in coupling phase).

4. Detection of negating QTL (QTL in repulsion phase).

5. MQM gives (compared to CIM) a reduction in type I and type II error [77].

6. A pragmatic permutation strategy for controlling the false discovery rate (FDR) and prevention of locating false QTL hot spots, as discussed in Breitling *et al.* [46]. Marker data is permuted, while keeping the correlation structure in the trait data.

7. High-performance computing by scaling on multi-CPU computers, as well as clustered computers, by calculating phenotypes in parallel, through the Message Passing Interface (MPI) of the SNOW package for R [85].

8. Visualizations for exploring interactions in a genomic circle plot (Fig. 3.2) and *cis*- and *trans*-regulation (Fig. 3.3).

A 40-page tutorial for MQM explores both the automated procedure and the manual procedure of adding and removing cofactors in an A. *thaliana* recombinant inbred line

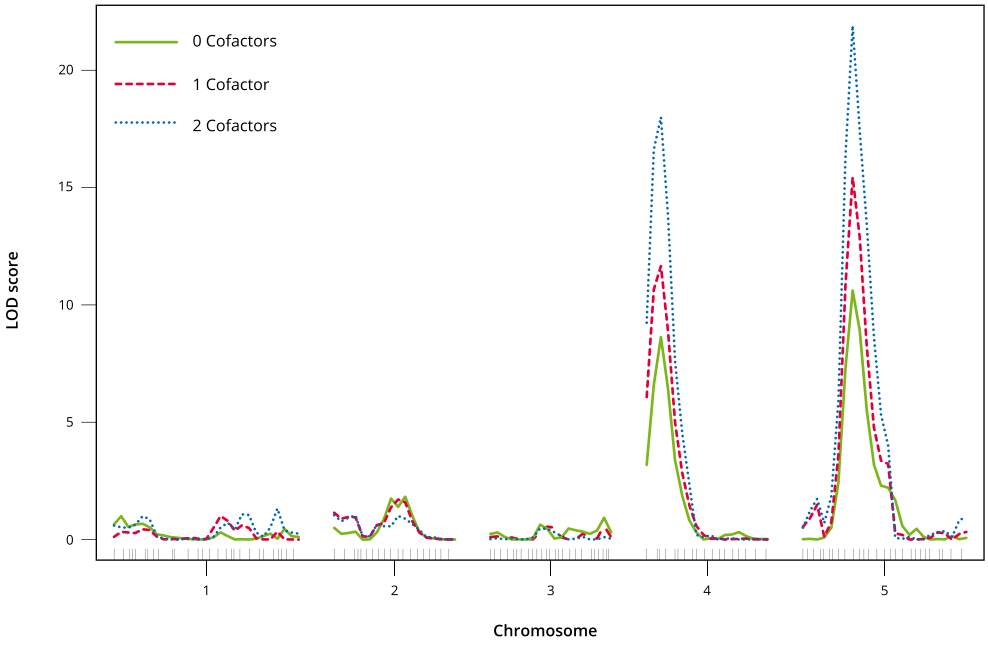


***Figure 3.1** – Three way comparison of MQM performance in A. thaliana [84]. LOD score increases when cofactors are added manually to the model. Here, adding more than two cofactors does not improve the model any further (as discussed in the online MQM tutorial).*

(RIL) metabolite (mQTL) data set with 24 metabolites as phenotypes [84]. In addition, the tutorial visually explains the effects of data augmentation, cofactor selection, model selection, and tweaking of input parameters, such as cofactor significance. Genetic interactions (epistasis) are explored through effect plots, and an example is given of parallel computation. The tutorial is part of the software distribution of R/qtl and is available online.

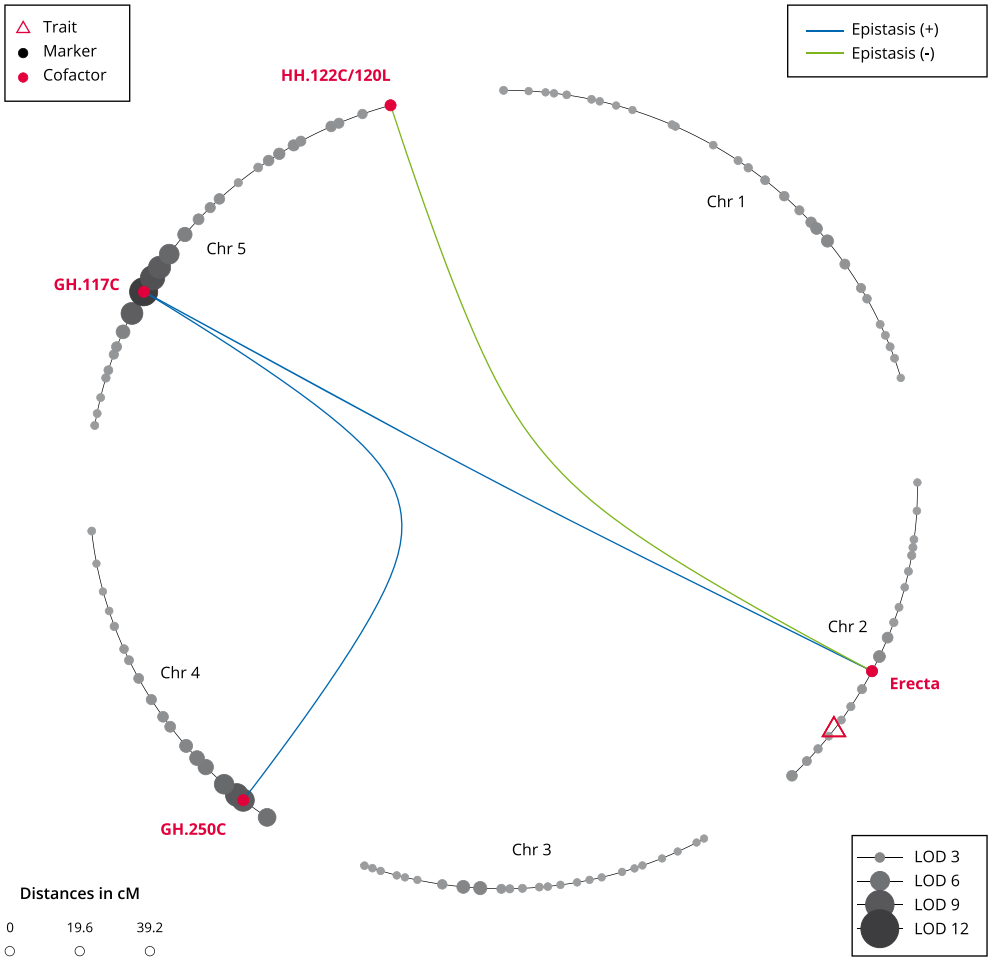


Figure 3.2 *– Circular genome interaction plot of the A. thaliana glucosinolate pathway [84]. LOD scores shown at marker positions are scaled (grey circles), with selected cofactors (red circles) and epistasis between multiple cofactors (green and blue splines).*

3.2.2 Conclusions and discussion

MQM for R/qtl is a significant addition to the QTL mapper's toolbox. R/qtl provides the user with the most frequently used statistical analysis methods: single-marker analysis, interval mapping, Haley-Knott regression [36], CIM [37] and MQM [79]. MQM has improved handling of missing data and allows more powerful and precise detection of QTL, compared to many other methods. Not only is this new implementation of MQM available in the statistical R environment, which allows scripting for pipe-lined setups, it is also highly scalable through parallelisation and paves the way for high-throughput

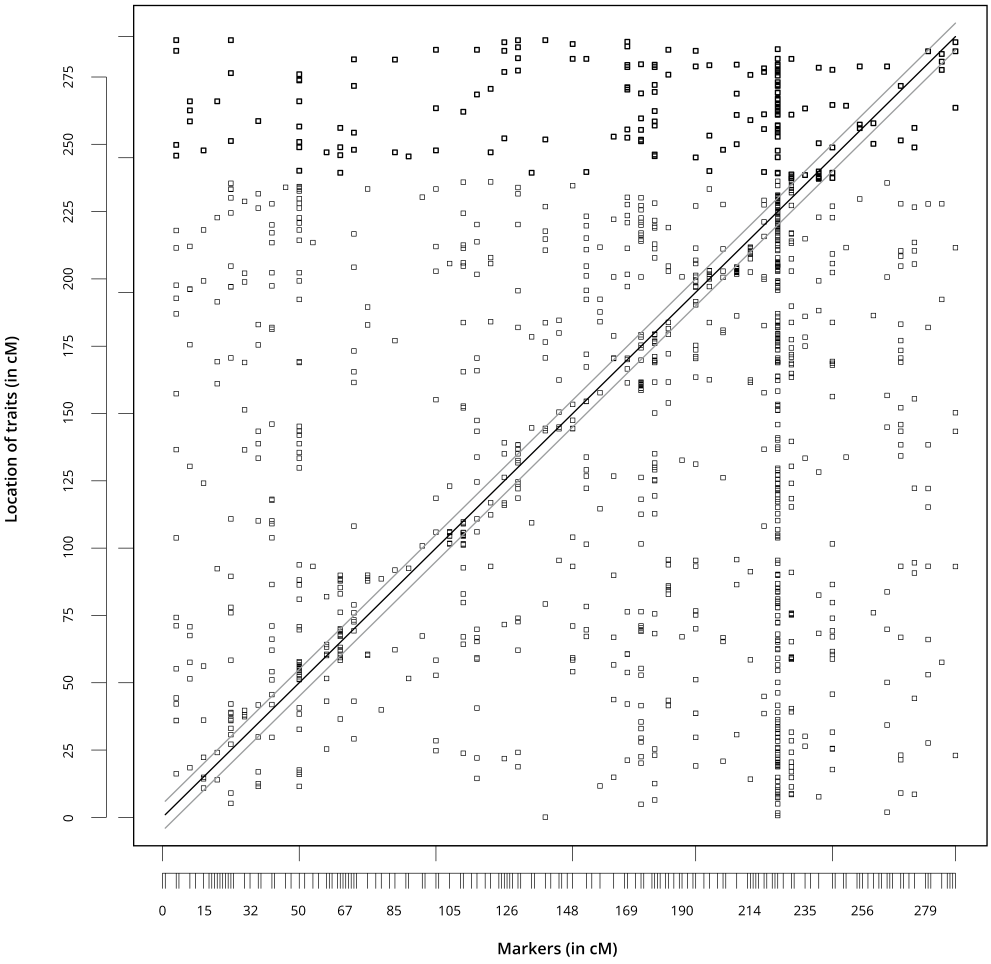


Figure 3.3 *– Cis-trans plot of significant QTL (squares) showing cis-acting QTL (diagonal) and a trans-band (vertical, chromosome 5) in Caenorhabditis elegans [86].*

QTL analysis. With MQM, R/qtl is a free and high-performance comprehensive QTL mapping toolbox for the analysis of experimental populations. R/qtl now includes permutation strategies for determining thresholds of significance relevant for QTL and QTL hotspots; the first step towards causal inference and network analysis.

3.3 Mapping classical phenotypes

Perfect timing of germination is required to encounter optimal conditions for plant survival and it is the result of a complex interaction between molecular processes, seed characteristics and environmental cues. To detangle these processes we made use of natural genetic variation present in an A. *thaliana* Bayreuth × Shahdara RIL population. For a detailed analysis of the germination response we characterized rate, uniformity and maximum germination and discussed the added value of such precise measurements. The effects of after-ripening, stratification and controlled deterioration as well as the effect of salt (NaCl), mannitol, heat, cold and ABA with and without cold stratification were analyzed for these germination characteristics. Seed morphology (size, length) of both dry and imbibed seeds were quantified by using image analysis. For the overwhelming amount of data produced in this study we developed new approaches to perform and visualize high-throughput QTL analysis. We show correlation of trait data, (shared) QTL positions and epistatic interactions. The detection of similar loci for different stresses indicate that often the molecular processes regulating environmental responses converge into similar pathways. Seven major QTL hotspots were confirmed using a HIF approach. QTLs co-locating with previously reported QTLs and well characterized mutants are discussed. A new connection between dormancy, ABA and a cripple mucilage formation due to a natural occurring mutation in the MUM2 gene is proposed, which is an interesting lead for further research on the regulatory role of ABA in mucilage production and its multiple effects on germination parameters.

3.3.1 Background

Colonizing plants are subject to a wide variety of environmental conditions. For successful adaptation to new habitats the timing of developmental transitions is especially important. Seed germination is one of these important transitions as it determines the seasonal environment experienced in further plant life [87]. Natural populations that develop under distinct environmental conditions may reveal genetic adaptation, which can be used to disentangle the signaling routes that are involved. Seed germination is described by three phases of water uptake. In phase 1 the seed imbibes and reinitiates metabolic processes followed by a lag phase (phase 2). Further water uptake results in protrusion of the radicle through the testa and endosperm

(phase 3). The moment of radicle protrusion through the endosperm is considered to be the moment of germination sensu stricto [88]. To characterize the genetic variation of germination related traits we focused on the effect of the environment that a seed perceives during germination rather than the effect of the environment during maternal plant growth, which has been the subject of other studies [89, 90].

Seed content (e.g. oil) is often used as commodity and modifications to the content can therefore be regarded as seed quality parameters as well. To prevent confusion we will use the term seed performance to indicate that the focus of our study was restricted to seed germination characteristics.

The production of high quality crop seed not only entails knowledge about maternal plant growth, harvesting and storage of seeds, but also of germination conditions [91]. To obtain better germination and field performance, many seed companies rely on enhancement methods, such as seed priming and coating and/or pelleting, but these methods are reaching their limits. Dissecting the molecular mechanisms underlying seed germination and its tolerance to the environment may unlock the full genetic potential and enable targeted breeding for seed performance. In this study we used a recombinant inbred line (RIL) population derived from two A. *thaliana* ecotypes: Bayreuth (Bay-0) which originates from a fallow land habitat in Germany and Shahdara (Sha) which grows at high altitude in the Pamiro-Alay mountains in Tadjikistan [67]. The Bay-0 × Sha RIL population has been used in many previous studies to map QTL positions for root morphology [92, 93], anion content [94], nitrogen use efficiency [95], cell wall digestibility [96], carbohydrate content [97], sulfate content [98], leaf senescence [99], morning-specific growth [100] and cold-dark germination [101]. We have used the natural variation present in this RIL population to map the response of germination characteristics to environmental conditions to which a seed is exposed.

Freshly harvested viable *Arabidopsis* seeds often don't germinate even when placed under conditions favorable for germination. This event, called primary dormancy, is shown to be subject to natural variation [102]. In many *Arabidopsis* ecotypes, this primary dormancy is released after a period of dry storage at room temperature. Another dormancy breaking treatment is cold stratification were seeds are imbibed in water and stored at 4 °C in the dark for four days before putting them into optimal conditions for germination [88]. Unfavorable conditions during seed germination may result in a changed rate or even failure of germination. In *Arabidopsis*, it has been shown that the responsiveness to temperature is closely related to the level of after-ripening [103]. High salt concentrations induce osmotic stress and ion toxicity resulting in both a delay and reduction of maximum germination [104]. Often, these different environmental stresses are interconnected and will cause osmotic and associated oxidative stress [105, 106].

The plant hormone Abscisic Acid (ABA) plays a predominant role in plant responses to different environmental stresses and can activate various signal transduction pathways leading to a global change in transcription [107, 108]. Exogenous application of ABA during germination results in a distinction between testa and endosperm rupture. At certain concentrations the testa will rupture but germination sensu stricto (radicle protrusion through the endosperm) will be inhibited. This phenomenon, caused by reduced weakening of the endosperm cap, is the consequence of a complex interplay between ABA, GA and ethylene signals [109]. In this report, we determined germination sensu stricto for primary dormancy in freshly harvested seeds, germination of fully after-ripened seeds with and without a preceding cold stratification period (see material and methods for conditions), and germination under various stress conditions (low/high temperature, salt/osmotic stress and ABA) to assess natural variation in the Bay-0 × Sha RIL population. Additionally, seed morphology (size and length) and flowering time were phenotyped as they have been shown to be strong determinants of plant trait variation [90, 110, 111]. We correlated these traits to our germination related traits to evaluate

Trait	G_{max}	AUC	t_{50}	t_{10}	U_{8416}
AR.NS	0.82	0.97	0.86	0.79	0.82
AR.NS.Cold	0.51	0.77	0.73	0.66	0.48
AR.NS.Mannitol	0.61	0.79	0.70	0.55	0.62
AR.NS.NaCl	0.90	0.94	0.80	0.76	0.43
AR.WS	0.63	0.19	0.78	0.72	0.72
AR.WS.NaCl	0.91	0.93	0.86	0.78	0.70
Fresh.NS	0.92	0.94	0.81	0.70	0.76
Fresh.WS	0.40	0.81	0.87	0.84	0.76

Table 3.1 *– Overview of the broad sense heritability scores. Included are those traits for which different blocks were tested (trait code descriptions can be found in Table 3.2, for germination parameters see Figure 3.4). Broad-sense heritability were calculated with the QTL data analysis tools in Genstat 14, using the preliminary single environment analysis and adding the block as an additional fixed term.*

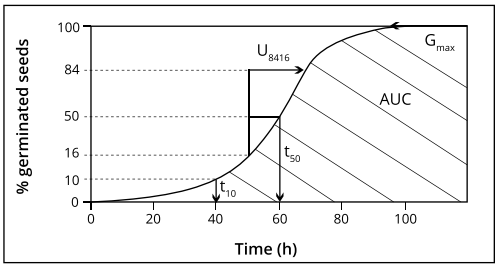


Figure 3.4 *– A hypothetical germination curve, t_{10}: time when 10% of seeds have germinated, t_{50}: time when 50% of seeds have germinated. U_{8416}: uniformity of germination, time difference between 16% germination and 84% germination. G_{max}: % of total germinated seeds, AUC: Area under the curve.*

possible causality. In total this analysis resulted in 327 trait scores over different harvests. Evaluation of these high numbers of phenotypes demanded methods of QTL analysis that extended beyond mapping of individual traits and that allowed comprehensive and comprehensible visualization.

Analysis of natural variation that is captured in well-defined recombinant inbred populations has shown to be a powerful tool to detect important loci that influence the traits under study [112]. To uncover the loci with genetic variation a statistical framework is needed. For this, any programming language can be used which supports statistics. In the life sciences the statistical language R is often the prime candidate. R is open source, contains the latest in statistical analysis methods and has a large community for help and support (www.r-project.org). Furthermore, it has the R/qtl package [15], which contains an array of different QTL mapping methods, including Single Marker Mapping, Composite Interval Mapping (CIM) and Multiple QTL Mapping (MQM) [16]. Although all possibilities to perform a detailed QTL analysis including data preprocessing and output formatting are present in R, it requires extensive knowledge of the R-syntax to combine all necessary steps in a single analysis protocol that can loop through hundreds or thousands of traits.

This type of automated analysis combined with efficient/automated data visualization (Fig. 3.5, Fig. 3.6 and Fig. 3.7) is a necessary step to keep up with the increasing rate of biological data production. For using single trait mapping the effect of a certain treatment, e.g. germination at high temperature, must be corrected by the germination characteristics under control conditions. Here, we subtracted the observed germination under stress conditions from values for germination under control conditions. This correction can lead to complicated interpretation, especially when the environment under study affects loci with already strong effects under control conditions. Further, it can reduce statistical power due to summation of the error components. Therefore we performed an additional analysis using a QTL by environment (QTL×E) approach [113, 114]. Instead of considering individual responses, one can then treat the stress conditions as a set of environmental perturbations and evaluate a single trait (such as germination percentage). Because several environments are taken into account simultaneously, the statistical power to detect loci that are affected by several environments increases, and interpretation becomes more intuitive as the need for correcting the stress response by the control response is eliminated [115].

The Bay-0 × Sha RIL population consists of 420 lines that were genotyped in the F6. This relatively low degree of inbreeding provoked residual heterozygosity present at almost all genome positions. This residual heterozygosity can be used to confirm QTL positions, as it provides a possibility to study both parental alleles at the locus of interest

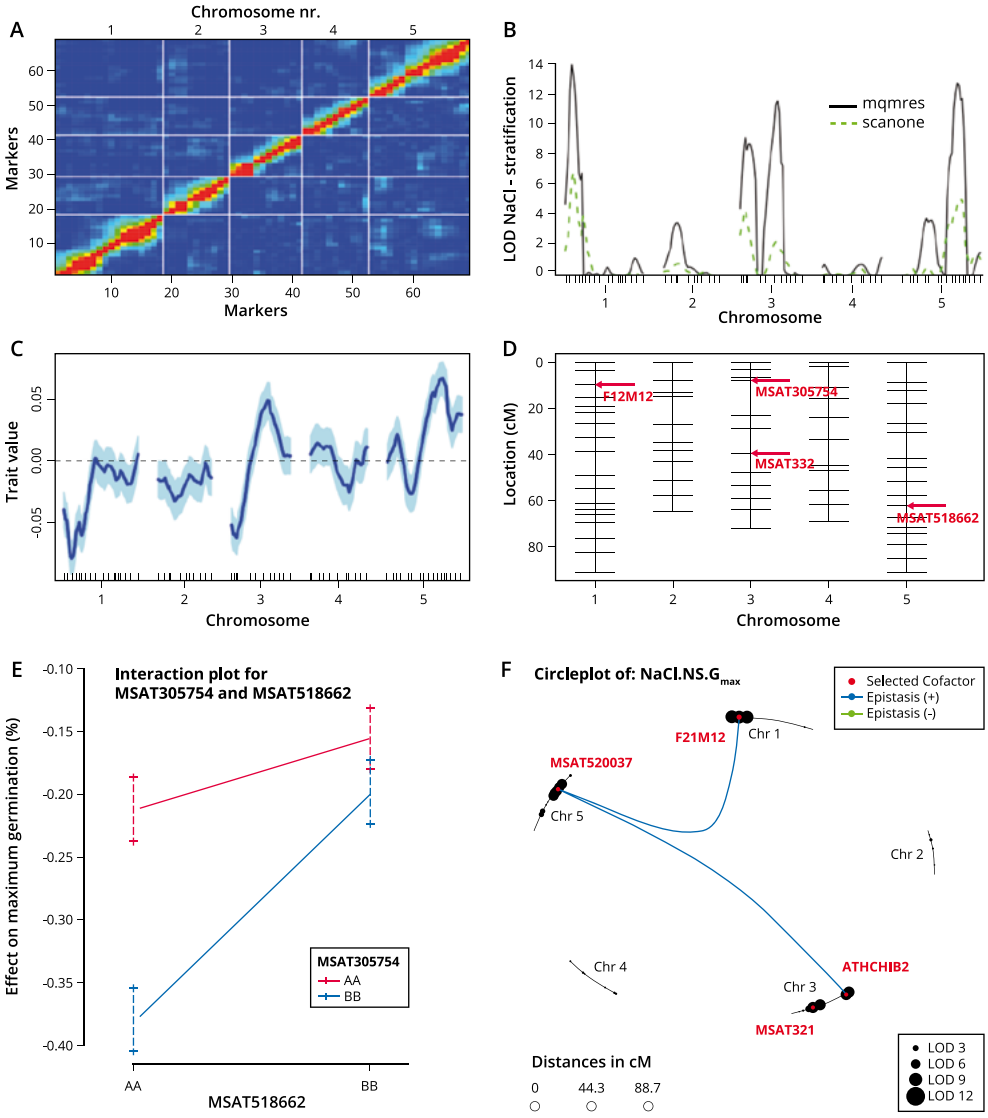


***Figure* 3.5** – *R/QTL output for the effect of 100mM NaCl on the maximum germination without stratification.* **A.** *Pairwise recombination fractions;* **B.** *LOD profile comparison between MQM and Haley-Knott (scanone) interval mapping;* **C.** *Genome wide additive effects based on raw phenotype data* **D.** *Genetic map with the significant QTL markers labelled;* **E.** *Interaction plot showing the effect size comparison between marker MSAT305754 and MSAT518662 at the Sha (AA) and Bay-0 (BB) alleles;* **F.** *Circle plot showing interactions between all significant markers.*

in an elsewhere homozygous background. In contrast to conventional near isogenic lines (NILs) the genetic background of heterogeneous inbred families (HIFs) consist of a mix of the two parental genomes. The availability of a genome wide set of HIF lines for the Bay-0 × Sha RIL population provides a fast and accurate tool to confirm detected QTL loci.

3.3.2 Results

Single trait QTL mapping
To evaluate the response of germination to a certain treatment, we first subtracted the observed germination at test conditions from germination at the proper control conditions. For example, the effect of NaCl on germination after cold stratification is determined by subtracting G_{max} on NaCl from G_{max} on water. This subtraction was reversed for the rate and uniformity parameters to correct the reversed nature of these parameters (e.g. slower germination results in a larger t_{10} and t_{50}). Table 3.2 gives an overview of all corrections that have been applied.

An analytical protocol was designed, using the popular R/qtl package of R to analyze trait data of recombinant inbred populations with the multiple QTL model approach [16]. When performing a detailed QTL analysis it is important that several steps are performed or checked. Missing genotypic data is imputed and a recombination frequency plot is generated (Fig. 3.5A). In the next step, quality of the trait data is investigated. Outliers are detected and removed using a Z-score transformation with a user defined threshold. As an extra control the results of MQM mapping were always compared to standard interval mapping, using the parametric model with Haley-Knott regression [36] (Fig. 3.5B). The whole genome additive effect was estimated based on the nontransformed data as half the difference between the phenotypic averages for the two homozygotes (Fig. 3.5C).

R/qtl MQM uses a backward elimination of cofactors. As a rule of thumb one can select a maximum of `N-20` initial cofactors with this procedure [77], with `N` being the number of lines in the RIL population. In our script, a cofactor file can be provided with the selection of the initial cofactors. When no cofactors are provided, the analysis will be performed without cofactors resulting in an analysis comparable with the composite interval mapping (CIM) method. For the analysis of the Bay-0 × Sha population we pre-selected 39 out of 69 markers as possible cofactors. Cofactors were selected based on their quality (least amount of missing data or heterozygous status) and physical cM position, attempting to obtain intervals of about 10 cM. Although the procedure allows the selection of all 69 markers as cofactors, this does not improve mapping and only lowers statistical power due to the multiple testing correction in the permutation analysis. The provided cofactor file is used to perform automated backward elimination of cofactors.

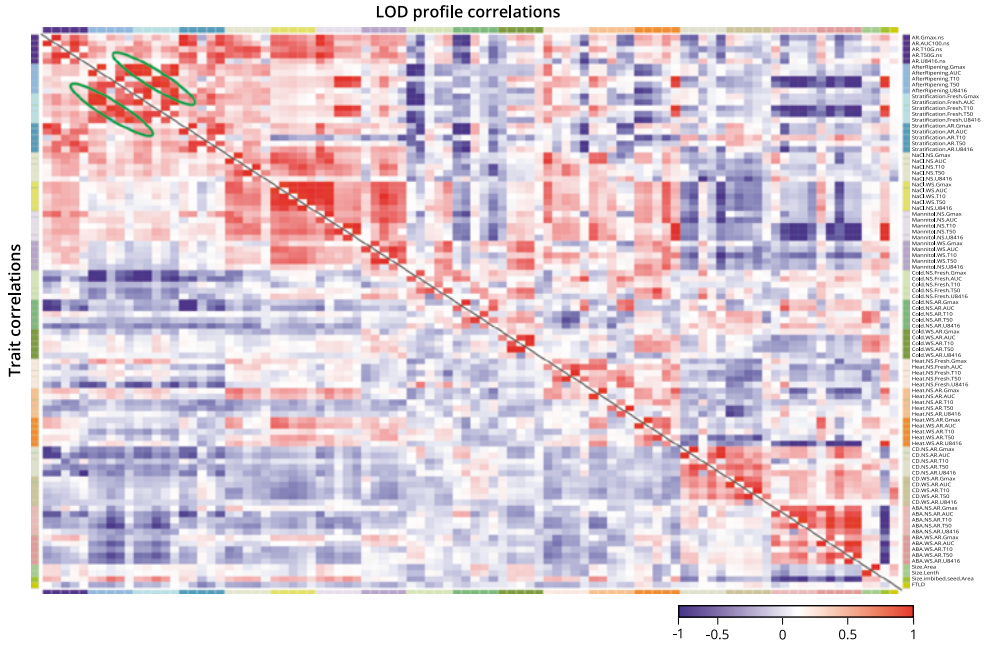


Figure 3.6 *– Correlation between all trait values (bottom left panel) and between all LOD profiles (top right panel). Traits are indicated by the color code depicted in Table 3.2. Green circles indicate an example of the close correlation between "after-ripened" and "fresh with stratification" trait values.*

Backward elimination is performed to remove cofactors that do not significantly contribute to the fit of the initial model. This is achieved by comparing Akaike's information criterions (AIC) of the different models [38]. Using the final selected QTL model, the mapping LOD scores are calculated for all genetic markers. Plots showing all significant markers are produced automatically (Fig. 3.5D). We have used the procedure described to map all 327 individual measurements but to enhance readability we only show average values for each trait (94 traits).

QTL×QTL interactions

Epistatic interactions between QTL can help to elucidate meaningful co-localizations and will enable an efficient design of follow up experiments. Besides the visualization of the epistatic interactions per trait (Fig. 3.5F) our script creates an output that can help to visualize all detected epistatic interactions in a single plot. This output file in sif format summarizes all marker-trait relations (Fig. 3.8) and detected epistatic interactions (Fig. 3.9). Among others, clear hotspots of epistatic interactions between QTL loci on

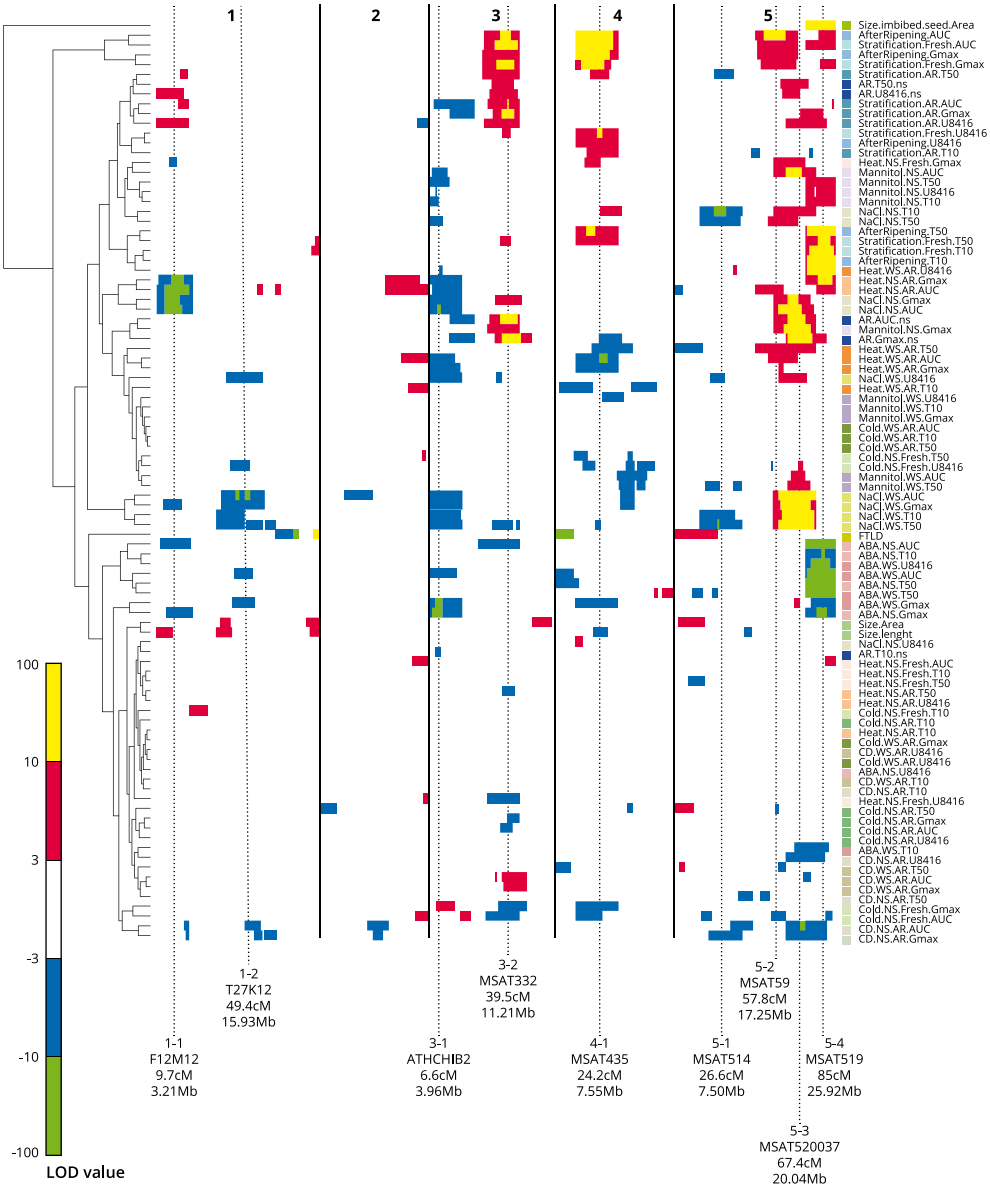


Figure 3.7 – *A clustered heat map showing the LOD profiles of the measured traits is automatically produced. Columns indicate chromosome position along the 5 chromosomes; rows indicate individual trait LOD profiles. A false color scale is used to indicate the QTL significance. Positive values (yellow and red) represent a larger effect of the treatment in Shahdara, negative values (blue and green) in Bayreuth.*

Trait Group	Stratification		Harvest	Description	Codes
Germination	N		ABCD	After-ripened seed germination	AR.NS
After ripening	N		ABCD	Delta between freshly harvested seed germination and after-ripened seed germination	AR.NS - Fresh.NS
Fresh	Y		ABCD	Delta between freshly harvested seed germination and freshly harvested seed germination	Fresh.WS -Fresh.NS
AR	Y		ABCD	Delta between after-ripened seed germination and after-ripened seed germination	AR.WS - AR.NS
NaCl	N		ABCD	Delta between after-ripened seed germination on 100 mM NaCl and after-ripened seed germination on water	AR.NS - NaCl.
NaCl	Y		ABCD		AR.WS - NaCl.WS.
Mannitol	N		AD	Delta between after-ripened seed germination on -0.5 mP Mannitol and after-ripened seed germination on water	AR.NS - AR.Mann.NS.
Mannitol	Y		AD		AR.WS - AR.Mann.WS.
Cold Fresh	N		D	Delta between freshly harvested seed germination at 10 °C and freshly harvested seed germination at 20 °C	Fresh.NS - Fresh.Cold.NS
Cold	N		AD	Delta between after-ripened seed germination at 10 °C and after-ripened seed germination at 20 °C	AR.NS - AR.Cold.NS
Cold	Y		D		AR.WS - AR.Cold.WS
Heat Fresh	N		D	Delta between freshly harvested seed germination at 30 °C and after-ripened seed germination at 20 °C	Fresh.NS - Fresh.Heat.NS
Heat	N		D	Delta between after-ripened seed germination at 30 °C and after-ripened seed germination at 20 °C	AR.NS - AR.Heat.NS
Heat	Y		D		AR.WS - AR.Heat.WS
CD*	N		D	Delta between after-ripened seed germination after controlled detoriation and after-ripened seed germination on water	AR.NS - AR.CD.NS
CD*	Y		D		AR.WS - AR.CD.WS
ABA	N		D	Delta between after-ripened seed germination with 0.5 µM ABA and after-ripened seed germination on water	AR.NS - AR.ABA.NS
ABA	Y		D		AR.WS - AR.ABA.WS
Seed size	N		ABD	Seed size and length of dry seeds	Size.Area
Seed size, imbibed	N		ABD	Seed size of imbibed seeds	Size.imbibed
Flowering time	N		ABD	First open flower in long day (16D/8N) conditions	FTLD

Table 3.2 *– Overview of traits in this study and the harvest(s) used for the measurement. The indicated color code is used in all figures throughout this chapter. For each mentioned experiment* G_{max}*, AUC,* t_{50}*,* t_{10} *and* U_{8416} *were determined. Abbreviations in codes are as follows:* AR*, after-ripened;* CD*, controlled deterioration;* NS*, no stratification;* WS*, with stratification;* Δ*, difference.*

chromosome 3, 4 and 5 (resp. ATHCHIB2 + MSAT332, MSAT435 and MSAT520037 + MSAT519) were observed for germination on salt (yellow lines) and dormancy (blue lines). Next to the importance of detecting possible interacting loci this QTL×QTL analysis provides additional arguments for co-locating QTL to be of similar genetic origin. Overall, the creation of this type of summarizing figures is greatly facilitating the interpretation of large data sets.

QTL × environment interaction

To obtain a parameter for the response, we had to correct all values with their proper control condition values. This sometimes led to complex interpretation, which can be circumvented by using the non-corrected germination parameters and model them over the various environmental conditions that were tested. Because several environments are taken into account simultaneously, the statistical power to detect loci that are affected by several environments increases, and interpretation becomes more intuitive as the need for correcting the stress response by the control response is eliminated. By using this approach the sensitivity of a specific QTL for environmental conditions can be determined for each separate germination parameter. Results are summarized in Figure 3.10. The final model P-value profiles (top panel, Fig. 3.10) clearly show the great consistency between the 5 germination parameters that we measured. However, a closer look also reveals loci that are affecting different germination curve parameters. For example, the QTL on top chromosome 5 is not detected by measuring maximum germination but is well defined when using t_{50} or t_{10} as parameter. As expected, the parameter AUC (Area Under the Curve) is outperforming the others as it represents a combined value for maximum germination percentage, rate and uniformity. For comparison of the environment-specific QTL effects for the 5 different germination parameters (5 lower panels, Fig. 3.10) the effects could be compared with germination under control conditions. For example, after-ripened seeds without stratification (AR.NS) can guide as reference for the stress treatments (AR.NS.ABA, AR.NS.CD, AR.NS.Cold, AR.NS.Heat, AR.NS.Mannitol, AR.NS.NaCl). The same analogy holds true for after-ripened seeds without stratification (AR.NS) and freshly harvested seeds without stratification (Fresh.NS). In this way stress specific QTLs on chromosome 2 and top chromosome 3 can easily be identified. Interestingly, some QTLs, including germination at low temperature (top chromosome 1) and germination in the presence of exogenous ABA (bottom chromosome 5) displayed opposite effects on germination when compared to the other treatments.

QTL confirmation

Taking advantage of the residual heterozygosity present in the F6 generation of the Bay-0 × Sha population, combined with the large population size, we were able to confirm several QTL following the heterogeneous inbred family (HIF) approach. In

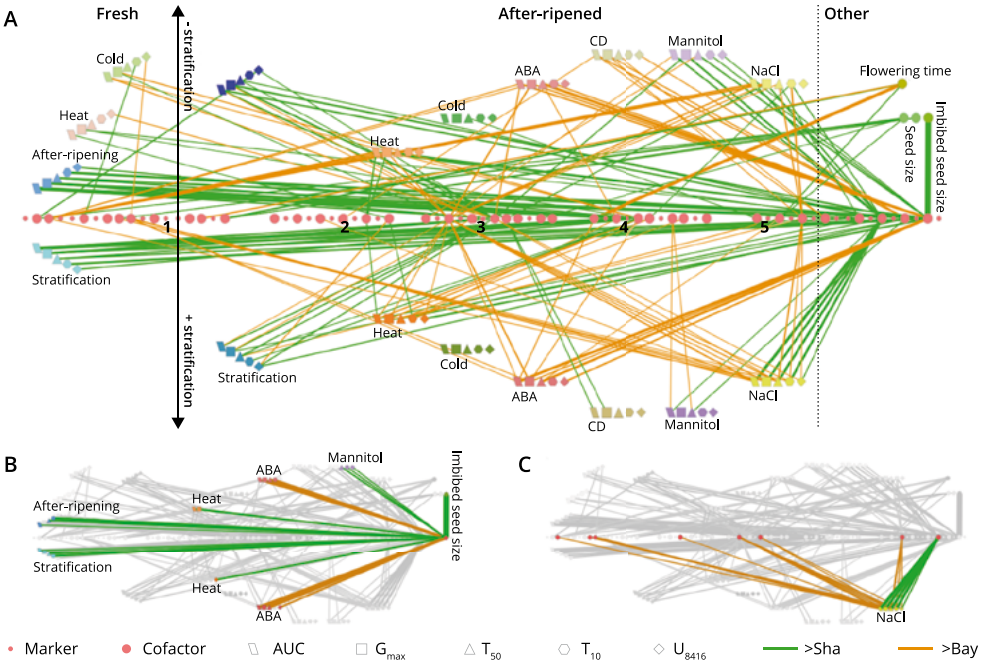


***Figure* 3.8** – *Cytoscape marker-trait network.* **A.** *Significant QTL positions are indicated by a connection between traits and markers, edge colors indicate the direction of the QTL effect, line width indicates the LOD score;* **B.** *Sub-network showing all traits with a significant QTL at marker* MSAT519; **C.** *Sub-network showing all markers with significant QTL for germination on NaCl with stratification.*

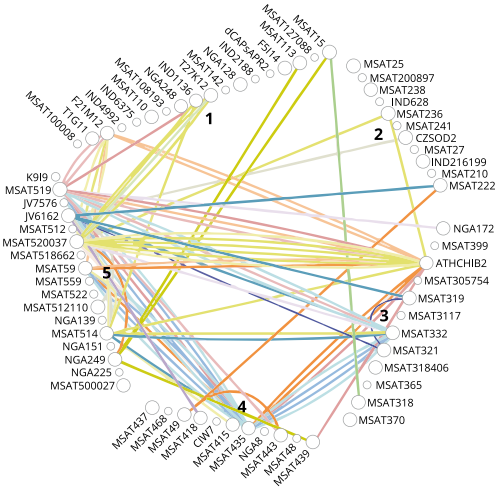


***Figure* 3.9** – *Epistatic interaction network created with Cytoscape from the interaction. sif output file. Nodes indicate markers (small circles) and selected cofactors (large circles). Edges represent the detected significant epistatic interactions (edge colors represent traits, see Table 3.2 for colors).*

short, RIL lines which are heterozygous at the locus of interest were selected in the next generation for lines homozygous for both parental alleles. These 'families' are near isogenic lines (NIL) which can be used to confirm the observed allelic effects (Fig. 3.11A). We applied this strategy for 7 of the major QTL that we detected in this study and tested the 5 germination parameters for 11 different conditions. For a single parameter (G_{max}) and a single HIF (line HIF103) the analytical procedure is summarized in Figure 3.11B. We detected a vast QTL for imbibed seed size at the bottom of chromosome 5, which could be confirmed by the use of HIF103. Upon imbibition seeds swell due to rapid water uptake and possibly because of the expansion of the inner mucilage layer. In Sha, which is a natural mutant for the MUM2 gene [116], this swelling did not occur. Also the HIF lines at the MUM2 position showed a clear difference in swelling phenotype which was still significant 24 hours after imbibition (Fig. 3.12).

3.3.3 Conclusions and discussion

When analyzing large (RIL) populations, it is hardly feasible to manually count all germination experiments several times a day to obtain germination curves. Therefore, previous studies mostly restricted to counting end-point germination [101, 102, 104, 117, 118, 119, 120, 121]. A germination curve allows QTL mapping under conditions where rate and uniformity are delayed, but maximum germination is not affected. Therefore, we used the Germinator package [122] that enabled measurement of cumulative germination data and extracting 5 germination parameters that describe the resulting germination curve. In the present study we describe several germination QTLs that were not detected before in the Bay-0 × Sha population. We observed interesting co-localizations for several germination traits and identified the loci that show large effect epistatic interactions. Among these were new loci and loci similar to the ones already found in other RIL populations (see [69] - Table 4 for the major identified QTL loci).

Dormancy

Primary dormancy has been studied extensively in various RIL populations [102]. These authors quantified primary dormancy with the DSDS50 parameter (days of dry storage to reach 50% germination), which is a good measure for after-ripening related dormancy breaking. Although we only compared the germination characteristics of freshly harvested seeds with those of after-ripened seeds and fresh seeds with and without stratification, we detected large genetic variation. Both dormancy breaking treatments showed strong QTL at positions 3-2, 4-1 and 5-2, co-locating with DOG6, DOG18 and DOG1, respectively (See [69] - Table 4). DOG18 was not detected in a Landsberg erecta × Shadara (Ler-0 × Sha) population and showed a stronger dormancy in Ler-0 as compared with An-1, Fei-0 and Kas-2 [102]. We detected stronger dormancy in Sha as compared to Bay-0 at the DOG18 locus. This suggests that both Ler-0 and Sha contain an allele

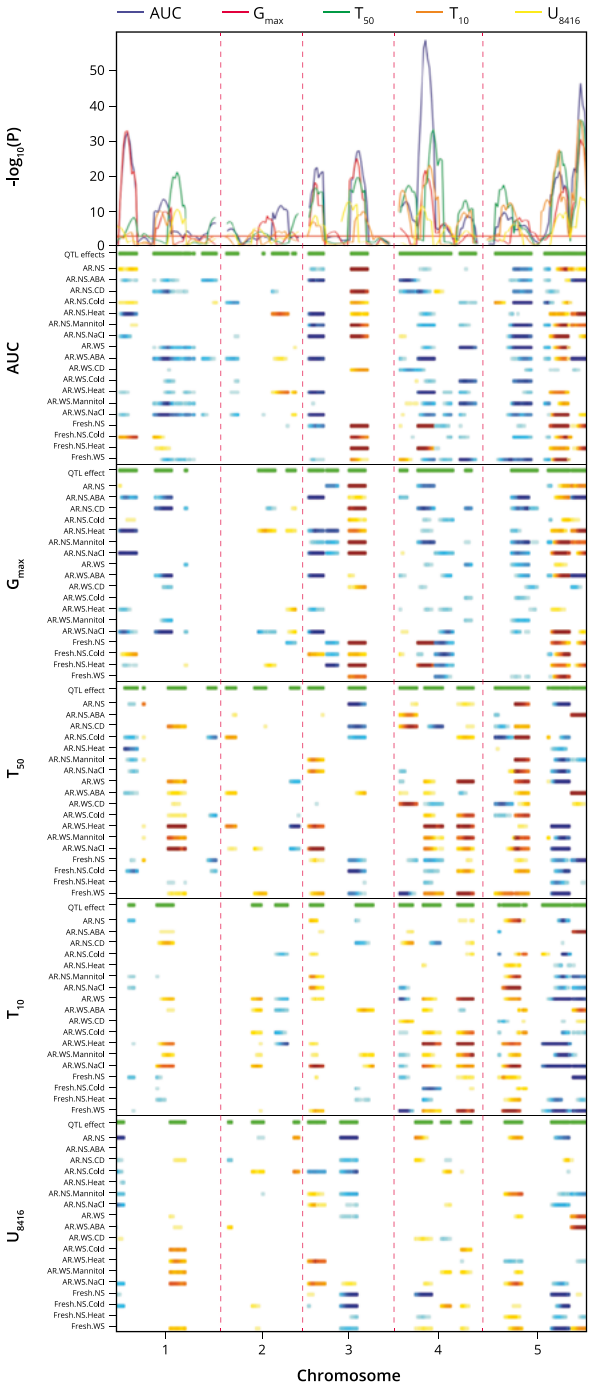


Figure 3.10 *- Genome scan for QTL × environment effects for seed germination. The P-values for the main effects of the different germination parameters are shown in the top panel. The red horizontal line is the genome wide significance threshold. The 5 bottom panels show the environment specific QTL effects. The green line indicates significant environment specific effects. For both G_{max} and AUC a bigger effect of the Sha allele is indicated in yellow-red (Bay-0 in cyan-blue). The color scale is opposite for the t_{50}, t_{10} and U_{8416} parameters due to the inverted nature of these parameters.*

of similar strength which is stronger when compared to An-1, Fei-0, Kas-2 and Bay-0. Remarkably, for both the DOG6 and DOG18 location the sensitivity to ABA was higher in Bay-0, whereas dormancy was deeper in Sha, which resulted in a directional change of the QTL effect. The more dormant Sha parent contains higher initial ABA levels and apparently, after-ripening and stratification reduce the ABA sensitivity to a greater extent as compared to the Bay-0 parent. This effect was not observed for the DOG1 locus. Further, we identified a strong effect of the dormancy-breaking treatments on the initiation (t_{10}) and rate (t_{50}) of germination at the bottom of chromosome 5 (marker MSAT519, 85 cM). The same was observed for germination on mannitol and germination at higher temperature. A QTL with opposite effect at this position was found for germination on ABA. Interestingly, these co-located with a QTL found for imbibed seed size.

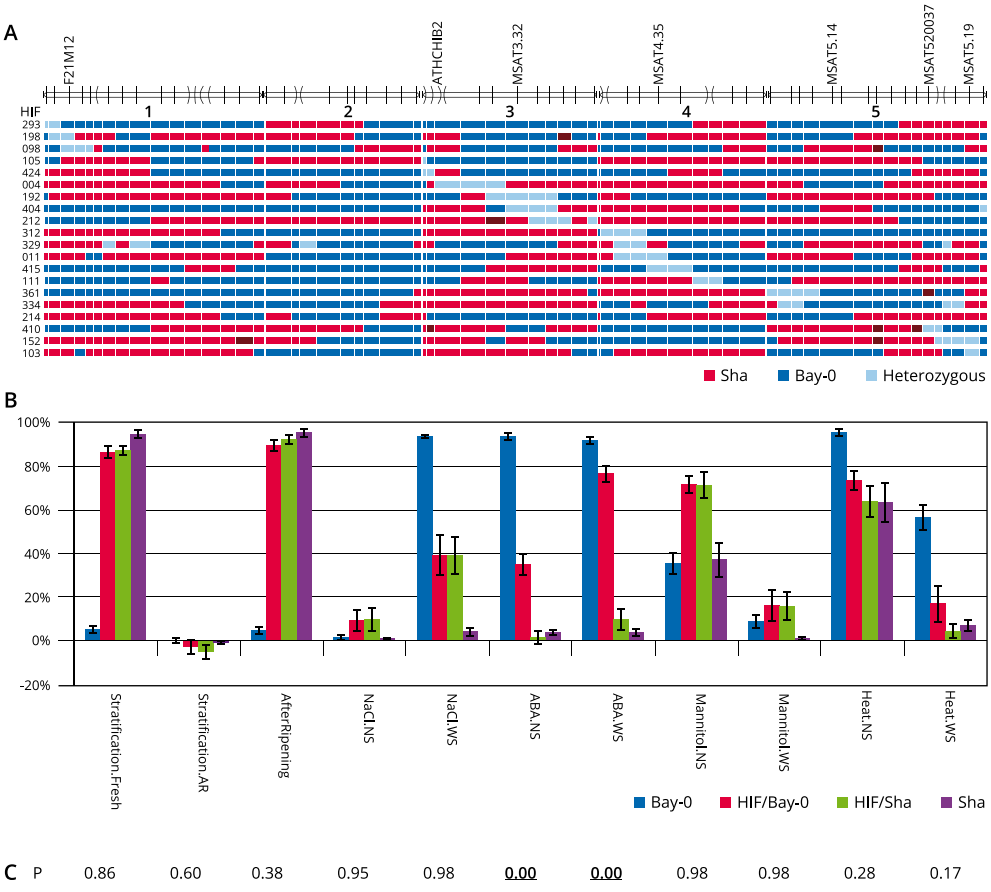
Water uptake

Initiation and rate of germination are highly influenced by the overall water potential of the seed. The mucilage layer surrounding the seed appears to play an important role in the process of water uptake [123]. Sha is a natural mucilage mutant due to a mutation in the MUM2 gene, which changes the hydrophilic potential of rhamnogalacturonan I [116]. Although mucilage has been reported to be dispensable for germination and development under lab conditions [124], a link with germination under reduced water potential conditions was shown by Penfield, *et al.* [123]. They showed reduced maximum germination of a mucilage-impaired mutant only on osmotic PEG solutions. In our study, other traits that co-located on the MUM2 locus were delayed initiation and rate of germination on osmotic mannitol solution but also on water, which clearly shows the advantage of determining a detailed germination curve. We also observed a very strong QTL for swelling of the seed in the first hours of imbibition (imbibed seed size) at the MUM2 location. Interestingly, exogenous ABA can be used to stimulate mucilage production and ABA-1 mutants are affected in mucilage production [125]. This indicates a regulatory role of ABA in mucilage production and fits with our observation of the co-localization of a QTL for initiation and rate of germination with a QTL with opposite effect for ABA sensitivity. Therefore, we hypothesize that Sha has a slower initiation and rate of germination, combined with reduced ABA sensitivity due to its mutation in the MUM2 gene. This observation may open new research strategies to define the regulatory role of ABA in mucilage production and its multiple effects on germination parameters.

Salt, heat and ABA

At the top of chromosome 1, underlying marker F12M12, we detected a strong QTL for maximum germination in the presence of 100 mM NaCl or 0.5µM ABA. A similar locus has been identified and fine-mapped in a Ler-0 × Sha population [126]. They identified a premature stop codon in the Response to ABA and Salt 1 gene (RAS1; At1g09950)

in Sha that led to a truncated protein and showed its role as a negative regulator of salt tolerance during seed germination and early seedling growth by enhancing ABA sensitivity. Here we show that a similar locus is also inferring tolerance to germination at 30 ℃. This suggests an additional role for the RAS1 gene. Increased heat tolerance due to modulation of ABA sensitivity has been shown before for other loci [127, 128].



C

P	0.86	0.60	0.38	0.95	0.98	**0.00**	**0.00**	0.98	0.98	0.28	0.17

Figure 3.11 *- Example of the confirmations performed with the HIF approach.* **A.** *The blue/red bars indicate the allelic confirmation of the HIF lines used (blue=Bay-0, red=Sha, light blue=heterozygous). The 5 chromosomes are indicated at the top with the nearest genetic marker for the 7 major loci.* **B.** *example analysis for HIF103 (segregating at MSAT5.19, bottom chromosome 5). Indicated is the maximum germination (G_{max}) for 11 conditions. Error bars represent standard error of at least 6 replicates. Responses are calculated by subtracting the test sample from the control sample as indicated in Table 3.2.* **C.** *T-test significance values for the response values (significant (`P < 0.05`) values are bold).*

Interestingly, our present study showed a strong effect of stratification which resulted in a strong reduction of significant linkage for NaCl, heat and ABA sensitivity at the F12M12 locus. A specific QTL for germination on NaCl preceded by a cold stratification period was found at the middle of chromosome 1 (marker T27K12). Also at this locus we found colocation with sensitivity for germination on ABA after stratification. Further fine-mapping at this locus might help to elucidate the effect of stratification on ABA mediated abiotic stress tolerance, as well as the apparent overlap of dormancy and stress responses. Especially interesting is QTL 5-1 (See [69] - Table 4, and Fig. 3.10) which mainly influences rate and initiation of germination. We detected this QTL for t_{50} in after-ripened seeds with stratification treatment, but also for t_{10} and t_{50} for germination on salt, regardless of a preceding cold stratification and for maximum germination after an accelerated aging treatment. One of the genes underlying this QTL interval is a nicotinamidase gene (NIC2, At5g23230), the mutant of which has retarded germination and impaired germination potential [129]. These authors suggested that NIC2 is normally metabolizing nicotinamide during moist chilling or after-ripening, which relieves inhibition of poly(ADP-ribose) polymerase (PARP enzyme) activity and allows DNA repair to occur prior to germination. Both accelerated aging and germination under salt stress conditions might require optimal functioning of this DNA repair mechanism. Further research is needed to determine whether NIC2 is causal for this QTL.

Detection of epistatic interactions in genetic studies can enhance the understanding of underlying molecular mechanisms. Recently, N. Galpaz and M. Reymond [104] showed strong epistasis in the genetic network controlling germination under salt stress in *Arabidopsis*. Due to careful dissection of the epistatic relationships they were able to show that three detected QTL rely on the presence of a Columbia allele at a QTL on top of chromosome 1. This observation led to the hypothesis that RAS1 [126] functions as a switch of the genetic network by regulating the expression of the other QTL. In another study it was found that epistasis significantly influences both fitness and germination in *Arabidopsis* [87] and novel allele combinations were identified that resulted in higher fitness. In our study we detected clear hotspots of epistatic interactions between QTL loci on chromosome 3, 4 and 5 (ATHCHIB2, MSAT332, MSAT435, MSAT520037 + MSAT519, respectively). This observation strengthens the hypothesis that some of the traits with strong QTL co-localizations indeed rely on the same underlying genetic networks.

Concluding remarks

We analyzed natural variation for many seed germination characteristics and showed their correlation, (shared) QTL positions and epistatic interactions, using a high-throughput phenotyping approach and subsequent high-throughput QTL mapping. Using the HIF approach, confirmation of some major QTL hotspots was demonstrated, which allows a fast but solid confirmation of a QTL position. Together with results from

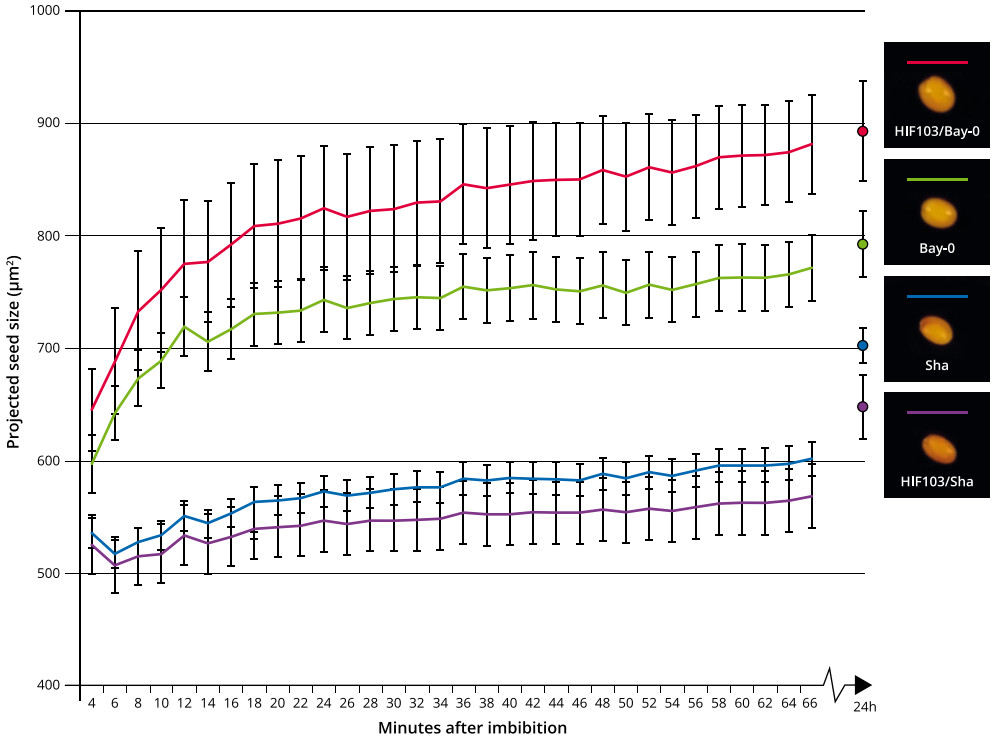


***Figure 3.12** – Different increases in seed size during the start of imbibition for Bay-0 (green), Sha (blue), HIF103/Bay-0 (red), and HIF103/Sha (purple) seeds. Shown is the average projected seed size area of 10 seeds. Error bars represent SE values. Photographs show 24h imbibed seeds.*

several other studies focusing on genetic variation in seed traits, this study has generated an extensive QTL database for Arabidopis and proposed a method of analysis to visualize the genetic landscape of seed performance. This database is a solid resource for further study. For most of the found loci in this and other studies further characterization, and in most cases fine mapping, must be undertaken to elucidate the causal molecular mechanisms. Further, we have designed a free available analysis protocol to perform detailed high-throughput QTL analysis based on the R/qtl MQM routine. In this era of large-scale phenotyping we regard a detailed analysis of QTL, QTL×QTL and QTL × environment interaction as indispensable steps to allow visualization and interpretation of multiple traits.

3.4 Metabolites in a designGG experiment

A complex phenotype such as seed germination is the resultant of several genetic and environmental cues and requires the concerted action of many genes. The use of well-structured recombinant inbred lines in combination with omics analysis can help to disentangle the genetic basis of such quantitative traits. This so called genetical genomics approach can effectively capture both genetic (G) and epistatic interactions (G×G). However, to understand how the environment interacts with genetic information (G×E) a better understanding of the perception and processing of environmental signals is needed. In a classical genetical genomics setup this requires replication of the whole experiment in different environmental conditions. A novel generalized setup overcomes this limitation and includes environmental perturbation within a single experimental design.

We developed a dedicated QTL mapping procedure to implement this approach and used existing phenotype data to demonstrate its power. Additionally, we studied the genetic regulation of primary metabolism in dry and imbibed *Arabidopsis* seeds. Many changes were observed in the metabolism which are both under environmental and genetic control and their interactions. This concept offers unique reduction of experimental load with minimal compromise of statistical power and is of great potential in the field of systems genetics which requires a broad understanding of both plasticity and dynamic regulation.

3.4.1 Background

The use of natural variation to disentangle the genetic mechanisms underlying differences in phenotypes has been very successful both in crop plants and in the model plant *Arabidopsis* (*Arabidopsis thaliana*) [112]. Most of the variation within wild or domesticated plant species is of quantitative nature determined by genetic polymorphisms at multiple loci. Such quantitative trait loci (QTL) can be analyzed efficiently using experimental mapping populations such as recombinant inbred lines (RILs) derived from directed crosses. Nowadays, many well structured RIL populations are available, often accompanied with detailed studies of phenotypic variation [130]. The complexity of quantitative traits is further determined by the interactions between genomic loci (i.e. epistasis) and between the genotype and the environment (genetic × environmental (G×E)). While epistasis can be effectively identified in QTL analyses, albeit with lower power than main effects, the detection of G×E interactions requires experimentation in multiple conditions of interest. Because of the large population sizes often needed to obtain sufficient statistical power for QTL detection, G×E interactions are usually ignored in experimental setups. However, a better understanding of the perception and processing of environmental (E) signals is greatly needed, because

interactions provide important insights in adaptation mechanisms and evolutionary constraints such as balancing and disruptive selection. To obtain a more detailed view of the molecular mechanisms underlying phenotypic variation, genetical genomics studies, in which molecular traits are genetically analyzed, have been successfully applied to enhance a directed strategy to identify causal relationships [35, 131, 132, 133]. The observed phenotype is often the resultant of a functional cascade of gene transcription followed by protein translation and modification, which finally leads to a highly dynamic metabolome underlying emergent properties [134]. With the technological advances made in genomic analytical platforms, such as transcriptomics, proteomics, and metabolomics, the large-scale, high-throughput analyses needed for quantitative genetic approaches have become feasible [31].

Incorporating developmental and environmental perturbation in the often expensive and laborious omic analyses, an alternative experimental setup, coined 'generalized genetical genomics' (GGG), using balanced fractions of a RIL population has been proposed [68]. It provides a cost-effective experimental setup for hypothesis-generating research in multiple environments. Such an approach aims for the creation of subpopulations of RILs, one for each environment to be tested, with an optimal distribution of parental alleles over all available markers [68]. When these subpopulations are subjected to environmental perturbation, the emerging phenotypes can be explained by several sources of variation: genetic variation, environmental variation, and G×E variation. Whenever the resulting phenotype is not or only mildly affected by interactions (G×E), the analysis of the different subpopulations can be combined, gaining the full power of a complete population. However, when a trait shows strong G×E interaction (e.g. those that only express genetic variation in specific environments), the power to detect QTL is dependent on those subpopulations expressing the genetic variation. Although G×E interactions have been detected previously in genetical genomics studies for expression [86, 135, 136] and metabolite content [137] by analyzing all lines in a population under different environments, the GGG concept offers an effective way of studying a combination of genetic and environmental perturbations and is of great potential in the field of systems genetics, in which a broad understanding of both plasticity and dynamics is required [138]. The fundamental basis of the experimental design and data analysis using a full model ($Y = E + G + G{\times}E + \varepsilon$), where Y is the observed phenotype and ε is residual error, is generally valid and frequently used [86, 136, 139]. As a proof of principle, we present experimental data on the genetic regulation of primary metabolism in dry and imbibed *Arabidopsis* seeds using a GGG design and discuss the application and implications of such a strategy.

Plants are extremely rich in biochemical compounds, and major roles in plant development, adaptation, and defense have been identified for biosynthesis pathways

and their products [165]. The biosynthetic pathways of primary metabolites are well studied and often well conserved between different taxa [140]. Nonetheless, quantitative variation for many of these compounds can be observed between natural variants, which might be reflected in their different growth characteristics. The analysis of single-gene mutants, for example, has unraveled many key components in biochemical pathways and has demonstrated their role in phenotypic traits [141]. In *Arabidopsis*, genetic variation for many of its metabolic compounds has been observed [133, 142, 143], but G×E interactions were ignored in these studies and only addressed by Chan *et al.* [144]. Metabolic profiling at different growth stages has further revealed important fluxes that regulate plant development and adaptation [145]. Using the accumulated historical mutations that occur in natural variants in combination with metabolic profiling in a generalized design offers the unique possibility of identifying genetic effects over a series of developmental stages. Here, we report on the interaction of four different physiological environments (i.e. developmental stages) in dry and imbibed seeds with two founder genotypes in a RIL population. To detect the majority of the most prominent primary metabolites, we used gas chromatography-mass spectrometry of polar extracts [146, 147]. These include essential metabolites such as sugars, amino acids, and organic acids, which are key compounds in reserve storage, catabolism, growth, and energy metabolism.

The switch from a dry seed, which is equipped for optimal survival and storage of reserves, toward an imbibed seed, in which energy needed for germination is released and which prepares for autotrophic production, is remarkable. Reserves that have been stored during seed maturation are degraded and remobilized during germination [148, 149], a process that is heavily influenced by the capacity of carbon/nitrogen partitioning of a maturing seed [150]. *Arabidopsis* mutants affected in their oil reserve content or its mobilization show delayed but not full inhibition of germination [149, 151, 152, 153]. This suggests an additional metabolic switch that occurs during seed desiccation after seed maturation involving a change from accumulation of oil and storage proteins to the synthesis of free amino acids, sugars, fatty acids, and their degradation products functioning to prepare for rapid metabolic recovery during imbibition [154, 155]. Imbibition of mature seeds specifically shows reduction of the metabolites that accumulate during the desiccation period. Upon germination, an increase of many metabolites, including amino acids, sugars, and organic acids, can be observed again, which reflects the increase of autotrophic activity [154]. Profiling the primary metabolome over different developmental stages in a mapping population is therefore expected to reveal the dynamics of genetic regulation of many of these important processes. We will demonstrate here that much of the observed variation in biochemical profiles can be attributed to genotype-by-environment interactions, which can be effectively identified in a GGG approach.

3.4.2 Results

Experimental design
Previous studies which focused on the comparative analysis of developmental and metabolic variation suggest a link between central metabolism and plant physiology, but genetic co-regulation is not frequently observed [143, 156]. That said, in several studies in *Arabidopsis* a major metabolite QTL cluster is associated with the ERECTA locus, representing a strong regulator of development which is known for its pleiotropic effects [215]. To circumvent this strong bias we used two natural variants, Bayreuth-0 (Bay-0) and Shahdara (Sha), which are not polymorphic for the ERECTA locus. The Bay-0 × Sha RIL population [67] has previously been shown to contain genetic variation for seed germination [69] and other physiological traits [92, 93, 94, 99, 100, 101, 157], anion strength [94], carbohydrate content [97], gene expression [35] and primary [133] and secondary metabolite levels [158].

Powerful mapping of genetic variation in a RIL population is dependent on the size of the population, the level of recombination and on an evenly genomewide distribution of the parental alleles. In the present study, a core set of the Bay-0 × Sha RIL population [67] consisting of 165 lines and optimized for the aforementioned factors is used. This core population was carefully divided in four sub-populations optimized for the distribution of parental alleles using the R-package designGG, aiming at the most accurate estimate of genetic effect and G×E effect [68].

Comparison of different designs using classic phenotypes
Standard QTL mapping procedures can efficiently capture genetic variation and epistasis, but do not take environmental perturbation into consideration. Appropriate modeling of the genetic variance-covariance (VCOV) in the data is of great importance when combining information from different environments in QTL analysis [139]. Linear models are particularly well suited for this. Here environmental differences are incorporated as an additional variable in a generalized design (GGG design). To enable mapping of the observed trait variation and taking the four developmental stages into consideration an R-script was developed which uses functions and data structures from the R/qtl package [15, 16]. The R-script uses a linear model to calculate the likelihood of genotype to phenotype linkage for each marker with the following formula:

$$y_i = \beta_0 + \beta_1 e_i + \beta_2 g_i + \beta_3 (g_i \times e_i) + \varepsilon_i$$

Where y_i is the ith observation of the studied phenotype, variable g_i is the genotype, e_i is a vector with seed conditions, and $g_i \times e_i$ the interaction term. The values β_j represent parameters to be estimated, and ε_i is the error term. The simplified description `Y = E +`

G + G×E + ε of this linear model will be used henceforward. Separate likelihood estimates (-$\log_{10}$(probability), henceforth LOD scores) are generated for the environmental (E), genetic (G) and genetic × environmental (G×E) effects.

To validate the use of a GGG design, we studied the genetic (G) and the interacting effects between genetics and environment (G×E) on phenotypes in four different environmental conditions (E). These phenotypes were obtained by studying different germination parameters under different environmental conditions [69]. In total we compared the power of different designs by performing QTL analysis for 96 classic phenotypes under 4 different environments (Table 3.3) [69]. Furthermore, we also investigate the interacting effect between genotype and environmental. The full model mapping (Y = E + G + G×E + ε) was applied to full-block design, random design and GGG design. Single maker mapping (Y = G + ε) was applied to single block design. The number of detected QTL and interacting QTL (FDR = 0.05, based on >10,000 runs permutation) with the different designs are shown in Table 3.3. In the full-block design, all samples were allocated to the four conditions. Obviously, this is the most expensive way of performing the experiment as the required resources and effort is quadrupled (4 × N). As a result of the size of the experiment, the power of detecting genetic effects is the best for this design. Unfortunately, we cannot afford such expensive experiments in many situations due to limited resources and/or time. The single block design only focuses on one of the four conditions, as in most published genetical genomics studies to date. In this way the effect sample size for the selected condition is N and we will have equal power as full-block design for detecting the genetic effects for this particular condition. Clearly, this design will miss the information from the other three conditions, and interacting effects between genetic and environmental factors cannot be investigated. In order to study both genetic and interacting effects with a limited budget, the random and the GGG design allocate the N different samples to the four environments evenly, measuring N/4 samples in each condition. Although the possibility to detect genetic effects is only slightly better for the GGG design, the detection of interacting QTL is clearly improved in the GGG design as compared to the random design. These results show that the optimal allocation of samples as in the GGG design clearly improves the ability to detect both genetic and interacting effects and that the GGG design results in the maximisation of detected variation in relation to the necessary resources with only a minimal compromise of statistical power as compared to the full-block design.

Metabolic analysis

To study the metabolic status of *Arabidopsis* seeds during germination, four biologically important developmental stages of seed germination with expected variation in metabolite levels to different extent were selected. The first two stages, being freshly harvested primary dormant (PD) and after-ripened (AR) non-dormant dry seeds,

respectively, are expected to comprise a very similar metabolome as most, if not all, metabolic fluxes are arrested in the dry seed. The oil rich (~40%) *Arabidopsis* seeds [159] typically desiccate to moisture contents below 5% which results in an arrest of all enzymatic reactions due to the lack of free water. The other two stages represented early imbibition of seeds, imbibed for 6 hours (6H), and seeds at radical protrusion (RP), respectively. Full rehydration of dry seeds typically completes in less than 2 hours and although developmental differences are not yet expected, many metabolic processes will have started after 6 hours of imbibition [160, 161]. Radicle protrusion marks the endpoint of germination sensu stricto and is known to be accompanied by a major switch of both the transcriptome and metabolome [154, 160]. These four developmental stages are anticipated to vary to different degrees in their metabolic profiles: hardly any difference between dry seed samples, some differences between dry and imbibed seeds, very pronounced differences between dry seeds and seeds at radicle protrusion.

Design	**QTL Interacting**	**QTL**
Full-Block	96	30
Best power for G effect		
Most expensive		
Best power for G×E effect		
Single-Block	93	0
Same power for G in selected condition		
Less expensive		
Missing G×E effect		
Random	78(75)	17(5)
Limited power G effect		
Less expensive		
Limited power G×E effect		
designGG	81(67)	27(12)
Optimal power for G effect		
Less expensive		
Optimal power for G×E effect		

***Table 3.3** – Comparison of different experimental designs to study genetic and G×E effects on classic phenotypes in four different conditions. In total there are 164 genetically different RILs, and the data were analyzed in four different ways. The last two rows compare the number of QTLs for the main genetic effect and G×E interacting effect detected using different design strategies. The numbers in parentheses indicate the QTLs that share confidence intervals (1.5 drop-off) with the full-block design.*

To determine the metabolic status of genetic variants in these different developmental stages, all individuals in the four subpopulations and their parental accessions were subjected to GC-TOF-MS. Each sample consists of the polar fraction of a methanol extract of a bulk of approximately 700-1,000 seeds (20 mg). Samples were analyzed in random order and interspersed with pooled sample controls to control for experimental errors. The metabolic profiling of the segregating RILs was performed and the use of segregation population provides an intrinsic replication for each genotypic marker [31]. In total 7,537 mass peaks were detected, representing 161 metabolites according to centrotyping based on retention time and correlation structure [162]. In total 63 metabolites could be annotated using an in-house constructed library and a publicly available mass spectra library [163]. Parental accessions were measured in duplicate for all four developmental stages allowing us to model the influence of condition and accession using a multi-factor univariate analysis of variance (ANOVA).

$$y_i = \beta_0 + \beta_1 \text{condition}_i + \beta_2 \text{accession}_i + \varepsilon_i$$

Analysis of variance for the parental samples identified 108 metabolites showing significant variation (FDR < 0.05) between developmental stages (E) and 85 showing variation between the parents (G) with an overlap of 54 metabolites showing variation between both variables in an interactive way (G×E). For 37 metabolites no significant variation was detected between the parental accessions or in any of the developmental stages. A self-organizing map (SOM), created from the metabolites showing significant variation between the parents, groups different metabolites according to their accumulation pattern over different genotypes and developmental stages (Fig. 3.13). Clearly different patterns of variation can be observed, namely genetic in panel A and H; environmental in panel C and D; genetic + environmental in panel B and G and genetic × environmental in panel E and F, illustrating the complex regulation of metabolic processes and the need for sophisticated analysis methods, like PCA or Multiple QTL mapping [16].

Because metabolite levels are varying between both parents and between the chosen seed germination stages, a segregation of metabolic accumulation can be expected in the RIL population of 164 lines. A principle component analysis of the metabolic profiles, revealing the internal structure in the data, shows that the first component clearly separates 6-hour imbibed seeds and seeds at radicle protrusion from both primary dormant and after-ripened seeds, explaining 37% of the total variation. This confirms the large metabolic changes accompanying the transition from dry arrested seeds to the imbibed and germinating developmental stages. As expected, no obvious differences could be detected between the metabolomes of primary dormant and after-ripened dry seeds. The second component, explaining 11% of the total variation, sharply

separates the parental accessions, indicating that this component explains most of the genetic variation in metabolic profiles. These results demonstrate that Bay-0 and Sha possess substantial genetic variation for the accumulation of primary metabolites which segregates in their recombinant offspring and which is strongly influenced by the developmental stage used for profiling.

Transgressive segregation (when the expression of the offspring exceeds the expression of the parents) was visualized by comparing parental and RIL metabolite level

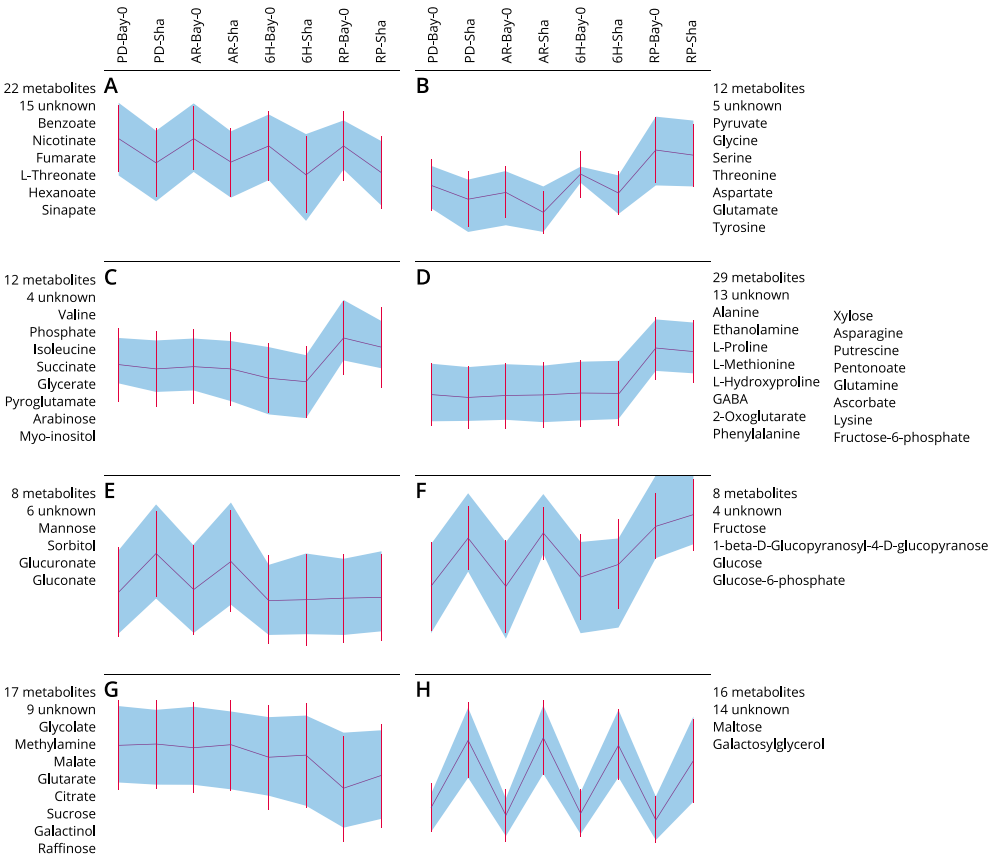


Figure 3.13 *– Self-organizing map, grouping different metabolites according to their accumulation pattern over different genotypes and developmental stages of significantly variable metabolites (`ANOVA p < 0.05`) measured in the parental lines Bay-0 and Sha in four developmental stages. PD=Primary dormant, AR=After-ripened, 6H=6 hour imbibed, RP=seeds at radicle protrusion. Two independent biological replicates were measured for each combination of parent and developmental stage.*

distributions (Fig. 3.14). Some positive and negative transgression is observed for most of the metabolites in which the metabolite accumulation in a RIL is respectively higher or lower compared to the respectively highest or lowest parent. In addition, 15 metabolites were detected in RILs which were not present in either parent. This suggests that new allele combinations in the RIL population resulted in enhanced accumulation or even novel formation of metabolites.

QTL mapping of metabolites in a generalized genetical genomics design

In the experimental setup of this study, the environmental variation is defined as variation observed between the four developmental stages (PD, AR, 6H, and RP). Significance thresholds, determined by permutation analysis ($n = 1{,}000$, $P < 0.01$) for each metabolite, ranged from LOD 3.43 to LOD 3.50 and was stringently set to LOD 4 for all analyses. Mapping resulted in 120 significant QTLs in the genetic component for 83 metabolites and 31 G×E QTLs for 27 metabolites, ranging from one to four QTLs per metabolite. Thirteen of the G×E QTLs are significant in the genetic component as well. For 66 metabolites, no significant QTL was detected.

To test the performance of the generalized mapping procedure, QTLs detected in individual environments using the linear model $Y = G + \varepsilon$ were compared with QTLs detected in the combined mapping approach (using the linear model $Y = E + G + G{\times}E + \varepsilon$; Fig. 3.15). QTLs were binned in upper or lower chromosome arms to reduce the effects of small positional shifts. Results were plotted in a network, with nodes representing QTLs connected with edges to nodes representing the mapping populations in which they were detected (Fig. 3.15). QTLs are grouped in three sections according to their detection in the different mapping procedures. The middle section shows 73 QTLs that were detected in both the $Y = E + G + G{\times}E + \varepsilon$ model and in one or more single-environment mappings using the $Y = G + \varepsilon$ model. This shows that most of the G variation present in the single environments can effectively be captured by using the generalized model.

The presence of 60 QTLs that were only significantly detected in the $Y = E + G + G{\times}E + \varepsilon$ model (right section) shows the combined power of the generalized approach and the usage of more genotypes. These QTLs are not detected in the single-environment mapping in which only 41 individuals were used. Combining all data across all environments in the linear model increases power to detect QTLs, but it should be noted that there are also 20 minor QTLs (left section) that are only significant in the single environment mapping using the $Y = G + \varepsilon$ model. These QTLs are not detected in the $Y = E + G + G{\times}E + \varepsilon$ model. This can be explained by two factors: (1) environments in which the QTL is not expressed introduce noise in the experimental data and thereby decrease mapping power, and (2) deviations from a balanced allele distribution in the different

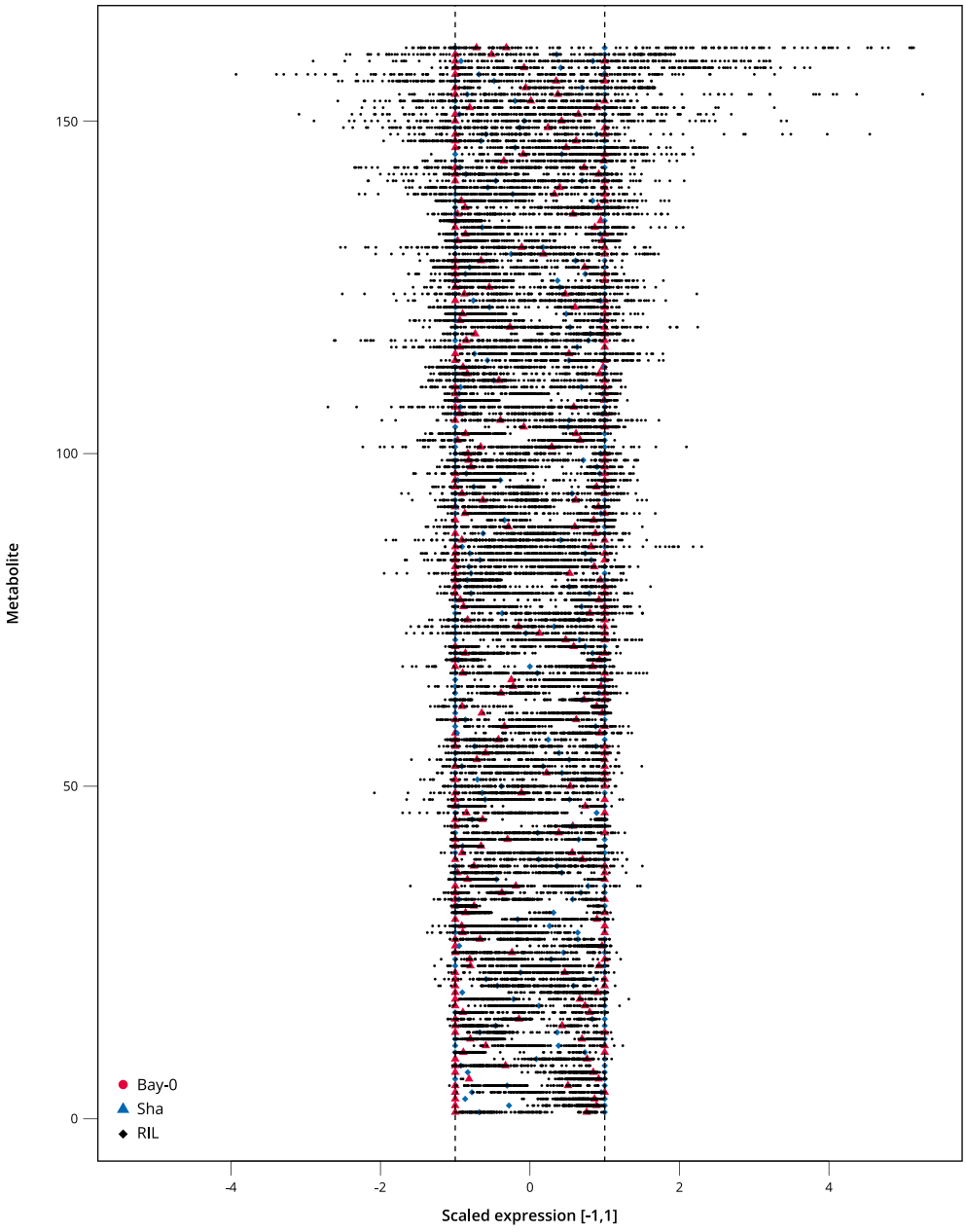


Figure 3.14 *– Positive and negative transgression is observed for most of the metabolites in which the metabolite accumulation in a RIL is respectively higher or lower compared to the respectively highest or lowest parent.*

subpopulations can introduce some stochasticity around the threshold level, although this is not the case in our data.

Importantly, all major-to-moderate-effect-size QTLs could be detected using the generalized model, even when these QTLs were not detected in the separate environment models. Although it is difficult to compare power with the latter models, because population sizes differ, the generalized design efficiently identifies all relevant QTLs, which were detected by the four separate models, and in addition, it detects G×E interactions. In a general exploratory study, the reduction in experimental burden therefore amply outweighs the incidental failure to detect the limited number of small-effect QTLs. The application of a GGG design can thus be an important advancement in evolutionary and ecological studies assessing the contribution of genetic and environmental effects to natural variation in life history traits.

For breeding purposes, the allelic effect size is an important measure, and differentiation of the environment in which the allelic effect is expressed can be very useful. In the generalized setup, the allelic effect size of those metabolites with significant QTLs is separated per environment. For every QTL that is consistently detected in all four conditions, a LOD score for G effect (Fig. 3.16, x-axis) is obtained from full-model mapping. For these QTLs, normalized allelic effect sizes are calculated by Z-score transformations for each environment (Fig. 3.16, y-axis). QTLs detected (Fig. 3.16A) show an expected linear relationship between LOD score and effect size in all measured environments. This correlation is much weaker for QTLs detected in the G×E component of the linear model (Fig. 3.16B) because the QTL is not expressed in all environments. QTLs of metabolites with strong G×E interaction, therefore, display larger effect sizes in fewer environments compared to QTLs of similar significance levels expressed in all conditions.

Clearly, the choice of environments used in these studies is crucial [138]. Limited power can be expected when environments vary too much and no overlapping genetic variation is present, and contrarily, there is hardly any additive value of the design when using very similar environments. In this study, we carefully selected four biologically relevant developmental stages of seed germination with expected variation in metabolite levels to different extent and consider them as an environmental factor in the follow-up statistical analysis. The selected developmental stages start from PD dry seeds to seeds at the point of RP. The first two stages, being freshly harvested PD and AR non-dormant dry seeds, respectively, are expected to comprise a very similar metabolome, as most, if not all, metabolic fluxes are arrested in the dry seed. The other two stages represent 6H seeds and seeds at RP, respectively.

Genetic regulation of metabolic traits

One of the most rewarding benefits of the generalized approach is the possibility to analyze metabolic fluxes over different environments or developmental stages in addition to the effect of genetic variation. The acquired information of both sources of variation can be effectively displayed in so-called flash cards in which line graphs illustrate the genetic and environmental effect and detected QTLs are plotted in heat bars (Fig. 3.17). The individual components of the linear model $Y = E + G + G \times E + \varepsilon$

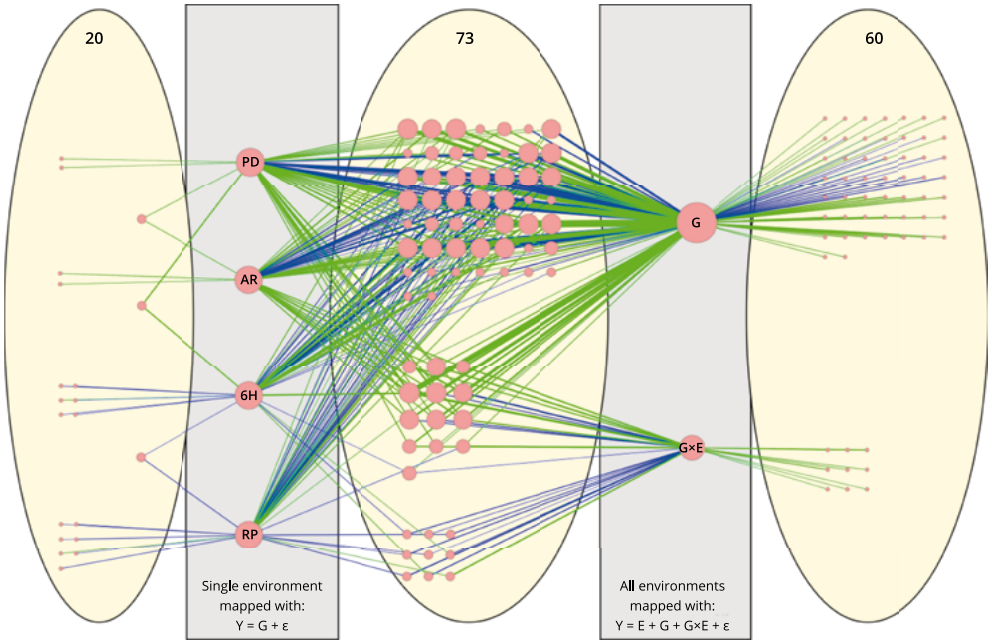


Figure 3.15 *– Comparison of QTLs detected within single environments (PD, AR, 6H and RP) by using the simple* $Y = G + \varepsilon$ *model with QTLs detected when combining environments via the full* $Y = E + G + G \times E + \varepsilon$ *model. QTLs were binned to two regions per chromosome (i.e. starting and ending regions). When comparing QTLs of a single trait from two models, they are considered as shared ones if QTLs fall in the same region. In total we found 73 QTLs shared between two models, as shown in the middle ellipse. There are 20 and 60 QTLs that are only detected in simple and full model, respectively. Nodes indicate metabolite QTLs and node size shows the degree of connectivity. Nodes are connected by edges which show the link between a QTL and a mapping population (single environments versus multiple environments). Separate nodes are created for the genetic (G) component and the genetic × environmental (G×E) component. Edge line color represents direction of the QTLs, green for higher levels in Sha; blue for higher levels in Bay-0. Edge width indicates increasing LOD scores.*

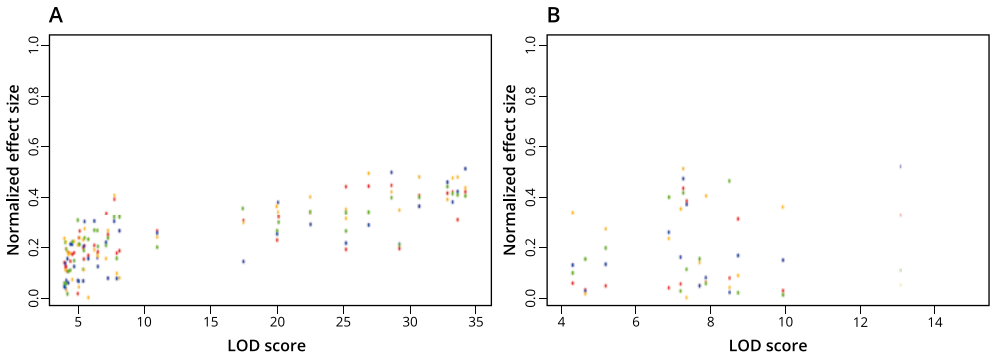


Figure 3.16 *– Effect sizes for each individual developmental stages are plotted against the derived LOD score.* **A.** *Normalized allelic effect size per environment against LOD scores from the genetic (G) component and* **B.** *Normalized allelic effect size per environment against LOD scores from the genetic × environmental interaction (G×E) component. Colors indicate the developmental stages (red = primary dormant (PD); blue = after-ripened (AR); green = 6 hours imbibed (6H); orange = seeds at radicle protrusion (RP).*

provide the valuable measures for the various sources of variation. For example lysine content strongly increases in germinating seeds, indicated by a significant LOD score of 16.1 for the environmental effect, but no genetic variation for lysine could be detected (Fig. 3.17A). For this metabolite genetic variants vary indistinguishable from each other over different environments. In contrast, fumaric acid shows little variation between the developmental stages (`LOD = 0.6`), but displays strong genetic variation explained by a highly significant QTL (`LOD = 6.5`) for the genetic effect at the center of chromosome 2. Higher levels for fumaric acid are detected in all developmental stages for those lines harboring the Bay-0 allele (Fig. 3.17B). An example of the additive effect of environmental and genetic factors is the decrease in levels of malic acid in imbibed seeds. Here a strong environmental effect (`LOD = 13.2`) is accompanied with an additional genetic effect explained by a genetic QTL (`LOD = 6.9`) at the bottom of chromosome 1. Note that the genetic effect here is similar in all environments (Fig. 3.17C). This is not the case for gluconic acid which levels are strongly affected by the interaction between the genotype and the environment. A strong G×E QTL (`LOD = 10`) is detected at the top of chromosome 4. The Sha allele at this position causes higher levels of gluconic acid in dry seeds, but not in imbibed seeds (Fig. 3.17D). This strong negative environmental effect (`LOD = 6.6`) is also responsible for the apparent directional shift of the G×E QTL effect.

Similar to the self-organizing maps in Figure 3.13 flashcards can be instrumental in the identification of metabolic relationships with the added value of genetic regulatory

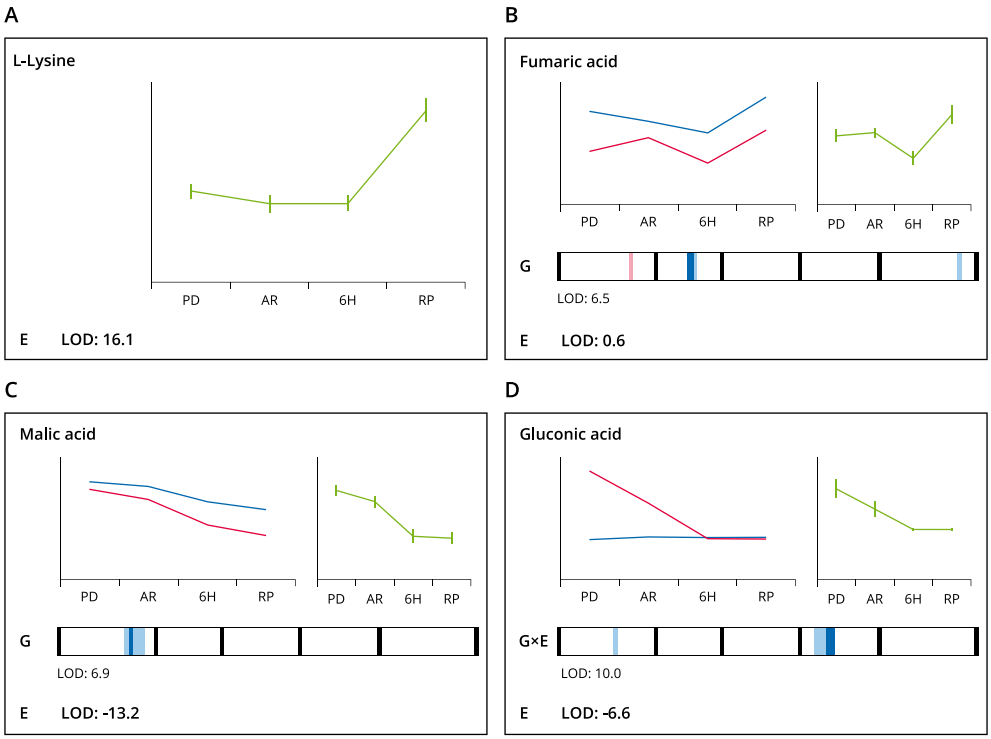


Figure 3.17 *– Normalized metabolite changes during 4 developmental stages (PD, AR, 6H and RP). Each panel represents a single metabolite and contains information about environmental variation (green line plot, average over all lines within a single developmental stage) and genetic variation (blue lines represent the metabolite levels for lines carrying the Bay-0 allele for the most significant QTL and red lines those for the Sha allele carrying lines). QTL profiles for metabolites with either genetic (G) or genetic × environmental (G×E) variation are indicated at the bottom of each panel by a heat bar representing the 5 chromosomes, and a false-color scale is used to indicate the QTL significance. For genetic QTLs, positive values (light and dark blue) represent a larger effect on the metabolite content for the Bay-0 allele, and negative values (light and dark red) represent a larger effect on the metabolite content for the Sha allele. Interpretation of the color scale for G×E QTLs is less intuitive because strong negative environmental effects can result in inversion of the QTL LOD score (e.g. gluconic acid). The presented effectplot (left line plot) shows the true allele effect. Environmental (E) variation is expressed as LOD score in the lower left corner. Depending on the most significant variation either, genetic (G) or interaction (G×E) effects are also indicated with LOD scores in the lower left corner below or above the heat bar respectively.* ***A.*** *L-Lysine showing only Environmental (E) variation;* ***B.*** *Fumaric acid: showing Genetic (G) variation;* ***C.*** *Malic acid showing both Environmental and Genetic variation (G+E);* ***D.*** *Gluconic acid showing interaction between Environment and Genetic variation (G×E).*

information. This is illustrated by integrating flashcards of all metabolites that were identified in this study with a general *Arabidopsis* metabolic pathway diagram (www.KEGG.jp). For instance, several pathways in carbohydrate metabolism, such as the biosynthesis routes for galactose, pentose phosphate, starch/sucrose and amino and nucleotide sugars, are highly interconnected and are therefore subject to co-regulation mechanisms. A number of compounds involved in different subparts of the carbohydrate network module (e.g. glucose-6-phoshate, maltose, mannose, glucuronic and gluconic acid) indeed share a strong QTL at the top of chromosome 4. This suggests that the observed variation for these compounds has a single genetic basis, possibly affecting competition for a general precursor or directing feedback loops. In addition many of these compounds show strong positive or negative correlation due to environmental control. Genetic co-regulation was also observed for amino acid metabolism. Amino acids are substrate for the synthesis of aminoacyl-tRNAs which in turn are essential substrates for translation [164]. A single G×E QTL at the bottom of chromosome 1 was detected for eight amino acids explaining a large part of the observed genetic variation. The joined analysis of environmentally and genetically induced variation in metabolic profiles can thus identify causal relationships between different modular parts of metabolic networks and associate these connections with relevant biological processes.

Regulatory hotspots and physiological co-regulation

As noted, the accumulation of several metabolites maps to identical positions suggesting that these might be regulated by a common genetic factor. Although co-locating QTLs can be the result of independent closely linked genetic factors, such coinciding QTLs are expected to occur more or less randomly by chance. Any deviation from expected frequency distributions along the genome thus hints at genetic co-regulation [46]. When plotted against their genomic position eight of such suggestive QTL hotspots can be

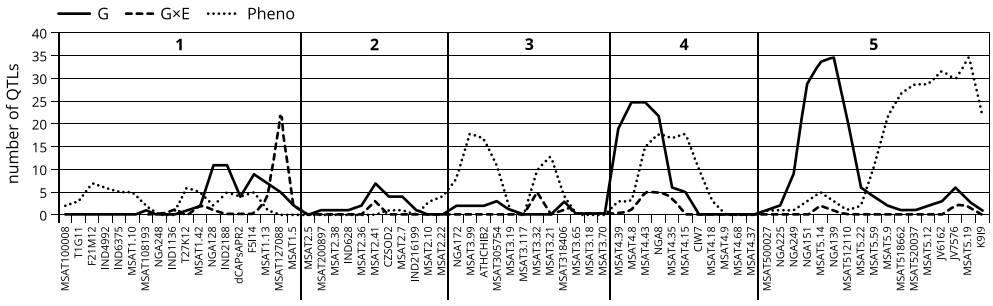


Figure 3.18 *– Number of significant QTLs plotted against the genetic location. Metabolic QTLs are represented by the solid (genetic component; G) and dashed (genetic × environmental component; G×E) lines. Germination related QTLs [69] are shown by the dotted line.*

seen (Fig. 3.18) of which the two major ones (Chromosome 4-MSAT4.8 and Chromosome 5-NGA139) co-locate with previously identified hotspots for metabolic regulation [133, 142, 143, 158]. Interestingly, both loci have been shown to play a role in glucosinolate biosynthesis. The AOP locus at chromosome 4 regulates side chain modification while the MAM locus at chromosome 5 determines chain elongation, but these compounds are not targeted for in Gas Chromatography-Mass Spectrometry (GC-MS) analysis which predominantly detects primary metabolites. As for many glucosinolates, for some metabolites, including GABA and maltose, QTLs were detected at both positions. In other cases a single QTL was detected at chromosome 4 or 5, e.g. glucose-6-phosphate and tyrosine, respectively. Although the identified primary metabolites are not directly connected with the glucosinolate biosynthesis pathway, such associations have been reported before [133]. These results might suggest alternative functions for AOP and MAM or a role in resource competition and allocation in central metabolism. This suggestion is further supported by the fact that these loci link to flowering time and the circadian clock regulation in the Bay-0 × Sha population [144]. It also cannot be ruled out that other genes overlapping the AOP or MAM regions are causal for the observed variation.

Since many metabolites appear to be co-regulated, the strong impact of some loci on central metabolism might also exert its effect on physiological traits. Recently, the genetic landscape of seed germination in the same population has been described for which seed germination parameters were acquired under a wide range of environmental conditions [69]. A comparison between variation in germination characteristics and metabolite levels might reveal compounds involved in the process of germination. Although no clear collocation of hotspots for germination and metabolite QTLs could be observed, incidental coincidence between isolated QTLs of both types of traits did occur. For instance, genetic variation for seed size co-locates with a large metabolic QTL cluster on the lower arm of chromosome 1 (75 cM). This cluster contains many QTLs for amino acids, but also for components of the tricarboxylic acid (TCA) cycle (e.g. fumarate and malate). In plants, leucine, isoleucine and valine can be broken down, and the end products of their catabolic pathways enter the TCA cycle to generate energy. It has been shown that these amino acids promote their own degradation, but only during seed germination, senescence, or under sugar starvation [165]. This suggests that the degradation pathways provide alternative carbon sources for the plant in extreme conditions. In addition, branched-chain amino acids and their derived alpha-keto acids are cytotoxic and prevent accumulation through degradation which may be an important detoxification mechanism [166]. Higher levels of both fumarate and malate, as a result of the degradation of a surplus of amino acids, might thus be indicative for larger seed sizes. A second QTL for seed size on chromosome 5 co-locates with a QTL of opposite effect for GABA accumulation. Interestingly, Bay-0 alleles at both QTLs

confer larger seed size, suggesting directed evolution (directed evolution is a non-natural selection inadvertently introduced by the researchers), as was also observed in a different population [167]. However, where levels of fumarate and malate are increased in larger seeds, the accumulation of GABA is decreased. GABA is known to be involved in a range of cellular processes [168] and is rapidly accumulated in response to biotic and abiotic stresses [151]. It has been postulated that it has roles in herbivore deterrence, pH and redox regulation, energy production and maintenance of carbon/nitrogen (C/N) balance [152]. In a recent study, GABA levels in seeds were shown to increase by expressing glutamate decarboxylase (GAD) under a seed maturation-specific phaseolin promoter [169]. In accordance with our findings this resulted in smaller seed size and reduced seed vigor in T3 plants. No opposite seed size effect could be detected at a GABA QTL with increased levels due to the Bay-0 allele at the top of chromosome 4, but co-locating genetic variation for germination on ABA, heat sensitivity and dormancy was observed at this position. These cases illustrate the power of joined genetic analyses of metabolic and physiological traits for the generation of hypotheses that can help in the functional annotation of plant metabolites and their possible role in the regulation of important physiological processes.

Confirmation of metabolic QTLs

To independently confirm the effect of a single locus, it must be isolated and tested in an isogenic background. Several methods can be followed to perform such an independent confirmation of QTLs. A powerful approach is the use of residual heterozygosity in early generations of RILs. The Bay-0 × Sha RIL population (420 lines in total) was genotyped at F6, in which approximately 97% homozygosity is reached in each line. This resulted in the presence of residual heterozygosity in at least a single RIL at almost all genome positions. Those heterozygous regions are segregating in a Mendelian fashion in the next generation and can be used to confirm QTL positions, as it provides a possibility to study both parental alleles at the locus of interest in an otherwise homozygous background [170]. In a heterogeneous inbred family (HIF), those heterozygous regions are fixed, and two separate lines containing the alleles of both parents, respectively, are maintained.

HIF312 and HIF214 are segregating for regions at the top of chromosomes 4 and 5 (Fig. 3.19A), respectively, and cover the region in which the two major metabolite hotspots were detected. AR dry seeds were used to profile the HIFs for metabolic content because many of the QTLs detected in this region showed a large-effect size at the dry seed stages. Significant differences between parental alleles using four replicates were defined by a two-tailed Student's *t*-test (`P < 0.05`). In total, 34 out of 64 QTLs could be confirmed using this approach. For maltose, for instance, two QTLs with opposite direction were found (Fig. 3.19B), which could both be confirmed using the two distinct HIFs (Fig. 3.19C). In a number of cases, a HIF effect was observed that was not detected significantly in the

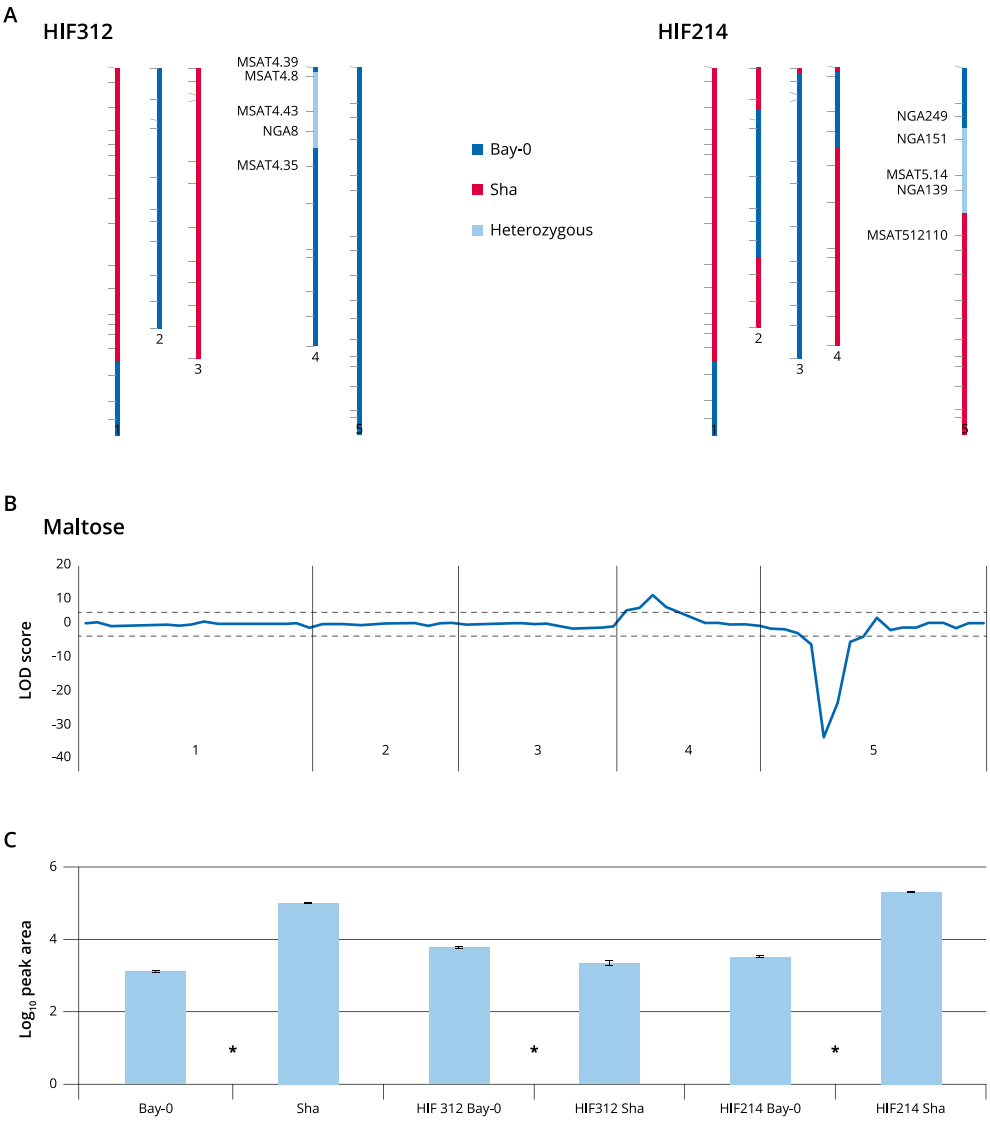


Figure 3.19 *– QTL confirmation for maltose using the heterogeneous inbred family (HIF) approach. Two QTL regions (top chromosome 4 and top chromosome 5) were analyzed using after-ripened (AR) seeds of lines HIF312 and HIF214 (A). The QTL profile for maltose (B) shows two significant QTLs (dashed line indicates the LOD 4 significance threshold). The lower panel (C) shows the parental levels for maltose and the confirmation for both QTLs by the segregating HIF lines (either fixed for Bay-0 or Sha alleles at the heterozygous interval). Significant differences (t-test* `P < 0.05`*) are indicated with * in-between the two contrasting samples.*

RIL population (e.g. digalactosylglycerol). This might be the result from the higher power in near isogenic lines due to the absence of epistatic interactions [171]. Nonetheless, a substantial number of QTLs could not be confirmed by the HIF lines. The enrichment for small-effect QTLs in the unconfirmed class suggests that four replicates generate insufficient power to identify significant differences for these metabolites in the HIF experiments, although we cannot rule out that they are false positives from the QTL analysis. Furthermore, QTLs depending on epistatic interactions cannot be detected in some near-isogenic lines. In addition, a number of QTL support intervals are broader than the region covered by the HIF, and thus, the causal genetic polymorphism within the QTL interval, but outside the region covered by the HIF, would have been missed.

The analyses of the HIF lines indicate that most of the large-effect QTLs can be accurately detected using a generalized genomics approach. Although an underestimation of small-effect QTLs can be expected, this is largely compensated by the higher power to detect G×E interactions.

4 High-throughput QTL mapping using correlated traits

This chapter describes a new methodology to be used in quantitative genetics called Correlated Traits Locus (CTL) mapping, a method complementary to QTL mapping. Where QTL associates differences in mean, CTL associates differences in correlation to genetic variation, i.e. CTL mapping identifies regions in the genome for which one genotype leads to correlated expression between a pair of traits, while the other genotype shows no (or significantly different) correlation.

Originally published as:

Danny Arends, Pjotr Prins, Yang Li, Lude Franke and Ritsert C. Jansen
CTL mapping
Draft

Danny Arends*, Harm-Jan Westra*, ..., Ritsert C. Jansen and Lude Franke
Cell-type specific eQTL analysis without the need to sort cells
Submitted

4.1 CTL mapping

4.1.1 Introduction

QTL mapping of a gene expression identifies regions in the genome at which different genotypes lead to differences in gene expression levels. In a similar fashion, abundance of thousands of proteins and metabolites can be measured to map protein QTL (pQTL) and metabolite QTL (mQTL). Deep sequencing, chromatin, and methylated DNA immunoprecipitation are just a few of the latest technologies that add to the arsenal of tools available for the study of the genetic variation underlying quantitative phenotypes. Together, eQTL, mQTL, and pQTL are referred to as xQTL [70]. Different xQTL can localize to confirm each other, for example, within the *Arabidopsis thaliana* glucosinolate pathway [172]. Such inference can lead to dissecting pathways and gene networks, currently an active field of research [173].

Advances in QTL mapping focus on increasing QTL detection power and precision by using more and more advanced models to explain observed variance. A higher amount of explained variance will result in a more reliable causal inference [174]. Methods developed to improve this explained variance are: Bayesian interval mapping, a framework to add prior knowledge [75, 175], Multiple QTL model (MQM) mapping tries to improve power by fitting a genetic model using backward elimination of pre-selected genetic loci [16, 38] and machine learning approaches, such as Random Forest [176], support vector machines (SVM), neural networks, and more recently vQTL mapping [177] fall into this class of methods developed to improve power and attribute more variance to genetic factors.

Another class of methods used in QTL mapping are the multivariate mapping approaches. These approaches combine variance information from multiple traits to increase the number and significance of detected QTL. Methods like principal component analysis (PCA) and differential expression (DE) analysis fall into the class of multivariate mapping methods. A review of many of these methods can be found in Gilbert and le Roy [178].

Recently the field of analyzing differential networks has gained more and more attention. In these approaches the differences between genetic networks are studied [179, 180]. Several approaches have been used to identify differential correlations between experimental conditions for large-scale omics data sets using topological overlap [181]. However current approaches for detecting differential correlations only focus on the detection of differences in correlations between two or more experimental conditions [180, 181, 182]. We however believe that a major source of complementary information is available.

Here, we present Correlated Traits Locus (CTL) mapping, a method complementary to QTL mapping. Where QTL associates differences in mean, CTL associates differences in correlation to genetic variation, i.e. CTL mapping identifies regions in the genome for which one genotype leads to correlated expression between a pair of traits, while the other genotype shows no (or significantly different) correlation. CTL information complements QTL information, and provides insights into the genetic regulation of correlated traits hidden in a traditional QTL mapping approach.

CTL mapping using the R/ctl package is performed in the same way as QTL mapping using the R/qtl [15, 16] package. This means data and results from R/qtl are directly usable in the R/ctl package. The results section shows a small code example, to show similarities between CTL mapping and QTL mapping using R/qtl.

4.1.2 Calculating a CTL

A CTL is calculated at marker M by analyzing all possible trait-trait combinations. The input for the calculation is the genotype of marker M and the traits measured on the individuals. Rather than taking the mean trait value for each of the genotypes, as is done with QTL mapping, correlation is calculated between T1 and T2 only for individuals with the AA allele. Then we do the same for the individuals with the BB allele. In pseudo code:

```
foreach T1 in Traits
   foreach T2 in Traits
      corAA = cor(T1|AA, T2|AA)
      corBB = cor(T1|BB, T2|BB)
```

In words: at marker M, split the individuals in two groups conditional on their genotypes. Now, for each pair of traits (T1 and T2) calculate the correlation between T1 and T2 using individuals with an AA allele (`corAA`) or with a BB allele (`corBB`). When `corAA` or `corBB` is high it implies that T1 and T2 are regulated together conditional on the genotype.

This calculation of differences in correlation using pairwise traits conditional to genotype is repeated for all markers.

For the sake of simplicity the example given here deals with a cross with only 2 alleles (AA and BB). CTL mapping however is not limited to only two alleles (see the discussion section).

Locating genetic markers acting on trait pairs

In itself, the difference in trait-trait correlations observed at specific markers is

interesting. The correlation suggests that the pair T1 and T2 are connected traits. For molecular data T1 and T2 could be acting in tandem, they could be in the same pathway, they could have the same regulator or they could be connected in some other way. Naturally, there is the possibility the correlation is there by chance, or by sequence polymorphisms [73]. However, when the traits T1 and T2 are highly correlated for both AA and BB, the genomic location at marker M may not be so meaningful in the context of genetics, but when correlation between expression levels varies significantly between genotype AA and BB (e.g. `corAA >> corBB`), some form of regulation at marker M in the genotype is implied.

Take the differential

Therefore, to calculate the CTL effect size, we add an extra step. CTL effect size is based on the difference between `corAA` and `corBB`, i.e., the CTL effect is the delta of the correlations for the two genotypes AA and BB:

$$CTL = corAA - corBB$$

Again, this CTL effect size is calculated for every genomic position. When at a certain genomic location traits T1 and T2 correlate highly in AA, but for BB do not (or significantly less) correlate there may be an effect of interest. This, in essence, is a CTL. The CTL suggests the genomic location is operating on both traits in tandem for genotype (AA), but not in the other genotype (BB). Therefore, a hidden factor X at the CTL location (M) may be involved controlling the correlation between the two traits. Finally, to correct for differences in sample size the full calculation for the CTL at marker M reads:

$$CTL = \frac{(Z(corAA) - Z(corBB))}{stderr}$$

For calculation of Z and stderr see the section 'CTL analysis on N-genotypes' in the discussion.

Assigning significance

When scaling the difference between two Z-values using the standard error we obtain a T-statistic which follows a normal distribution [183] allowing us to calculate an exact P-value.

This P-value still needs to be corrected for multiple testing by using a Bonferroni correction or a multi-trait permutation approach [46] to estimate the null distribution. When doing permutations in each round the link between genotype and trait is

broken, by redistributing at random genotypes to the individuals while not allowing for duplicates. After 10,000+ permutations each CTL score is transformed into a P-value.

Observing that a trait might show many other traits with a CTL at a marker we can also use an alternative approach: Don't assign significance to the individual trait-trait connections, but summarize the effect across all traits, then use Quantile-Based Permutation Thresholds [184]. Using this approach for CTL mapping will add power to detect sets of co-localizing CTLs, but will obscure the individual trait-trait connections.

When CTL scores observed in real data are higher than any CTL score obtained during permutation, a Generalized Pareto Distribution (GPD) is used to estimate the extreme tail of the null distribution [185], this allows likelihood estimates for the extreme scores observed to be estimated.

QTL and CTL analysis of two traits
To describe the possible relationships between QTL and CTL we generate possible scenarios as shown in Figure 4.1.

A: The two traits have a QTL and a CTL at the given marker (1 scenario). This example shows the extreme scenario when the traits are well correlated in one genotype and entirely uncorrelated in the other. The locus affects both the mean and the correlation. They could have a regulatory factor in common, the effect of which is detected in QTL and CTL analysis.

B: The two traits have a CTL at the given marker, but only one trait has a QTL at that marker (2 scenarios). The locus affects the correlation, and the mean of one trait only. They could have a regulator in common in which case the latter trait is downstream of the other trait. This can happen if the locus leads to functionally different transcripts at equal expression levels for the trait without QTL, and this effect is propagated to the trait with (therefore) a QTL.

C: The two traits have a CTL but no QTL at the given marker (1 scenario). This example shows the extreme scenario when the traits are well correlated in one genotype and entirely uncorrelated in the other. The locus affects the correlation only. They could have a regulatory factor in common, the effect of which is detected in CTL analysis only: e.g. if the locus leads to co-regulated transcription in one genotype only.

D: The two traits have no CTL at the given marker (4 scenarios). The locus does not change the correlation, it changes the mean if one or both traits have a QTL. The example shows the extreme scenario when both traits have a QTL but are entirely

uncorrelated: they could be downstream of the same or a different factor at the given region.

E: The two traits have a CTL at the given marker, but the trait values are at the noise level in one genotype for one or both traits (3 scenarios). This expression at noise level for one trait could result from e.g. hybridization failure for one genotype. One or both traits can have a possibly artificial QTL and the two traits have a possibly artificial CTL. The example shows the scenario when only one trait has a QTL.

4.1.3 Inference of a hierarchical relationship between traits

Using information of co-localized CTL and QTL
Observing a significant CTL between T1 and T2 means that there is a factor (e.g. genetic regulator) located underneath the CTL peak influencing the correlation between T1 and T2. This information of shared genetic control, therefore, indicates that T1 and T2 are involved in the same biological pathway [181].

The absence of QTL in trait T1 and the presence of QTL in trait T2 indicate that T2 is downstream of T1 in the hierarchical network. The reasoning is that the variation caused by a genetic factor in T2 (QTL) has not propagated to T1 [172]. The reversed situation can also happen: presence of QTL in trait T1 and the absence of QTL in trait T2 indicate that T2 is upstream of T1 in the hierarchical network.

The presence of QTL in both traits T1 and T2 further confirms that T1 and T2 are related in the same pathway, and it even allows us to go from inferring hierarchical relationship to causality using methods such as conditional correlation [174, 186].

It should be noted that the co-localized CTL information does not exclude the existence of intermediate factor(s) between T1 and T2 in the pathway although the relative strength of correlation provides us information on the distance among them in the network, e.g. higher correlation can indicate a shorter distance between T1 and T2 in the network.

Using information of co-localized CTLs
When the CTLxy of trait X and trait Y co-localizes with CTLxz of X and Z and CTLyz of Y and Z, these three traits are possibly involved in the same network. The relative effect sizes of these CTLs can be used to infer the order of X, Y and Z in the hierarchical network. For example, when CTLxy and CTLyz are larger than CTLxz, we can conclude that they follow the order of X-Y-Z in the network, i.e. Y is in between of X and Z.

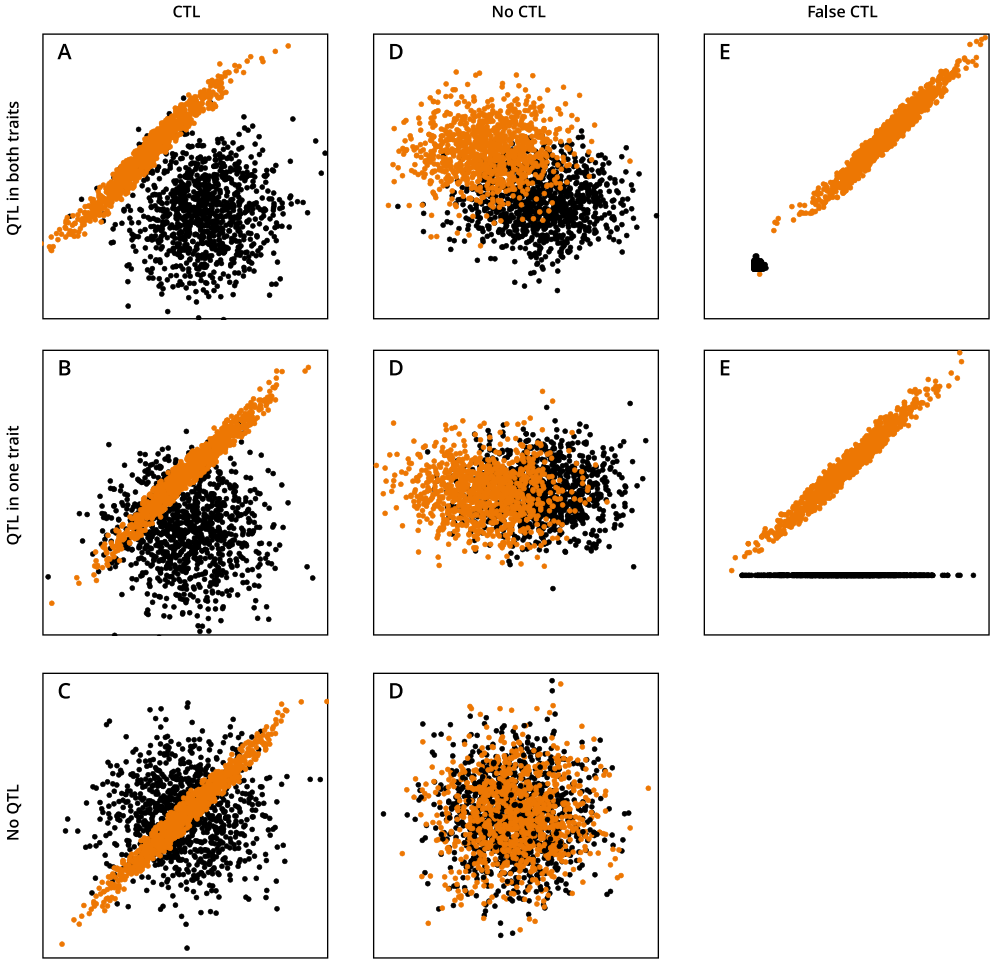


Figure 4.1 – *QTL and CTL analysis of two traits: a schematic of possible scenarios at a given marker. A trait can have a QTL or no QTL at the given marker, so there are four possible combinations in QTL analysis with two traits. The trait pair can also have a CTL or no CTL and in the combined QTL and CTL analysis the total number of possible combinations is eight. Three additional scenarios refer to special cases when one trait or both traits are not expressed in one genotype. For sake of simplicity it is assumed in this figure that there are two genotypes only, e.g. as in a BC or RIL population. (We also ignore possible info on cis- and trans-mapping for the moment.)*

Additionally when a QTL shows two or more co-localizing CTL we discovered a set of possible downstream targets for the QTL. Which, when sample sizes permit, can be further annotated using gene ontology [187] or untangled by using methods like hierarchical inference (this paper) or causal inference [86, 188] to obtain an even more detailed view of genetic regulation within the set.

Visualizing CTL information
Information obtained by CTL mapping can be visualized in several different ways. The author found the most accessible representation is a user-customizable network view. User customization helps researchers to add their own and/or literature information to a network allowing for a rich interpretation of the created network. Networks generated from CTL mapping can be visualized by software tools including Cytoscape [189, 190].

CTL information allows us to draw lines from traits via markers to other traits reconstructing the underlying genetic wiring. To create hierarchical networks we transform the significant CTLs into network edges. This is done by using a user defined genome wide FDR followed by transforming only significant trait-marker-trait interactions into a .sif network file.

CTL analysis on N-genotypes
CTL mapping on crosses with more than two genotypes (e.g. F2 or 4-way), or CTL mapping on GWAS data where SNPs are used for association mapping, involves a slightly different approach than outlined before. Here we detail our method while we assume not two but k genotypes. For each genotype i, a sample size of n_i individuals is available.

When k independent correlation coefficients are to be compared a Fisher's Z-transformation is applied to each of the individual correlation coefficients (R_i):

$$Z_i = 0.5[\ln(1 + R_i) - \ln(1 - R_i)]$$

To perform CTL mapping when three or more genotypes are present, one may evaluate the Chi-square test statistic [191, 192]. This has the following null hypothesis (H_0): all correlation coefficients are almost equal. The test statistic is then calculated summing over all k:

$$Chi^2 = \sum_{i=1}^{k} [(n_i - 3) \times Z_i^2] - \frac{[\sum_{i=1}^{k} (n_i - 3) \times Z_i]^2}{\sum_{i=1}^{k} (n_i - 3)}$$

Where $i = 1, 2, 3, .., k$.

The Chi-square test statistic has (k–1) degrees of freedom, where k is the number of genotypes. We transform the obtained P-value by using the $-\log_{10}(\text{P-value})$ and store the obtained LOD score for this trait-trait combination, then continue to the next genetic marker.

Slope and correlation
There exists a relationship between the slope of the regression line and the correlation:

$$cor = \beta \times \frac{SD_x}{SD_y}$$

From this formula we can see that there is no difference between using correlation (cor) or the regression coefficient (β) when the standard deviation in x (SD_x) is equal to the standard deviation in y (SD_y), in other words: (SDx/SDy=1). However, when there is a difference between the standard deviation ratio, correlation and slope have two different interpretations, meaning that observed slope differences should be interpreted as different a phenomenon than difference in correlation.

4.1.4 Power analysis

With the use of simulations the statistical power of our method can be determined when different effect sizes and/or genotype ratios are encountered by the algorithm. A summary of the power analysis results are shown in Fig. 4.2, and results are discussed in the caption of this figure. Code to simulate different effect sizes and genotype ratios is part of the CTL mapping R-package and can be used to estimate (beforehand) the required sample size and/or the best breeding strategy to obtain the most power for detecting CTL.

4.1.5 CTL power calculation applies to QTL and GWA

The power calculation that comes with the CTL method is a measure of the statistical power built into the experimental design of an experiment. This applies not only to CTL, but has interesting implications for QTL and GWA studies in general. I.e., the CTL power calculation is a measure of the statistical power of the experiment under analysis (QTL, CTL or GWA).

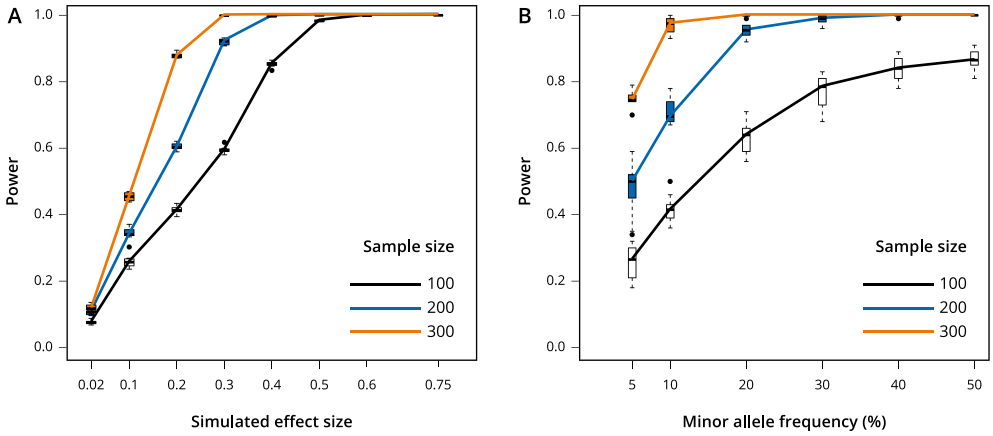


***Figure 4.2** - Power analysis (using simulations done at a single marker) shows how much sample size is needed to obtain a given power at* `alpha = 0.05`. ***A.** Increased effect sizes are detected with more power even at low sample sizes. To detect small differences in correlation 0.1 - 0.2 we need a large sample size (1000+). **B.** The closer the minor allele frequency is to 50%, the more power to detect CTLs. This plot was created using a simulated effect size of 0.3 for all genotype ratio's. The effect of minor allele frequencies shows that when ratios are close to 5%, we need a three times larger sample size to detect the same effect.*

4.2 Cell type specific eQTL mapping in human GWAS data

Expression quantitative trait locus (eQTL) mapping on tissue, organ or whole organism data can detect associations that are generic across cell types. The number of available expression quantitative trait locus (eQTL) data sets from individual cell types is limited because purifying cell types from mixtures is often challenging. Because whole peripheral blood is easily accessible and comprised of many different cell types (Fig. 4.3), we describe a new method to focus upon specific cell types without first needing to sort cells. We applied the method to whole blood data from 5,683 samples and demonstrate that SNPs associated with Crohn's disease preferentially affect gene expression within neutrophils.

4.2.1 Background

In the past seven years, genome-wide association studies (GWAS) have identified thousands of genetic variants that are associated with human disease [47]. The realisation that many of the disease-predisposing variants are non-coding and that single nucleotide polymorphisms (SNPs) often affect the expression of nearby genes

(i.e. *cis*-expression quantitative trait loci; *cis*-eQTLs) [48] suggests these variants have a predominantly regulatory function. Recent studies have shown that disease-predisposing variants in humans often exert their regulatory effect on gene expression in a cell-type dependent manner [193, 194, 195]. However, most human eQTL studies have used sample data obtained from mixtures of cell types (e.g. whole blood) or a few specific cell types (e.g. lympoblastoid cell lines) due to the prohibitive costs and labor required to purify subsets of cells from large samples (by cell sorting or laser capture micro-dissection). In addition, the method of cell isolation can trigger uncontrolled processes in the cell, which can cause biases. In consequence, it has been difficult to identify in which cell types these disease-associated variants exert their effect. Here we describe a generic approach that deepens our interpretation of GWAS data to the level of individual cell types (Fig. 4.4).

4.2.2 Method

Our strategy was to: (1) collect gene expression data from an entire tissue; (2) predict the abundance of its constituent cell types (i.e. the cell counts) by using expression levels of genes that serve as proxies for these cell types; (3) run an association analysis with a term for interaction between the SNP marker and the proxy for cell count to detect cell-type mediated or -specific associations, and (4) test whether known disease associations

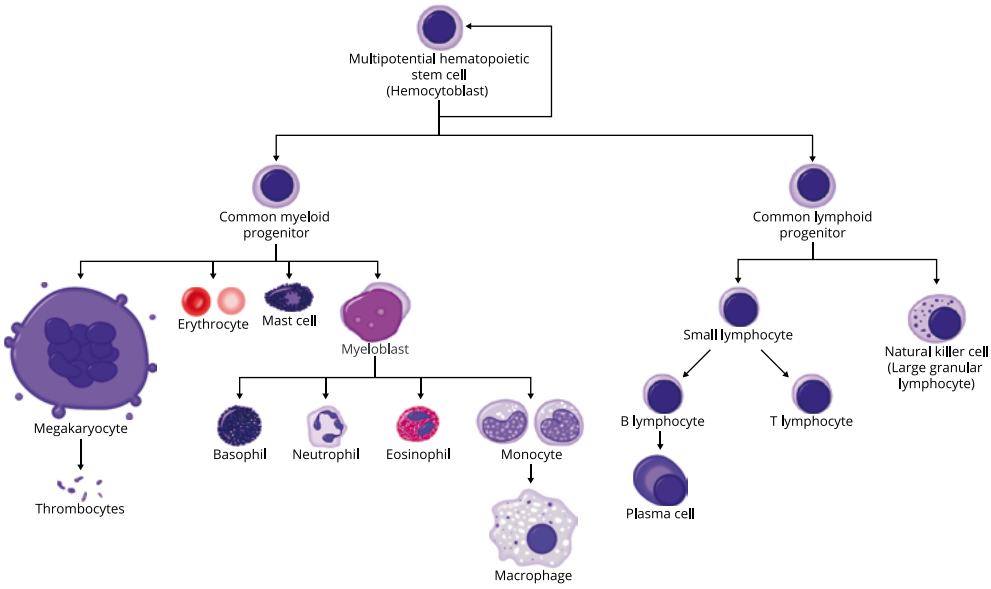


Figure 4.3 – *Overview of cells involved in hematopoiesis.*

are enriched for SNPs that show the cell-type-mediated or -specific effects on gene expression (i.e. eQTLs).

We applied this method to 5,863 unrelated, whole blood samples from seven cohorts: EGCUT [196], InCHIANTI [197], Rotterdam Study [198], Fehrmann [48], SHIP-TREND [199], KORA F4 [32, 50], and DILGOM [200]. Blood contains many different cell types that originate from either the myeloid (e.g. neutrophils and monocytes) or lymphoid lineage (e.g. B-cells and T-cells). Even though neutrophils comprise 62% of all white blood cells, no neutrophil eQTL data have been published to date, because this cell type is particularly difficult to purify or culture in the lab.

For the purpose of illustrating our new cell-type specific analysis in the seven whole blood cohorts, we focused on neutrophils. Direct neutrophil cell counts and percentages were only available in the EGCUT and SHIP-TREND cohorts, requiring us to infer neutrophil percentages for the other five cohorts (Fig. 4.4). We used the EGCUT cohort as a training data set to identify a list of 58 Illumina HT12v3 probes that correlated positively with neutrophil percentage (Spearman's correlation coefficient `R > 0.55`). We then summarized the gene expression levels of these 58 individual probes into a single neutrophil percentage estimate, by applying principal component analysis (PCA) and using the first principal component (confirmation of the accuracy of prediction in the SHIP-TREND cohort; Spearman `R = 0.81`). We then used this procedure in the other cohorts to predict the neutrophil percentage.

4.2.3 Results

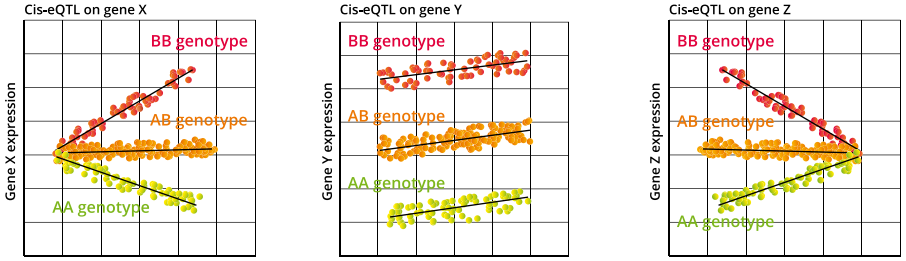
Identification of cell-type specific eQTLs in mixtures (e.g. whole peripheral blood) is most easily understood when assuming an extreme situation where in half of all blood samples the proportion of e.g. neutrophil granulocytes is very low, and in the other half of all blood samples this proportion is actually high. If an eQTL of interest is not showing any effect in the samples with a very low neutrophil granulocyte proportion, while showing a strong effect in the samples with a high proportion, this suggests the eQTL is specific for neutrophil granulocytes. In order to apply this reasoning on a whole blood data set, the neutrophil percentage of each sample should be known and is used as a covariate in a linear model that includes an interaction term (cell type percentage × genotype).

Here we limit our analysis to 13,124 previously discovered *cis*-eQTLs [201], (although a genome-wide application of our method might result in the detection of additional cell-type-specific *cis*-eQTLs that we might have missed by assuming a generic effect across cell types). We performed the eQTL association analysis with a term for interaction

1 In a **whole blood** dataset that has neutrophil percentage information (**EGCUT, n = 891**), correlate gene expression levels with neutrophil percentage and select Illumina probes with a high positive correlation (R2 > 0.3; n = 58).

2 In a whole blood dataset (without neutrophil percentage information), combine selected probes using principal component analysis (PCA). **The first principal component now acts as a proxy for neutrophil percentage.**

3 Map *cis*-eQTLs with a G x E interaction model, using the neutrophil proxy as an environmental factor.



4 We applied this method to seven independent whole blood studies and performed a meta-analysis (**n = 5,761**), testing **13,124 *cis*-eQTLs** identified in our previous study, controlling the FDR at 0.05.

Participating studies:

EGCUT SHIP-TREND Rotterdam Study InCHIANTI Fehrmann KORA-F4 DILGOM

Figure 4.4 *- Overview of the method to detect cell type mediated cis-eQTL in compound tissues.* ***A.*** *Starting with a data set that has cell count measurements, determine a set of probes strongly positively correlating to the cell count measurements. Calculate the correlation between these specific probes in the other data sets, and apply principal component analysis to combine them into a single proxy for the cell count measurement.* ***B.*** *Apply the neutrophil percentage predictor in the other cohorts.* ***C.*** *Use the proxy as a covariate in a linear model with interaction term in order to distinguish cell type mediated from non-cell type mediated cis-eQTL effects, and classify each cis-eQTL according to the direction of the interaction effect.*

between the SNP marker and the proxy for cell count within each cohort, followed by a meta-analysis (weighted for sample size) across all the cohorts. We identified 1,117 *cis*-eQTLs with a significant interaction effect (8.5% of all *cis*-eQTLs tested; false discovery rate (FDR) < 0.05; 1,037 unique SNPs and 836 unique probes). Out of the total number of *cis*-eQTLs tested, 909 (6.9%) had a positive direction of effect, which indicates that these *cis*-eQTLs show stronger effect sizes in neutrophils ('neutrophil-mediated *cis*-eQTLs'; 843 unique SNPs and 692 unique probes). Another 208 (1.6%) had a negative direction of effect (196 unique SNPs and 145 unique probes), indicating a stronger *cis*-eQTL effect size in lymphoid cells ('lymphocyte-mediated *cis*-eQTLs'; since lymphocyte percentages are strongly negatively correlated with neutrophil percentages). Overall, the directions of the significant interaction effects were consistent across the different cohorts, indicating that our findings are robust.

We validated the neutrophil- and lymphoid-mediated associations we detected in six small, purified cell-type gene expression data sets (Fig. 4.5) that had not been used in our meta-analysis. We generated new eQTL data from two lymphoid cell types (CD4+, $n = 309$ and CD8+ T-cells, $n = 309$) and one myeloid cell type (neutrophils, $n = 114$) and used previously generated eQTL data on two lymphoid cell types (lymphoblastoid cell lines, $n = 608$, and B-cells, $n = 283$) and another myeloid cell type (monocytes, $n = 282$). As expected, compared to *cis*-eQTLs without a significant interaction term ('generic *cis*-eQTLs', n = 12,007) the 909 neutrophil-mediated *cis*-eQTLs did indeed show very strong *cis*-eQTL effects in both of the myeloid data sets (Wilcoxon $P < 4.9 \times 10^{-31}$), and small effect sizes in the lymphoid data sets. Coversely, the 208 lymphoid-mediated *cis*-eQTLs had a pronounced effect in each of the lymphoid data sets (Wilcoxon $P < 7.8 \times 10^{-14}$; Fig. 4.5), while having small effect sizes in the myeloid data sets. These results indicate that our method is able to reliably predict whether a *cis*-eQTL is mediated by a specific cell type. Unfortunately, the cell type that mediates the *cis*-eQTL is not necessarily the one in which the *cis*-gene has the highest expression, making it impossible to identify cell-type-specific eQTLs directly on the basis of expression levels.

Myeloid and lymphoid blood cell types provide crucial immunological functions. Therefore, we assessed five immune-related diseases for which genomewide association studies previously identified at least 20 loci with a *cis*-eQTL in our meta-analysis. We observed a significant enrichment only for Crohn's disease (CD), (binomial test, one-tailed $P = 0.002$, Table 4.1): out of 49 unique CD-associated SNPs showing a *cis*-eQTL effect, 11 (22%) were neutrophilmediated. These 11 SNPs affect the expression of 14 unique genes (ordered by size of interaction effect: IL18RAP, CPEB4, RP11-514O12.4, RNASET2, NOD2, CISD1, LGALS9, AC034220.3, SLC22A4, HOTAIRM2, ZGPAT, LIME1, SLC2A4RG, and PLCL1). CD is a chronic inflammatory disease of the intestinal tract, and neutrophils are essential for killing microbes that translocate through the mucosal layer

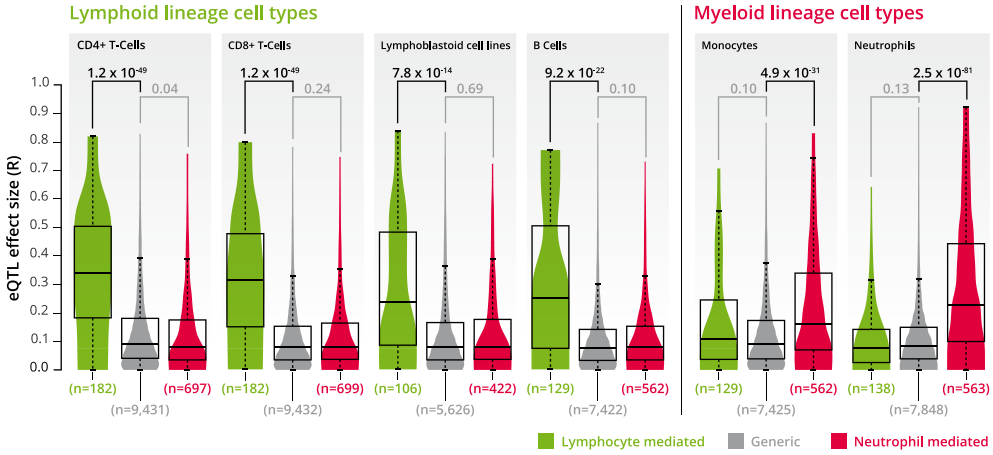


***Figure 4.5** – We validated the neutrophil and lymphoid mediated cis-eQTL effects in four purified cell type data sets from the lymphoid lineage (B-cells, CD4+ T-cells, CD8+ T-cells and lymphoblastoid cell lines) and two data sets from the myeloid lineage (monocytes and neutrophil granulocytes). Compared to generic cis-eQTLs, large effect sizes are observed for lymphoid mediated cis-eQTLs in lymphoid lineage cell types, while neutrophil mediated cis-eQTL effects have large effect sizes specifically in the neutrophil data sets. These results indicate that our method is able to reliably predict whether a specific cis-eQTL is mediated by cell type.*

of the intestine. The mucosal layer is affected in CD, but also in monogenic diseases with neutropenia and defects in phagocyte bacterial killing, such as chronic granulomatous disease, glycogen storage disease type I, and congenital neutropenia, leading to various CD phenotypes. In addition, pharmacological interventions for the treatment of CD have been developed to specifically target neutrophils, including Sagramostim and Natalizumab. Our new analysis shows clear neutrophil-mediated eQTL effects for many of the known CD genes, including the archetypal NOD2 gene, and our results provide deeper insight into the role of neutrophils in CD pathogenesis.

Large sample sizes are essential in order to find cell-type-mediated *cis*-eQTLs (Fig. 4.6): when we repeat our study on fewer samples by systematically excluding more cohorts from our study, the number of significant celltype-mediated eQTLs decreased rapidly. This was particularly important for the lymphoid-mediated *cis*-eQTLs, because myeloid cells are approximately twice as abundant as lymphoid cells in whole blood. Consequently, detecting lymphoid-mediated *cis*-eQTLs is more challenging than detecting myeloidmediated *cis*-eQTLs. As whole blood eQTL data is easily collected, we were able to gather a sufficient sample size in order to detect cell-type-mediated or -specific associations without requiring the actual purification of cell types.

Software availability
The source code and documentation for this type of analysis are available as part of the eQTL meta-analysis pipeline at: www.github.com/molgenis/systemsgenetics
Summary results are available from: www.genenetwork.nl/celltype

4.2.4 Discussion

Here we have shown that it is possible to infer in which cell types *cis*-eQTLs are operating, without the actual need to sort cells. We used whole peripheral blood eQTL data of 5,683 unrelated samples. By first estimating cell-type proportions and subsequent use of a G×E (i.e. the estimated cell-type proportions) interaction model we were able to demonstrate that hundreds of *cis*-eQTLs show stronger effects in myeloid than lymphoid cell types and vice versa.

Since we could subsequently replicate these results in 6 individual purified cell-type eQTL data sets (two reflecting the myeloid and four reflecting the lymphoid lineage), this indicates G×E analyses can provide important additional biological insights for many SNPs that have previously been found to be associated with complex (molecular) traits.

However, two main criteria apply in order to identify such G×E effects: first, sample size should be large. Although, to our knowledge, this is the largest eQTL study that has been conducted so far (5,683 samples), it is clear by sampling subsets of the participating cohorts (Fig. 4.6), that more G×E effects should be detectable with even larger sample sizes. Secondly, choosing the appropriate environmental factor is essential in order to find convincing G×E effects: although we found 1,117 *cis*-eQTLs that were mediated by myeloid or lymphoid cell types, a recent G×E study that aimed to detect *cis*-eQTLs that were mediated by either gender or age (assessed in 5,254 samples) found only

Disease	SNPs	AP	*cis*-eQTL SNPs	NM	Ratio	P-value
Crohn's disease	64	49	49	11	0.2895	0.0018
IBD	50	48	48	8	0.2000	0.0405
Rheumatoid arthritis	34	27	27	3	0.1250	0.3846
Multiple sclerosis	35	30	30	3	0.1111	0.4523
Type 1 diabetes	27	21	21	2	0.1053	0.5242

***Table 4.1** – Results of the different diseases tested for neutrophile mediated SNPs enrichment. AP = SNPs after pruning, NM = Number of neutrophil mediated SNPs, P-values reported are one tailed binomial pvalues. IDB = Inflammatory bowel disease.*

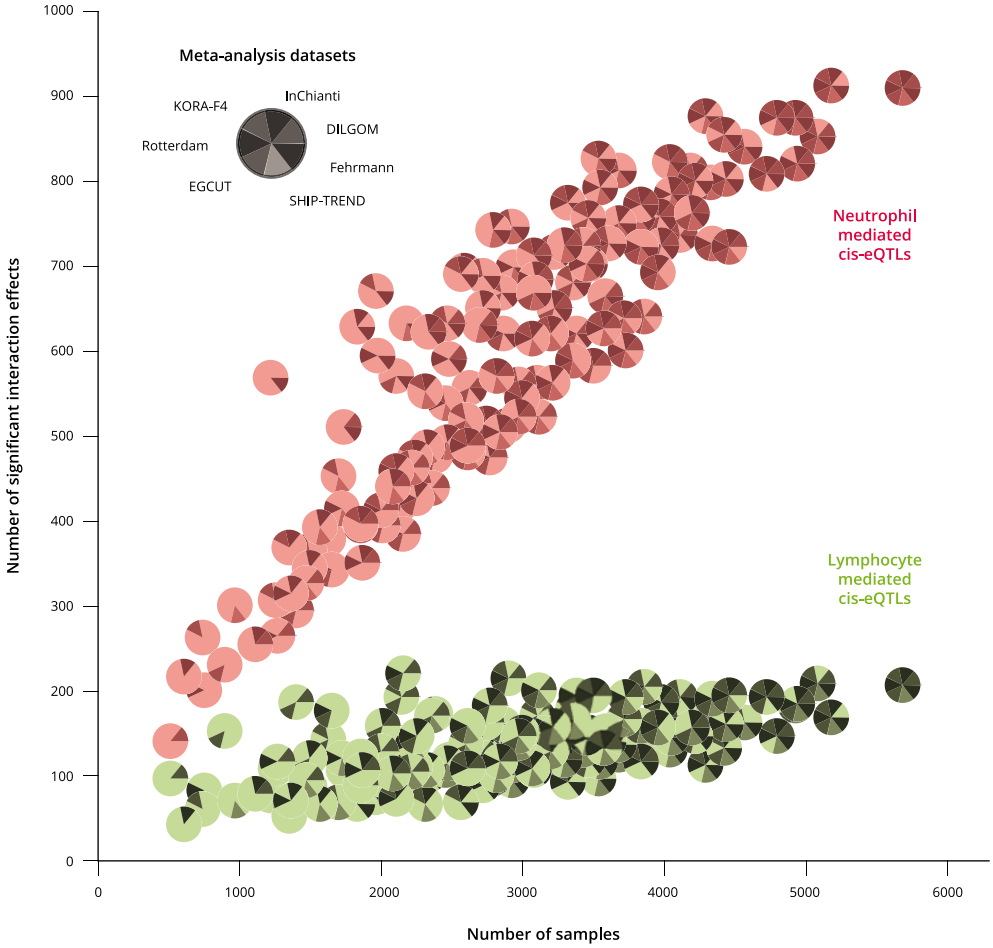


Figure 4.6 *– We systematically excluded data sets from our meta-analysis in order to determine the effect of sample size on the ability to detect significant interaction effects. The number of significant interaction effects dramatically decreases when the sample sizes become smaller. In general, lymphoid mediated cis-eQTL effects are harder to detect than neutrophil mediated cis-eQTL effects, due to their relatively small abundance in whole blood.*

five G×E effects that replicated in an independent cohort [202]. As such the choice of environmental factor is crucial in order to find G×E interaction effects.

Here, we have concentrated on identifying *cis*-eQTLs that are preferentially operating in either myeloid or lymphoid cell types. We did not attempt to assess this for specialized cell types within the myeloid or lymphoid lineage. However, this is well possible if cell-counts are available for these cell types, or if these cell-counts can be estimated by using the expression levels of genes that serve as proxies for those cell-counts. As such, identification of cell-type mediated eQTLs for previously unstudied cell types is possible, without the actual need to generate any new data. However, it should be noted that these individual cell types typically have a rather low abundance within whole blood (e.g. natural killer cells only comprise 2% of all circulating white blood cells). As a consequence, in order to have sufficient statistical power to identify eQTLs that are mediated by these cell types, very large whole blood eQTL sample-sizes are required (analogous to the substantially lower number of identified lymphoid mediated *cis*-eQTLs, as compared to the myeloid mediated *cis*-eQTLs, since neutrophils are twice as abundant as lymphoid cells in whole blood). Additionally, such cell types should show differences in abundance across different individuals, rendering the identification of *cis*-eQTLs that are specific for certain cell types impossible if those cell types have a near equal abundance in blood across each of the individuals.

In order to improve statistical power to detect cell-type mediated eQTLs, we corrected the gene expression for technical and batch effects (here we applied principal component analysis and removed per cohort the 40 strongest principal components that affect gene expression). Such procedures are commonly used when conducting *cis*-eQTL mapping [48, 195, 201, 203, 204, 205]. Although this correction might diminish the power to detect *trans*-eQTLs [201], this concern does not apply to our G×E *cis*-eQTL study.

We anticipate that with the (pending) availability of large RNA-Seq based eQTL data sets, statistical power to identify cell-type mediated eQTLs will improve further: since RNA-Seq enables very accurate gene expression level quantitation and is not limited to a set of preselected probes that interrogate well known genes (as is the case for microarrays), the detection of genes that can serve as reliable proxies for individual cell types will improve. Secondly, using RNA-Seq data, it is possible to assess whether SNPs that affect the expression of non-coding transcripts, affect splicing [205] or result in alternative polyadenylation [204], are mediated by specific cell types.

The method can be applied to eQTL data from any tissue, organ or whole organism, providing a computational alternative to sorting cells or performing laser capture micro-dissection. The only prerequisite for our method is the availability of a relatively

small training data set with cell count measurements in order to develop a reliable proxy for cell count measurements. Since the number of such training data sets is rapidly increasing and meta-analyses have proven successful [48, 201], our approach provides a cost-effective way to identify cell-type-mediated or -specific associations, and it is likely to reveal major biological insights.

5 High-throughput infrastructure for systems genetics

Modern high-throughput technologies generate large amounts of genomic, transcriptomic, proteomic and metabolomic data, creating a major challenge in bioinformatics from the size of data collected and the multitude of technologies used. We present the xQTL workbench system [70], a flexible and scalable platform to store and analyse all these new phenotype and genotype data using XGAP [206] dataformat and MOLGENIS software [207]. While we evaluated the system in multiple applications, only the WormQTL database [71] is presented here.

Originally published as:

Morris A. Swertz, Martijn Dijkstra, ..., **Danny Arends**, *et al.*
The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button
BMC Bioinformatics 1 Suppl 1(Suppl 12):S12 (2010)

Morris A. Swertz, K. Joeri van der Velde, ..., **Danny Arends**, *et al.*
XGAP: a uniform and extensible data model and software platform for genotype and phenotype experiments
Genome Biology 11(3):R27 (2010)

Danny Arends*, K. Joeri van der Velde*, Pjotr Prins, Karl W. Broman, *et al.*
xQTL workbench: a scalable web environment for multi-level QTL analysis
Bioinformatics 28(7):1042-4 (2012)

Danny Arends*, L. Basten Snoek*, K. Joeri van der Velde*, Yang Li*, *et al.*
WormQTL: Public archive and analysis web portal for natural variation data in Caenorhabditis spp
Nucleic Acids Research 41(DB issue):D738-43 (2012)

Modern genetic and genomic technologies provide researchers with unprecedented amounts of raw and processed data, and the need for software infrastructures to manage and process the large data sets produced is widely accepted [207, 208, 209]. For example, recent genetical genomics [207, 208, 209] studies have mapped gene expression (eQTL), protein abundance (pQTL) and metabolite abundance (mQTL) to genetic variation using genomewide linkage and genome-wide association experiments on various microarray, mass spectrometry and proton nuclear magnetic resonance (NMR) platforms and in a wide range of organisms, including human [197, 210, 211], yeast [42, 212], mouse [213], rat [214], *Caenorhabditis elegans* [174] and *Arabidopsis thaliana* [131, 143, 215].

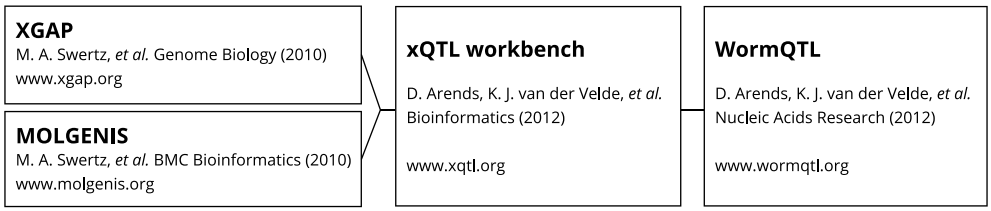


***Figure 5.1** – Overview of this chapter and how the different parts of this chapter are related to eachother.*

Understanding these and other high-tech genotype-to-phenotype data is challenging and depends on suitable 'cyber infrastructure' to integrate and analyze data [208, 216]: data infrastructures to store and query the data from different organisms, biomolecular profiling technologies, analysis protocols and experimental designs; graphical user interfaces (GUIs) to submit, trace and retrieve these particular data; communicating infrastructure in, for example, R [217], Java and web services to connect to different processing infrastructures for statistical analysis [84, 218] and/or integration of background information from public databases [219]; and a simple file format to load and exchange data within and between projects.

Many elements of the required cyber infrastructure are available: The Generic Model Organism Database (GMOD) community developed the Chado schema for sequence, expression and phenotype data [220, 221] and delivered reusable software components like gbrowse [222]; the BioConductor community has produced many analysis packages that include data structures for particular profiling technologies and experimental protocols [223].

Some integrated cyber infrastructures are also available: the National Center for Biotechnology Information (NCBI) has launched dbGaP (database of genotypes and

phenotypes) [224], a public database to archive genotype and clinical phenotype data from human studies; the Complex Trait Consortium has launched GeneNetwork [225], a database for mouse genotype, classical phenotype and gene expression phenotype data with tools for 'per-trait' quantitative trait loci (QTL) analysis.

However, a suitable and customizable integration of these elements to support high-throughput genotype-to-phenotype experiments is still needed [209]. dbGaP, GeneNetwork and the model organism databases are designed as international repositories and not to serve as general data infrastructure for individual projects. Many of the existing bespoke data models are too complicated and specialized, hard to integrate between profiling technologies, or lack software support to easily connect to new analysis tools. Customization of the existing infrastructures dbGaP, GeneNetwork, other international repositories [226, 227] or assembly of Bioconductor and generic model organism database components to suit particular experimental designs, organisms and biotechnologies still requires many minor and sometimes major manual changes in the software code that go beyond what individual lab bioinformaticians can or should do. This results in duplicated efforts between labs if attempted. Existing open source web-based tools for QTL analysis, such as webQTL [228] and QTLNetwork [229], are not easily extendable to different settings and computationally scalable for whole genome analyses.

Here we present xQTL workbench and its application in the WormQTL database. WormQTL (www.wormqtl.org) is an easily accessible database enabling search, comparative analysis and meta-analysis of all data on variation in *C. elegans*. All was built on top of xQTL workbench, a generic scalable web platform for the mapping of quantitative trait loci (QTLs) at multiple levels, for example: gene expression (eQTL), protein abundance (pQTL), metabolite abundance (mQTL) and phenotype (phQTL) data. xQTL workbench, MOLGENIS, XGAP and WormQTL are related to each other as outlined in Figure 5.1, and are described in detail below.

5.1 High-throughput data analysis (xQTL workbench)

Existing open source web-based tools for QTL analysis, such as webQTL [228, 231] and QTLNetwork [229], are not easily extendable to different settings and computationally scalable for whole genome analyses. xQTL workbench makes it easy to analyse large and complex data sets using state-of-the-art QTL mapping tools and to apply these methods to millions of phenotypes using parallelized 'Big Data' solutions [58]. xQTL workbench supports storing of data (raw, intermediate and final result), analysis protocols, history for reproducibility, and data provenance using the XGAP data model [206]. Use of

MOLGENIS [230] helps to customize the software. All is conveniently accessible via standard Internet browsers on Windows, Linux or Mac.

xQTL workbench is a scalable web platform for the mapping of QTLs at multiple levels: for example gene expression (eQTL), protein abundance (pQTL), metabolite abundance (mQTL) and phenotype (phQTL) data. Popular QTL mapping methods for model organisms and human populations are accessible via the web user interface. Large calculations scale easily on to multi-core computers, clusters and cloud. All data involved can be uploaded and queried online: markers, genotypes, microarrays, NGS, LC-MS, GC-MS, NMR, etc. When new data types come available, xQTL workbench is quickly customized using the MOLGENIS software generator.

5.1.1 Features

xQTL workbench provides visualization of QTL profiles, single and multiple QTL mapping methods, easy addition of new QTL analyses, scalable data management and analysis tracking.

1. **Explore QTL profiles -** Through the web interface, users can explore mapped QTLs, and underlying information by viewing QTL plots in an interactive scrollable and zoomable window. xQTL workbench has support for other common image formats, such as PNG, JPG, SVG and embedded postscript; useful for publishing scientific results online, and on paper. From the output, main database identifiers, such as gene, protein and/or metabolite identifiers are automatically linked-out to matching external web pages of public database such as NCBI, KEGG, and Wormbase.

2. **Single and multiple QTL mapping -** xQTL workbench wraps R/qtl [15, 16] in a web-based analysis framework offering all important QTL mapping routines, such as the EM algorithm, imputation, Haley-Knott regression, the extended Haley-Knott method, marker regression, and Multiple QTL mapping. In addition, xQTL workbench includes R/qtlbim, a library which provides a Bayesian model selection approach for mapping multiple interacting QTL [75], and Plink, a library for association QTL mapping on Single Nucleotide Polymorphisms (SNP) in natural populations [232].

3. **Add new analysis tools -** xQTL workbench supports flexible adding of more QTL analysis software: any R-based or command-line tool can be plugged in. All analysis results are uploaded, stored and tracked in the xQTL workbench database through an R-API. When new tools are added, they can build on the high-level multi-core computer, cluster and cloud management functions, based on TORQUE/OpenPBS and BioNode [173]. xQTL workbench can be made part of a larger analysis pipeline

using interfaces to R, Excel, REST and SOAP web services, and Galaxy [233].

4. **Track and trace -** When a new analysis protocol or R script is defined, this protocol can easily be applied to new data. Also, xQTL workbench keeps track of history. Re-use of analysis protocols can be done in an automated fashion. Previous analyses can be rerun without resetting parameters. xQTL workbench provides an online overview of past analyses e.g. which analyses were performed, by who, when, and the settings used.

5. **Scalable data management -** xQTL workbench has a consistency checking database based on XGAP specification [206], user interfaces to manage and query genotype and phenotype data sets, and support for various database back ends including HSQL (standalone) and MySQL. Phenotype, genotype and genetic map data can be imported as text (TXT), comma separated values (CSV), and Excel files. xQTL workbench handles and stores large data in a new and efficient binary edition of the XGAP format, named XGAPbin (extension .xbin), documented online. Such binary formats are essential when handling, storing and transporting multi-Gigabyte data sets.

6. **Customize to research needs -** Additional modules for new data modalities can be added using the MOLGENIS software generator [230]. The 'look and feel' of xQTL workbench is adaptable to the style of the institute or consortium by changing a simple template, which is described in the xQTL workbench documentation, enabling seamless integration into an existing website or intranet site such as recently done for EU-PANACEA model organism project and LifeLines biobank.

5.1.2 Reusable software (MOLGENIS)

xQTL workbench was implemented using MOLGENIS. MOLGENIS was born from the observation that bioinformaticians are under continuous pressure to both tackle the complexity and diversity of new biological systems and analytical methods and to translate these quickly into flexible informatics infrastructures, while keeping up with the unpredictable evolution of molecular biotechnologies and the increasing scale of experiments.

While standardization of tools and data formats in open source projects like the Generic Model Organism Database (GMOD) [221], and the Open Bioinformatics Foundation (OBF) [234], have been indispensable in reducing the development efforts needed via reusable and easy to integrate components, new research must also be quickly accommodated, for which efficient software variation mechanisms are needed.

Figure 5.2 outlines the 'model-driven' development method that several bioinformatics projects adopted in recent years to enable fast and flexible infrastructure development. We demonstrate step-by-step how bioinformaticians can use a domain-specific language to efficiently model the biological details of their particular biological system, and use MOLGENIS software generation tools to automatically generate a web application tailored to the experiments of their biologists, building on reusable components.

Next, we evaluate the results of these methods in the development of a range of MOLGENIS applications [68, 206, 207, 209, 219, 235], that is, software applications generated using the MOLGENIS toolkit. We found up to 30 times efficiency improvement compared to handwriting software, while providing a richness of features practically unfeasible to produce by hand but not yet provided by related projects. We conclude by inviting the bioinformatics community to add more MOLGENIS models, components and generators to quickly generate all the software infrastructures biologists want to have.

The MOLGENIS toolkit is based on the method of model-driven development which emerged in the late 1980s from the computer industry. Below we discuss the MOLGENIS' modeling language, reusable components and generators.

Modeling language
A custom MOLGENIS application can be defined in a single file. The file is written in MOLGENIS' modeling language. One can think of MOLGENIS' modeling language as a 'domain-specific language' [236], in this case to compose biosoftware infrastructures. The level of abstraction is raised, so no lengthy, technical or redundant details on how each feature should be implemented in general programming languages have to be given.

In most cases, knowledge of the DSL is all that is needed to produce a custom MOLGENIS application variant. The domain-specific language was implemented using XML so that model files can be edited using off-the-shelf XML editors. However, you may want to include hand-programmed components into a particular MOLGENIS instance. For example, for the eXtensible Genotype And Phenotype (XGAP) database application of MOLGENIS [206], we developed a 'MatrixViewer' that builds on the generated components, which saved us the work of writing the plug-in from scratch. This requires a model sentence that points to the 'plug-in' (allowing it to be seamlessly integrated) as well as hand-programming of the plug-in itself.

Reusable components
Each MOLGENIS application follows the widely accepted three-layered architecture design of web applications. MOLGENIS' reusable components provide building

blocks with a modular structure, which allows them to be assembled in diverse combinations, similar to prefabricated houses that are built from modular walls instead of bricks. Some building blocks are semi-finished and need to be 'completed' before use (which is automated in MOLGENIS via the generators and inheritance). We based the design of MOLGENIS on industryproven design patterns from the 'patterns for enterprise application architecture' (PEAA), a catalog of proven solutions for software design problems that we used as a guideline [237]. The logic of the reusable components is implemented using Java (www.java.sun.com); the HTML layout for the user interface is encoded in Freemarker templates (www.freemarker.sourceforge.net); and the database back end using MySQL, PostgreSQL or HSQLDB.

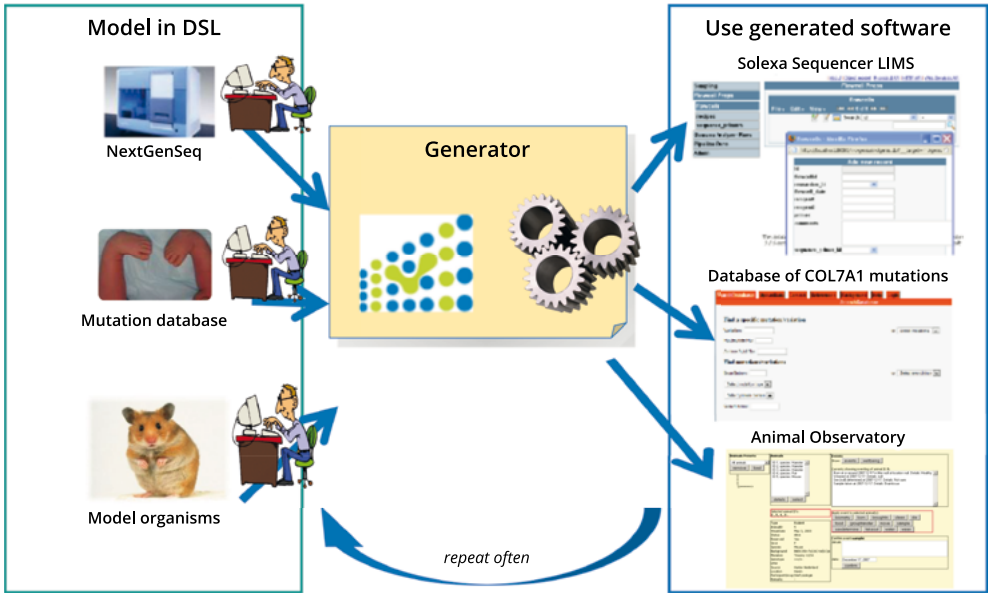


Figure 5.2 – *Many minor and major changes have to be written in software code before a 'standard' software infrastructure accommodates a particular research. Using 'modeldriven' development methods a bioinformatician only needs to model what is needed for his experiment using a therefore optimized domain-specific language (DSL). Generators quickly produce all the software logic to compose a full software infrastructure that accommodates these needs. When experimental needs change, a bioinformatician can (re)run the same generator with an adapted model file to quickly produce another variant of software infrastructure. This vastly reduces 'time-to-research' and enables bioinformaticians to quickly develop a suite of software infrastructures, with each variant accommodating a specific research task, while still on track to reuse, integrate and share the best standard features with other labs and bioinformaticians.*

Generators
The generators are compact specifications of how each database feature should be implemented. The MOLGENIS toolkit now has over 20 generators, but normal users will never need to take a look inside. However, for readers wanting to create their own generators, Figure 5.3 provides an example of the simple, textbased, generators we use. Each generator consists of two files: a Freemarker template that describes the code to be generated (similar to that shown in Figure 5.3A) and a Java 'Generator' class that controls the generating process. A new generator can be developed as follows: First write some examples of the desired programs by hand, where possible using similar patterns and mark which parts are variable between them. Then copy one of these examples into a generator template (text file) and replace all variable parts with 'holes' that are to be filled by the code generator based on parameters from DSL. At each generation, the template is then automatically copied and the 'holes' filled, based on parameters described in the domain-specific language, saving much laborious manual work.

Results
To start generating your own MOLGENIS application, you can download a ready-to-use 'workspace' from www.molgenis.org, which can be edited using the commonly used Eclipse integrated development environment (IDE) tool (www.eclipse.org). Extensive manuals are available to help install the Java, MySQL, Tomcat and Eclipse software needed and to learn how to walk through the Eclipse workspace to edit models and generate and run MOLGENIS instances; most new users can complete this part in about three hours. Detailed examples on how these features can be used to support actual microarray or genetical genomics experiments can be found in [68, 206, 219].

After completing a MOLGENIS model and running the generator as described above, you have a ready-to-use software application. The features you get when running the generated result as a web application: a fully functional system where researchers can upload, manage, browse and query their biological data that conform to the model, optionally enhanced with analysis tools to explore and annotate (depending on the plug-ins).

An important feature is human readable and printable documentation of your model, including a graphical overview showing relationships in UML which is of great use when still designing and discussing the model in a team. The next step is typically using the web user interface to populate and test your application with real data. To enable batch loading from a spreadsheet application such as Excel, the system comes with a tab-delimited import/export tool tailored for your data which you can use from the user interface as well as via a command-line tool; i.e., the headers of your Excel file have to match the fields you have defined in the model, In our experience, most computational

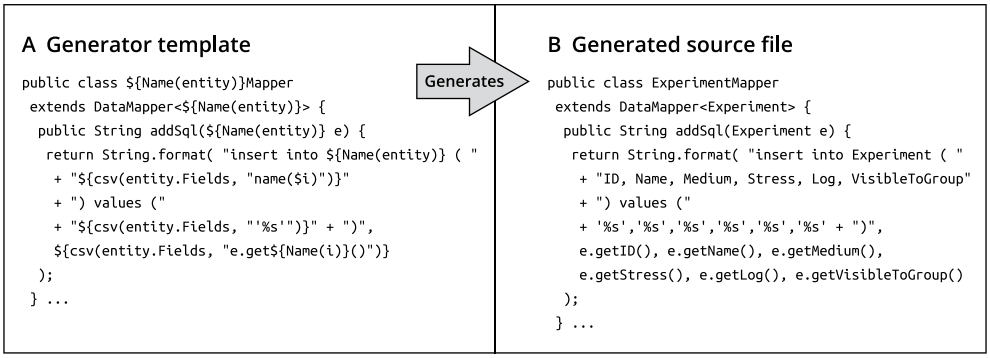


Figure 5.3 *- MOLGENIS generators are implemented as templates. This example shows the generator for a database component* (A)*. This template is applied to each <entity> in the model to generate many complete DataMappers that would otherwise need to be written by hand.* (B) *shows an example of the generated source files, in this case for <entity name="Experiment">. The command $Name(entity) translates to the name of the entity ("Experiment") and command $csv($entity.Fields, x) means that command 'x' is applied to each field of the entity and returned as a comma separated string (csv).*

biologists greatly appreciate the use of the R interface to load, analyze and re-store data from within the R statistical environment with web services to connect to workflow tools. Finally, advanced programmers may want to customize the layout or integrate their own scripts into the user interface, that is, create plug-ins that are seamlessly integrated with the generated software. Typical examples here are the integration of R scripts that produce graphical overviews of the data, enabling them to be run by non-technical research colleagues.

Applications of MOLGENIS

Since the earliest MOLGENIS application [207], we have successfully evaluated usage of the MOLGENIS toolkit to build a wide range of biomedical applications [68, 206, 219, 235, 238], ranging from sequencing to proteomics. A full list of MOLGENIS applications can be found at www.molgenis.org. Each of these MOLGENIS projects reported major benefits from the short cycle from model to running system to enable quick evaluation (500 lines of model XML replaces 15,000 lines of programming code) and use of the batch loading of data to evaluate how the newly built system works with real data. More often than not, MOLGENIS helped in finding inconsistencies in existing data that would otherwise have gone unnoticed, leading to experimental errors. In our experience, a typical MOLGENIS generator run gives you about 90% of the application that is desired 'for free', with the remaining 10% typically filled in using plug-ins that are written by hand. The MOLGENIS

toolkit has also been used to extend or replace existing software applications: the ExtractModel tool allows you to generate a MOLGENIS application from an existing database, which can then be run side-by-side with code developed previously, providing the best of both generated and handwritten worlds.

5.1.3 Extensible genotype and phenotype data (XGAP)

We present an extensible software model for the genotype and phenotype community, XGAP. Readers can download XGAP (www.xgap.org) or auto-generate a custom version using MOLGENIS with programming interfaces to R-software and web-services or user interfaces for biologists. XGAP has simple load formats for any type of genotype, epigenotype, transcript, protein, metabolite or other phenotype data. Current functionality includes tools ranging from eQTL analysis in mouse to genome-wide association studies in humans.

XGAP - A minimal and extensible object model

Another foundation of xQTL workbench was the development of an extensible data model for genotype and phenotype experiments (XGAP) that is designed as a platform to exchange data and tools and to be easily customized into variants to suit local experimental models.

Use of software domain-specific language and auto-generation, implemented using MOLGENIS, aims to ease and speed up customization/variation into new XGAP versions for new biotechnologies and alternative experimental designs while ensuring consistent programming interfaces for the integration and sharing of existing analysis tools. Standardized extension mechanisms should balance between format/interface stability for existing data types and tools, and flexibility to adopt new ones.

We developed the XGAP object model to uniformly capture the wide variety of (future) genotype and phenotype data, building on the generic standard model FuGE (Functional Genomics Experiment) [239] for describing the experimental 'metadata' on samples, protocols and experimental variables of functional genomics experiments, the OBO model (of the Open Biological and Biomedical Ontologies foundry) for use of standard and controlled vocabularies and ontologies that ease integration [240], and lessons learned from previous, profiling technology-specific modeling efforts [241].

Figure 5.4B shows the core components of a genotype-to-phenotype investigation: the biological subjects studied (for example, human individuals, mouse strains, plant tissue samples), the biomolecular protocols used (for example, Affymetrix, Illumina, Qiagen, liquid chromatography/mass spectrometry (LC/MS), Orbitrap, NMR), the trait data

generated (usually data matrices with, for example, phenotype or transcript abundance data), the additional information on these traits (for example, genome location of a transcript, masses of LC/MS peaks), the wet-lab or computational protocols used (for example, MetaNetwork [84] in the case of QTL and network analysis) and the derived data (for example, QTL likelihood curves).

We describe these biological components using FuGE data types and XGAP extensions thereof. Investigation holds all details of an investigation. Each investigation may apply a series of biomolecular [242] and computational [84, 218, 243, 244] Protocols. The applications of such Protocols are termed ProtocolApplications, which in the case of computational Protocols may require input Data and will deliver output Data. These data have the form of matrices, the DataElements of which have a row and a column index. Each row and column refers to a DimensionElement, being a particular Subject or a particular Trait.

Figure 5.4 (A,C) shows how the XGAP model can be extended to accommodate details on particular types of subjects and traits in a uniform way. A Trait can be a classical phenotype (for example, flowering - the flowering time is stored in the DataElement) or a biomolecular phenotype (for example, Gene X it's transcript abundance is stored in the DataElement). A Trait can also be a genotype (for example, Marker Y is a genomic feature observation that is stored in the DataElement).

Genomic traits such as Gene, Marker and Probe all need additional information about their genome Locus to be provided. Similarly, a Subject can be a single Sample (for example, a labeled biomaterial as put on a microarray) and such a sample may originate from one particular Individual. It may also be a PairedSample when biomaterials come from two individuals - for example, if biomaterial has been pooled as in two-color microarrays. An individual belongs to a particular Strain. When new experiments are added new variants of Trait and Subject can be added in a similar way.

Several standard data types were also inherited from FuGE to enable researchers to provide 'Minimum Information' for QTLs and Association Studies such as defined in the MIQAS checklist - a member of the Minimum Information for Biological and Biomedical Investigations (MIBBI) guideline effort [245].

Simple text-file format for data exchange

To enable data exchange using the XGAP model, we produced a simple textfile format (XGAP-TAB) based on the experience that for data formats to be used, data files should be easily created using simple Excel and text editor tools and closely resemble existing practices. This format is automatically derived from the model by requiring that all

annotations on Investigations, Protocols, Traits, Subjects, and extensions thereof, are described as delimited text files (one file per data type) with columns matching the properties described in the object model and each row describing one data instance. Optionally, sets of DataElements can also be formatted as separate text matrices with row and column names matching these in the Trait and Subject annotation files, and with each matrix value matching one DataElement. The dimensions of each data matrix are then listed by a row in the annotations on Data.

Application programming interfaces

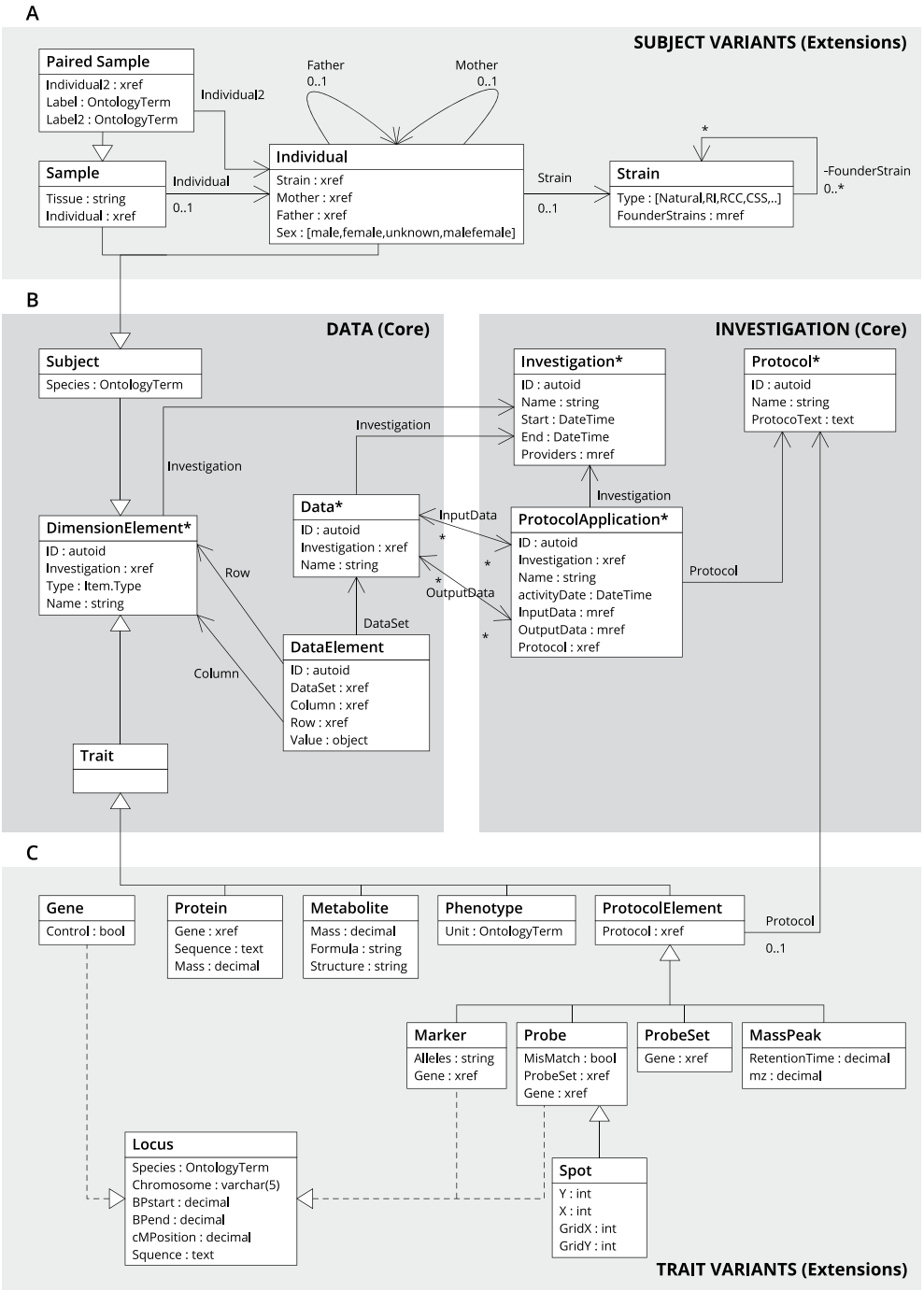
De facto standard analysis tools are emerging, for example, tools for transcript data [20, 21, 24] or metabolite abundance data [22] to mention just a few. These tools are typically implemented using the open source software for statistical analysis and graphics named R [217].

Bioinformaticians can connect their particular R or Java programs to the XGAP database using an API with similar functionality to the GUI, that is, using simple commands like 'find', 'add' and 'update' (R/API, Java/API). Scripts in other programming languages and workflow tools like Taverna [246] can use web services (SOAP/API) or a simple hyperlink-based interface (HTTP/API). On top of this, conversion tools have been added to the R interface to read and write XGAP data to the widely used R/qtl package [15, 16].

Based on these experiences, we expect use of XGAP to help the community of genome-to-phenome researchers to share data and tools, notwithstanding large variations in their research aims.

The XGAP data format can be used to represent and exchange all raw, intermediate and result data associated with an investigation, and an XGAP database, for instance, can be used as a platform to share both data and computational protocols (for example, written in the R statistical language) associated with a research publication in an open

***Figure* 5.4 –** *Experimental genotype and (molecular) phenotype data can be described using Subject, Trait, Data and DataElement; the experimental procedures used can be described using Investigation, Protocol and ProtocolApplication (B). Specific attributes and relationships can be added by extending core data types, e.g. Sample and Gene (A,C). The model is described in UML: Arrows denote relationships (Data has a field Investigation that refers to Investigation ID); Triangled lines denote inheritance (Metabolite inherits all properties ID, Name, Type from Trait, next to mass, formula and structure); Triangled dotted lines denote use of interface (Spot 'implements' properties of Locus); relationships are shown both as arrows and as properties ('xref' for one-to-many, 'mref' for many-to-many relationships). Asterisk* marks FuGE derived types.*


A
SUBJECT VARIANTS (Extensions)
Paired Sample
Individual2 : xref
Label : OntologyTerm
Label2 : OntologyTerm
Individual2
Father
0..1
Mother
0..1
Sample
Tissue : string
Individual : xref
Individual
0..1
Individual
Strain : xref
Mother : xref
Father : xref
Sex : [male,female,unknown,malefemale]
Strain
0..1
Strain
Type : [Natural,RI,RCC,CSS,..]
FounderStrains : mref
*
-FounderStrain
0..*
B
DATA (Core)
INVESTIGATION (Core)
Subject
Species : OntologyTerm
Investigation*
ID : autoid
Name : string
Start : DateTime
End : DateTime
Providers : mref
Protocol*
ID : autoid
Name : string
ProtocoText : text
Investigation
Investigation
Investigation
DimensionElement*
ID : autoid
Investigation : xref
Type : Item.Type
Name : string
Data*
ID : autoid
Investigation : xref
Name : string
InputData
OutputData
ProtocolApplication*
ID : autoid
Investigation : xref
Name : string
activityDate : DateTime
InputData : mref
OutputData : mref
Protocol : xref
Protocol
Row
Column
DataSet
DataElement
ID : autoid
DataSet : xref
Column : xref
Row : xref
Value : object
Trait
C
Gene
Control : bool
Protein
Gene : xref
Sequence : text
Mass : decimal
Metabolite
Mass : decimal
Formula : string
Structure : string
Phenotype
Unit : OntologyTerm
ProtocolElement
Protocol : xref
Protocol
0..1
Marker
Alleles : string
Gene : xref
Probe
MisMatch : bool
ProbeSet : xref
Gene : xref
ProbeSet
Gene : xref
MassPeak
RetentionTime : decimal
mz : decimal
Locus
Species : OntologyTerm
Chromosome : varchar(5)
BPstart : decimal
BPend : decimal
cMPosition : decimal
Squence : text
Spot
Y : int
X : int
GridX : int
GridY : int
TRAIT VARIANTS (Extensions)

format. We envision a directory service to which XGAP users can publish metadata on their investigations either manually or automatically by configuring this option in the XGAP administration user interface. This directory service can then be used as an entry point for federated querying between the community of XGAPs to share data and tools. Groups that already have an infrastructure can assimilate XGAP to ease evolution of their existing software.

Next to their existing user tools, they can 'rewire' algorithms and visual tools to also use the MOLGENIS APIs as data back end. Thus, researchers still have the same features as before, plus the features provided by the generated infrastructure (for example, data management GUIs, R/API) and connected tools (for example, R packages developed elsewhere). Moreover, much less software code needs to be maintained by hand when replacing handwritten parts by MOLGENIS-generated parts, allowing software engineers to add new features for researchers much more rapidly. We invite the broader community to join our efforts at the public www.XGAP.org wiki, mailing list and source code versioning system to evolve and share the best XGAP customizations and GUI/API 'plug-in' enhancements, to support the growing range of profiling technologies, create data pipelines between repositories, and to push developments in the directions that will most benefit research.

5.2 A worm database (WormQTL)

Here, we present an application of xQTL workbench: WormQTL (www.wormqtl.org) an easily accessible database enabling search, comparative analysis and meta-analysis of all data on variation in *Caenorhabditis spp*.

Over the past 30 years, the metazoan *Caenorhabditis elegans* has become a premier animal model for determining the genetic basis of quantitative traits [247, 248]. The extensive knowledge of molecular, cellular and neural bases of complex phenotypes makes *C. elegans* an ideal system for the next endeavor: determining the role of natural genetic variation on system variation. These efforts have resulted in an accumulation of a valuable amount of phenotypic, high-throughput molecular and genotypic data across different developmental worm stages and environments in hundreds of strains [174, 249, 250, 251, 252, 253, 254, 255, 256]. In addition, a similar wealth has been produced on hundreds of different *C. elegans* wild isolates and other species [257]. For example, *Caenorhabditis briggsae* is an emerging model organism that allows evolutionary comparisons with *C. elegans* and quantitative genetic exploration of its own unique biological attributes [258].

This rapid increase in valuable data calls for an easily accessible database allowing for comparative analysis and meta-analysis within and across *Caenorhabditis* species [259]. To facilitate this, we designed a public database repository for the worm community, WormQTL (www.wormqtl.org). Driven by the PANACEA project of the systems biology program of the EU, its design was tuned to the needs of *C. elegans* researchers via an intensive series of interactive design and user evaluation sessions on a mission to integrate all available data within the project.

All the software was built as open source, reusing and building on existing open source components as much as possible. WormQTL is freely accessible without registration and is hosted on a large computational cluster enabling high-throughput analyses.

As a result, data that were scattered across different platforms and databases can now be stored, downloaded, analysed and visualized in an easy and comprehensive way in WormQTL. Moreover, users can upload and share more R scripts as 'plug-in' for the colleagues in the community to use directly and run those on a computer cluster using software modules from xQTL workbench [70, 71]; this requires login to prevent abuse.

5.2.1 Results

WormQTL is an online database platform for expression quantitative trait loci (eQTL) exploration to service the worm community and already provides many publicly available data sets [86, 174, 251, 254, 255, 256, 260, 261]. New data sets can be uploaded using the XGAP plain file data format. Suitable help pages are provided. Currently, 38 public data sets have been loaded, of which the bulk is xQTL data on 500 strains (introgression lines, recombinant inbred lines (RILs), recombinant inbred advanced intercross lines and natural isolates), 55,000 transcripts, 1,594 samples and 1,579 markers. With this combination of classical phenotypes, molecular profiles and genetics data sets, WormQTL contains all the 'genetical genomics' experiments published to our current knowledge (except for some tiling data). Using WormQTL, researchers can explore many xQTLs across the various studies in different conditions and ages and compare classical QTLs with xQTLs. The main interfaces are 'Find QTLs', 'Genome browser' and 'Browse data'.

1. **Find QTLs -** QTL is genomic regions associated with phenotypic variation and can be used to study the genetic architecture of traits and to detect potential phenotypic regulators. Recently, the number of QTLs and especially eQTL studies in *C. elegans* has increased greatly. These eQTL studies consist of large data sets that, beforeWormQTL, were very difficult to access and perform a combined meta-analysis. Therefore, we provide easy access to most of the eQTL studies published, by

search, browse and plot functions (Fig. 5.5). We support relatively simple questions like 'does my gene have an xQTL?' to more advanced ones like 'how do these genes fit into an xQTL network?'. All the matching genes, markers and traits found in the data sets are returned, including links to WormBase and literature. Furthermore, WormQTL is the first portal for any species that allows comparison of eQTLs over multiple experiments and environments, giving insight in the plastic nature of genetic regulation.

2. **Genome browser -** To find the genes that have a QTL on your favorite position, click 'Genome browser'. Here, you can select from all the different releases of the University of California, Santa Cruz genome releases. You can add tracks from the designated experiments of interest. Then navigate to your favorite location (tip: use open in new window) and collect significant probe identifiers from that region.

3. **Browse data -** Complete data sets and accompanying gene, sample and trait identifier lists can be browsed via the 'browse data' user interface. External identifiers anywhere in the data are automatically recognized and enhanced as linkouts to background information, such as links to Wormbase, NCBI, KEGG or Ensembl. All the annotation lists and data matrices can be browsed and searched in a tabular form and can be downloaded as plain text or Excel files. Readers can also download data sets or submit new data sets using the XGAP data format following examples described in the WormQTL help section. Also all data can be accessed programmatically from with R (as whole matrix or per row) or using REST web services, including filtering of the annotations (genes, probes, markers and phenotypes) and services to 'slice' individual lines out of the complete data sets to speed up download and (parallel) analyses. Alternatively, readers can request a login to upload data and new analysis scripts directly.

Implementation

Using the XGAP data model, MOLGENIS generators automatically translate these models into a database, standard user interfaces for data queries and updates, upload/download tools for tab-delimited data and scriptable interfaces for programmers to users from within R and via web services. This greatly speeds up the initial software development and also enables rapid extension when, for example, new data types arrive. On top of this foundation, we built the WormQTL specific user interactions such as the 'Find QTLs' and the 'Genome browser' using the MOLGENIS 'plug-in' mechanism; and visualizations and plots using the R interface. xQTL workbench is a scalable web platform for the mapping of QTLs at multiple levels: for example, gene expression (xQTL), protein abundance (pQTL), metabolite abundance (mQTL) and phenotype (phQTL) data. The xQTL workbench provided a set of previously developed user interfaces to run R/qtl

mapping methods directly from within the WormQTL user interface. The ability to add new analysis procedures in R, data management and data format conversions, all greatly speed up the generation of new xQTL profiles.

All the data sets were downloaded from their original sources and then formatted using the XGAP data format. XGAP is a simple text file format that uses a directory of tab-delimited files or one Excel file with multiple sheets to load lists of annotations and data matrices. The annotations list all the background information needed to run and interpret the analysis, including genome position information such as markers, genes, probes and strains. The data matrices describe all the raw, intermediate and result data, such as gene expression, genotypes and QTL P-values, with the row names and column names cross linking to the annotations. For example, gene expression is a matrix of 'gene' × 'sample'. Subsequently these data sets were loaded using the MOLGENIS/xQTL data import wizards, which check the files for correctness and give informative feedback if the data are not yet in a format that WormQTL can understand [206]. All the annotations are stored in tables in the database; the large data matrices are stored in a optimized binary format to speed up analyses and queries. This format is documented in the WormQTL manual to ease the submission of new data sets from the community. Finally, all the QTL profiles were recalculated according to the specification of the original, or slightly modified when needed, such as to include a previously missing wrongly labelled sample correction. In this process, we greatly benefitted from the integration with xQTL workbench, which enabled us to re-run all these analyses on the computer cluster and add new R analysis procedures when needed, simply from the user interface.

5.2.2 Extensions into WormQTL-HD

Since its first publication, WormQTL has been expanded into a 'human disease' version. Interactions between proteins are highly conserved across species. As a result, the molecular basis of multiple diseases affecting humans can be studied in model organisms that offer many alternative experimental opportunities. One such organism - *Caenorhabditis elegans* - has been used to produce much molecular quantitative genetics and systems biology data over the past decade. We present WormQTL-HD (Human Disease), a database that quantitatively and systematically links expression Quantitative Trait Loci (eQTL) findings in *C. elegans* to gene-disease associations in man. WormQTL-HD, available online at www.wormqtl-hd.org, is a user-friendly set of tools to reveal functionally coherent, evolutionary conserved gene networks. These can be used to predict novel gene-to-gene associations and the functions of genes underlying the disease of interest. We created a new database that links *C. elegans* eQTL data sets to human diseases (34,337 gene-disease associations from OMIM, DGA, GWAS Central and NHGRI GWAS Catalogue) based on overlapping sets of orthologous genes associated to

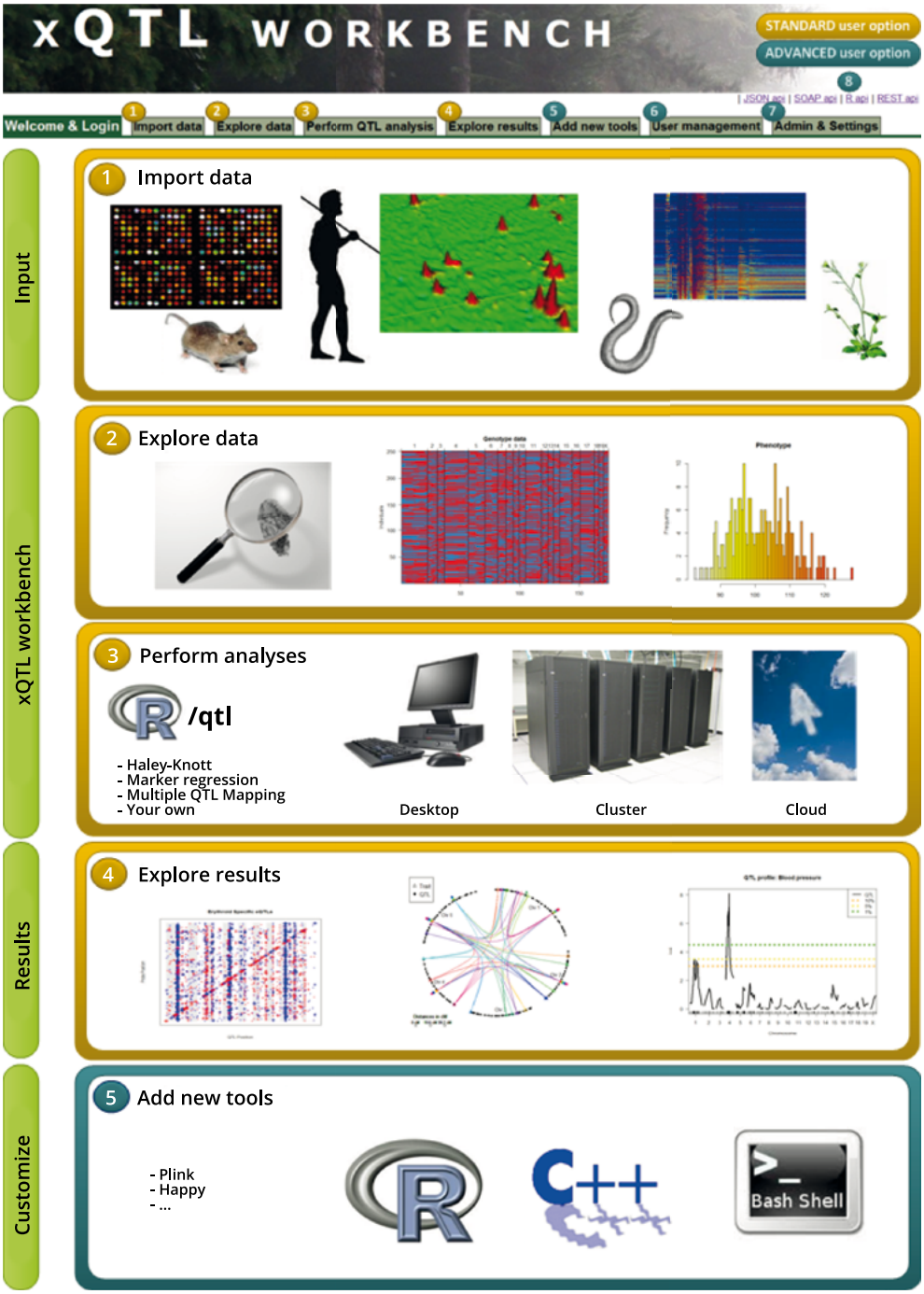
phenotypes in these two species. We utilized QTL results, high-throughput molecular phenotypes, classical phenotypes and genotype data covering different developmental stages and environments from WormQTL database. All software is available as open source, built on MOLGENIS and xQTL workbench.

5.2.3 Conclusion

xQTL workbench provides a total solution for web-based analysis: major QTL mapping routines are integrated for use by experienced and inexperienced users. Researchers can upload raw data, run analyses, explore mapped QTL and underlying information, and link-out to important databases. New algorithms can be flexibly added, immediately available to all users. Large analyses can be easily executed on a cluster or in the cloud. Future work includes visualizations and search options to explore the results. We also had an EU-SYSGENET workshop that envisioned further integration of xQTL with analysis tools like HAPPY, databases like GeneNetwork [225], and the workflow manager TIQS [262].

xQTL workbench was built on a minimal and extensible data infrastructure for the management and exchange of genotype-to-phenotype experiments, including an object model for genotype and phenotype data (XGAP-OM), a simple file format to exchange data using this model (XGAP-TAB) and easy-to-customize database software (XGAP-DB) that will help groups to directly use and adapt XGAP as a platform for their particular experimental data and analysis protocols. We successfully evaluated the XGAP model and software in a broad range of experiments: array data (gene expression, including tiling arrays for detection of alternative splicing, ChIP-on-chip for methylation, and genotyping arrays for SNP detection); proteomics and metabolomics data (liquid chromatography time of flight mass spectrometry (LC-QTOF MS), NMR); classical phenotype assays [86, 143, 211, 213, 263, 264, 265]; other assays for detection of genetic markers; and annotation information for panel, gene, sample and clone.

***Figure* 5.5** - *Screenshot of xQTL workbench with all features enabled,* **1.** *Import phenotype, genotype and genetic map data, examples are given per import type;* **2.** *Search through the whole database, explore and browse your data using MOLGENIS generated web interfaces;* **3.** *Run R/qtl QTL mapping, the general plug-in allows users to perform not only QTL mapping but also other analyses;* **4.** *Use default (or custom) plug-ins to explore results (e.g. Heatmaps, QTL profiles);* **5.** *Add new tools to the workbench (for Bioinformaticians);* **6.** *User management and access control of the system (Only for admins);* **7.** *Expert settings can be altered in the admin tab (Only for admins);* **8.** *Connect/share data using generated API's to R statistics, REST/JSON, SOAP.*


xQTL WORKBENCH
STANDARD user option
ADVANCED user option
JSON api | SOAP api | R api | REST api
Welcome & Login
1 Import data
2 Explore data
3 Perform QTL analysis
4 Explore results
5 Add new tools
6 User management
7 Admin & Settings
8
Input
xQTL workbench
Results
Customize
1 Import data
2 Explore data
3 Perform analyses
/qtl
- Haley-Knott
- Marker regression
- Multiple QTL Mapping
- Your own
Desktop
Cluster
Cloud
4 Explore results
5 Add new tools
- Plink
- Happy
- ...
C++
Bash Shell

Non-technical partners successfully evaluated the practical utility by independently formatting and loading parts of their consortium data: EUCASIMIR (for mouse), EU-GEN2PHEN (for human), EU-PANACEA (for *C. elegans*) and IOP-Brassica (for plants). A public subset of these data sets is available for download at www.xgap.org. When needed we could quickly add customizations to the model, building on the general schema, and then use MOLGENIS to generate a new version of the software at the push of a button, for example, to support NMR methods as an extended type of Trait [215]. Furthermore we successfully integrated processing tools, such as a two-way communication with R/QTL [15, 16] enabling QTL mapping on XGAP stored genotypes and phenotypes with QTL results stored back into XGAP.

In a recent perspective paper [259] we evaluated the general benefits and pitfalls of model-driven development, such as the ability to develop infrastructure in short cycles to get the application right, ensuring developers and biologists are thinking along the same lines, and increasing quality and functionality for all. We further evaluated applying this method to both microarray and genetical genomics experiments [207, 206].

Here we have presented MOLGENIS in detail for xQTL workbench and reported the results of using this method against a wider range of applications. We conclude that the use of model-driven methods enables bioinformaticians to build biological software infrastructures faster than before, with the additional benefit of much easier sharing of models, data and components.

Since its first creation, xQTL workbench has spawned a series of useful applications. We described the richest example to date, WormQTL, an easily accessible database enabling search, comparative analysis and meta-analysis of all data on variation in *Caenorhabditis spp*.

Recently, WormQTL-HD (Januari 2014) is a comprehensive and compendious database that enables molecular model organism data to be studied in the context of human diseases. Just as with WormQTL [71], we believe that WormQTL-HD will be continuously curated by the members of the *C. elegans* community.

We expect to further develop the xQTL workbench and WormQTL data and toolsets. There might be more ways in which researchers would like to search through the large amounts of data, for example, based on custom lists of gene identifiers, or by combining tools such as finding QTLs within specific regions. The QTL plots could be improved or replaced with interactive graphs that are more informative and would allow the users to continue 'drilling down' in the data instead of returning to the home page for a new analysis with a different tool. Furthermore, we envisage close integration with other

data sources and tools such as WormNet, R/qtl and GO Enrichment to provide even more biological context and analytical tools for the user.

We are committed to maintaining the data and software in the future and invite the community to add and share their new data and ideas. All software is available as open source on www.github.com/molgenis for others to reuse locally, and related technical documentation is available from: www.xqtl.org, www.rqtl.org and www.molgenis.org.

6 Conclusion, discussion and future perspectives

This chapter provides a short summary of the contents of the research findings, followed by a discussion about some of the relevant scientific issues that were addressed.

Parts of this chapter are adapted from:

Caroline Durrant, Morris A. Swertz, Rudi Alberts, **Danny Arends**, ..., Ritsert C. Jansen, Klaus Schughart, *et al.*
Bioinformatics tools and database resources for systems genetics analysis in mice – a short review and an evaluation of future needs
Briefings in Bioinformatics (2011)

6.1 General discussion

This thesis advocates that using high-throughput computational methodologies and more optimised algorithms such as Pheno2Geno or Multiple QTL mapping (MQM), combined with taking advantage of modern shared computational infrastructure, can provide a solution to the 'Big Data' challenges in systems genetics. Federated computing and collaborative research require tools such as the xQTL workbench, to work together and avoid duplicate efforts. This final chapter provides a short summary of the contents of the research findings, followed by a discussion of some of the relevant scientific issues that were addressed. Also, conclusions are drawn on the impact of the bioinformatics research and methodologies presented in this thesis, and on insights in systems genetics. The chapter concludes with an overview of future perspectives on research themes and methodologies in the interdisciplinary field of systems genetics and bioinformatics.

6.2 Highlight of the results

Chapter 2 details how Pheno2Geno was developed as an R-package for the high-throughput generation of genetic markers and maps from molecular phenotypes. Pheno2Geno selects suitable phenotypes that show clear differential expression in the founders. It uses mixture modeling to select phenotypes showing segregation ratios close to the expected Mendelian segregation ratios, and transforms them into genetic markers that are suitable for map construction and/or saturation. Pheno2Geno analyses the candidate genetic markers and excludes those showing multiple QTLs, epistatically interacting QTLs, and QTL by environment interactions to provide a set of robust markers for QTL mapping, protecting against genetic markers from a non-genetic origin.

Chapter 3 highlights the integration of the MQM algorithm into R/qtl. We validated the methodology with experimental data from a cross of *Arabidopsis thaliana* Bayreuth × Shahdara. Sections 3.3 and 3.4 show the performance of MQM on these experimental data. As the MQM algorithm is highly relevant to seed development, we discuss its use in this cross and try to find genetic factors underlying observed changes in the classical phenotypes and metabolome. We applied a generalised genetical genomics design [68] to optimise the statistical power to detect QTLs and QTL × environment interactions. This design approach allowed us to pick up more main effects and QTL × environment interactions as compared to a random design or a full-block design.

Chapter 4 describes our current work on the topic concerning correlation differences when mapping two traits onto the genome. Our observations of the patterns when performing CTL mapping on multiple phenotypes, revealed that some of these patterns

are similar to those when using a QTL × environment interaction model. We applied this interaction model to data from a human Genome Wide Association Study (GWAS). We detected cell type specific eQTLs in whole blood for neutrophils, and show that low effect QTLs observed in samples can be explained by compensating for the relative number of neutrophils observed or predicted. Cell type specific eQTL analysis will improve our power to detect QTLs and enables us to assign cell type labels to observed *cis*-eQTL effects.

Chapter 5 details our work to provide computational infrastructure for the Life Sciences. We advocate the use of generators to create software and propose a generic and extensible data model (XGAP) [206] to store phenotype and genotype data. Combining these two approaches, we developed several tools for 'Big Data' infrastructure. The combination of these tools is the xQTL workbench, a scalable web platform for the mapping of quantitative trait loci (QTLs) at multiple levels, such as gene expression (eQTL), protein abundance (pQTL), metabolite abundance (mQTL), and phenotype (phQTL) data. Popular QTL mapping methods for model organisms and human populations are accessible via the web user interface. The system also allows researchers to contribute new analytical routines. We conclude that good software infrastructure is of critical importance when scaling up large calculations to multi-core computers, clusters and cloud computing.

6.3 Pheno2Geno performance

Pheno2Geno is developed to optimise routines for generating genetics maps *de novo* or to saturate existing genetic maps. Pheno2Geno enables the user to generate biomarkers from microarrays, tilling arrays and RNA-Seq data using mixture models [62]. Important design requirements for Pheno2Geno: 1) as generic as possible; 2) suitable for multi-core machines; and 3) allowing clusters to generate genetic maps.

We developed Pheno2Geno to include a sensitive pre-selection of possible genetic markers with a simple *t*-test approach. This approach enables the user to quickly discard possible loci that do not show differential expression between parents. The method illustrates the trade-off between the risk of missing important markers found by other approaches, and the time consuming computational burden of analysing all loci. Pheno2Geno was developed in the context that it is better to have a map in reasonable time, than not having a map at all.

We applied mixture models to select good genetic markers that show segregation frequencies comparable to the expected segregation frequencies [60, 61]. We demonstrated that our approach was much more generically applicable to finding genetic markers than the method proposed by Michelmore *et al.* [60].

There are many tests to estimate a difference between two groups, including *t*-tests, linear models, mixture models etc. In a *t*-test only the mean values and variances are calculated. In our view, this provides the fastest way to decide if there is a difference. All other statistical methods mentioned above, come with added computational burdens such as ranking data or calculations that involve more statistics. The Pheno2Geno algorithm, however, is developed in such a way that it provides the user several other options for statistical testing (Wilcoxon Rank Testing, Mann-Whitney U Testing), and transformation of input signals (*Log*, *Sqrt*, *Reciprocal*, etc).

When creating maps for a recombinant inbred line (RIL) (or a Back-Cross) population, in case there is no parental information present, Pheno2Geno is still able to saturate a known low genetic map. Unlike other methods, Pheno2Geno does not require such parental information to detect the origins of the chromosomes. This is due to the fact that it flips anticorrelated chromosomes to find the optimal genetic map. A limitation of this procedure is that an initial genetic map has to be provided by the user.

The Pheno2Geno output structure is compatible with R/qtl [15, 16]. This facilitates an integrated experience between the map creation phase and the QTL mapping phase of the algorithm. Furthermore, when applicable, function parameters were harmonised between the two software packages. Compatibility of the outcome structures of Pheno2Geno and R/qtl also provide further high-throughput facilities for QTL mapping in inbred populations. With the upcoming new version of R/qtl (see section 6.6) the Pheno2Geno package will also require updating to accommodate new specifications. The expected changes include: a standalone package, further optimisation of parallelisation, performance improvements, and support for more crosses, e.g. diversity outbred and/or collaborative cross mice.

Taken together, the abovementioned features, i.e. pre-selection, generic software, optimised performance and R/qtl support, make Pheno2Geno a competitive package for genetic map creation in inbred populations. The Pheno2Geno algorithm also showed several drawbacks, which are discussed here.

As a consequence of the pre-selection of possible markers, using Pheno2Geno may induce that certain markers are missed. This is due to the fact that we discard phenotypes based on parental expression, and not apply mixture modeling to all phenotypes. If

mixture models would have been used to analyse all phenotypes, it is possible that more markers would have been detected.

Pheno2Geno currently does not support polyploid species such as corn, orchids and potatoes. For potato, fitTetra was recently developed [266]. To create genetic maps of this polyploid species, it uses a comparable strategy to Pheno2Geno. It can be expected that future releases of Pheno2Geno will include support for these species, as the expertise for polyploid map creation is present in our group and the topic is a subject of ongoing research.

Pheno2Geno is written in R. Despite the fact that this computer language is frequently used in the analysis of biological data, it is recognised that R is losing ground because of poor performance on extremely large data sets. We don't see this as a huge drawback because routines in Pheno2Geno are designed to minimise memory usage of the algorithm by providing 'lazy' equivalents of core functions, in such a way that data are only brought into RAM when necessary.

6.4 MQM revisited

The addition of MQM to the R/qtl toolbox has already shown to pay off in scientific terms: the paper that detailed the implementation has been cited more than 50 times since it was published in Bioinformatics in 2010 (Chapter 3.2). This indicates how much the research community values a comprehensive multiple QTL mapping methodology integrated into an already well-established QTL mapping framework (R/qtl). Several applications of MQM are known to our group. For example, dr. E. Lodder (Academic Medical Center, University of Amsterdam, The Netherlands) used MQM to investigate collagen deposits in mouse ventricular heart tissue (publication in press). She discovered two additional loci which are of interest for further investigation as potential new therapeutic targets for cardiac interventions. In this thesis we also show two examples concerning the classical phenotypes and metabolites in A. *thaliana*, in which we profiled MQM versus single marker mapping. Our analysis demonstrated that MQM detected more loci than the single marker mapping method, and the data showed more statistical confidence in the detected loci.

QTL mapping consists of two consecutive phases, modeling and mapping. In the modeling phase we look at the phenotype and assess the possible number of underlying QTL. We then use genetic loci to explain as much variation as possible. Subsequently, in the next phase of QTL mapping, the model is used and novel loci are included as assessed for their association with the phenotype of interest. We implemented multiple ways to

build initial models for mapping analysis, including: forward model selection, backward elimination of cofactors from a full model, and the application of the user's own custom cofactor model. How the different modeling techniques of MQM for systems genetics research are used precisely, and their pitfalls, are described in a detailed 40 page manual.

We have chosen to merge MQM with the R/qtl toolbox, as it has already proven itself in the mouse QTL mapping community. It is generally viewed as the standard for QTL mapping in all inbred species, and provides a unified data structure on which algorithms work. This allows quick incorporation of new algorithms and provides a common basis for data exchange between researchers.

MQM's core algorithm is written in C/C++ and is brought to R by creating a glue layer. Input to MQM is written in such a way that current R/qtl users can easily use the new MQM routine without the need to reformat the data. However, because the code is mostly written in C/C++ it is a trivial exercise to provide/interface MQM to different languages such as R, C/C++ and Ruby. This implies that the MQM routine is much more flexible to future changes in requirements or specifications, and/or language choice. It can also be extended to other tool boxes than R/qtl.

Earlier work has shown the following advantages of MQM compared to single marker QTL mapping [38, 77]:

- Higher power, as long as the QTL explain a reasonable amount of variation.
- Protection against over-fitting, because MQM fixes the residual variance from the full model, which allows the use of more cofactors than may be used in, for example, composite interval mapping (CIM) [37].
- Prevention of ghost QTL detection (between two QTL in coupling phase).
- Detection of negating QTL (QTL in repulsion phase).
- MQM gives (compared to CIM) a reduction in type I and II errors [77].

The research in this thesis has added several other advantages, most notable:

- A pragmatic permutation strategy for controlling the false discovery rate (FDR) and prevention of locating false QTL hot spots, as discussed in Breitling *et al.* [46]. Marker data are permuted, while keeping the correlation structure in the trait data. Because of the unified data structure in R/qtl, this permutation strategy can also be used

with all other QTL mapping functions in R/qtl, such as single marker mapping and composite interval mapping (CIM).

- High-performance computing by leveraging parallel computations using multi-CPU computers, as well as clustered computers, by calculating phenotypes in parallel through the Message Passing Interface (MPI) of the SNOW package for R [85].

- Visualisations for exploring interactions in a genomic circle plot (Fig. 3.2) and *cis*- and *trans*-regulation (Fig. 3.3).

The MQM routine for R/qtl also has drawbacks. It is computational much more intensive than single marker mapping. As it is designed in two languages, R and C/C++, the maintenance of code is therefore more difficult. Additionally, interpretation of the results is more complex when performing a multiple QTL scan. This is due to the fact that every phenotype is modelled in a unique way before scanning for QTL.

6.5 Visualising the output

As stated above, using MQM may pose problems to the user/biologist as the complexity of the model makes it much harder to interpret, as compared to the single marker QTL scan,which has a naïve model; one QTL for complex traits is by definition an oversimplification and (therefore) false. To alleviate the need for long and complex tabular output, we created circle plot visualisations (an example is shown in Fig 3.2). When scanning for interactions between multiple loci, it is common to compare everything versus everything. Since MQM provides us with a model of the main effect loci, we are able to quickly assess any interactions between them, because we only have to analyse a handful of possible interactions, i.e. between main effects. This allows us to quickly build up the circle plot with main QTL effects as well as interactions, which cannot be done for a single marker scan as no such model is present.

This type of visualisation provides researchers:

- Output per trait: an overview of interesting genetic loci (main effects) and the interactions between these main effect loci.

- A holistic overview of all traits to identify colocalization of main effect QTLs on the genome.

Furthermore, we added *cis*- and *trans*-visualisations which allow the same data to be explored in a 2D fashion providing comparable information to the holistic overview circle plot. Circle plots show the main effects and the interactions between genetic loci for a single phenotype, while *cis*-/*trans*-plots show main effects for many phenotypes, combined with genetic location information of the (expression) phenotypes mapped.

6.6 Limitations of R as an analysis platform for R/qtl

When dealing with high-throughput data obtained from RNA-Seq, exome sequencing or bisulphite sequencing, data size quickly becomes the limiting factor for software written in R [267, 268]. Because R/qtl-HD is aimed at QTL analysis of high-density, high-throughput data, R is not a suitable language for the high-throughput computational parts of the algorithms. R/qtl-HD is written in the high performance D programming language. D is chosen as primary language because it was developed to perform memory-safe and high-throughput numerical analyses. D uses a C like syntax [269], and therefore provides a familiar syntax to the R/qtl developers, most of whom have a C/C++ background, making D the language of choice for R/qtl-HD. Additional features of D include:

1. **Improved code quality -** This increases options for maintenance of the code. Also, a more readable code is more open to discussion and reasoning to optimise the code, which results in improved performance of the algorithms.

2. **C ABI compatible -** D directly allows to call C functions and a limited subset of C++. This allows to reuse previously written R/qtl code (written in R and C) [74].

3. **Built-in unit testing -** Unit testing is built into the D programming language, which facilitates programmers to use the power of unit- and regression testing to build a test suite without the need for external tools or frameworks.

4. **Concurrency and Actors -** The D language provides high level patterns such as *Actors* and *Message Passing* to deal with parallelisation of code. Unlike other languages, this is a built in feature of the language, which allows the standard D library to use safe lock free concurrency mechanisms [269].

5. **Static typing -** D requires types (e.g. *Integer*, *Float*, *Double*) to be declared and known at compile time. This improves readability of the code and reduces the risk for run-time type errors of the system.

6. **Compile Time Function Execution (CTFE) -** CTFE is a feature that allows to build look up tables at compile time. This greatly reduces the run time of algorithms, by pre-calculating common cases [270].

These are the main reasons for the R/qtl-HD development team to favour D over R. The current version of R/qtl provides many tools for genetic map construction, but also several historic methods for QTL mapping. R/qtl-HD will not provide a multitude of methods but focuses more on the high speed analysis of data using stable and proven methods such as Haley-Knott regression [36] and analysis of variance. Additional features such as genetic map construction and validation are not a high priority when converting the R/qtl software into a more optimised language. R/qtl-HD is designed for mapping huge volumes of phenotypes onto medium to large genetic maps [58].

Furthermore MQM for R/qtl will also be converted into the D programming language. The MQM routine has proven to be a useful addition to the R/qtl toolkit and with the new implementation of R/qtl as R/qtl-HD, we aim to improve its usefulness even more. The major limitation in MQM is the additional computational load when compared to single marker QTL mapping [16]. The R/qtl-HD development team is committed to: 1) add the MQM routine to the R/qtl-HD package, and 2) provide further optimisation to MQM allowing it to handle even larger volumes of phenotype and genetic data.

6.7 CTL & cell type specific eQTL

Presently, more and more research is being done to understand eQTLs and their effects on phenotypes. Recent papers reported the detection of some sex specific eQTLs [214, 271, 272], as they were present in females but not in males, or vice versa. Sex is obviously not the only covariate that can be used in these QTL × environment analyses; other possible covariates are population structure, age, and cell type. Our findings thus indicate the need to add cell type specific covariates in QTL mapping. One of the main advantages of using cell type as covariate is that we can compartmentalise environmental variation. This allows easier detection of associations between genetic loci and observed differences in phenotype, because some of the variation from non-genetic origin has been captured by the cell type covariate.

This method of mapping QTL × environment interactions is not new, although in our studies we applied it to cell types. However, this procedure was complicated by the fact that cell type abundance was only measured in two cohorts. The available cell counts were: lymphocyte, neutrophil, and RBC counts/percentages. As we used six cohorts in total, we had to infer the cell counts for the remaining four cohorts. Using the predicted

cell counts in these four cohorts, we were now able to perform our eQTL meta-analysis. This approach to handling cell type (for an eQTL) is similar to dealing with an environmental factor. When using a basic interaction model combined with cell type percentages, we can clearly distinguish between eQTLs which seem to be neutrophil specific and lymphocyte specific.

Large sample sizes are required to perform such interaction analyses with reasonable power to detect interaction effects. Currently, we can only obtain the required sample sizes by using meta-analysis of multiple eQTL GWA studies. The meta-analysis method increases our sample size dramatically as compared to a normal GWAS. This increase in sample size allows us to detect more eQTL, but also helps to reduce false positive eQTLs. Furthermore, we can use an additional level of information to assess our eQTLs. Looking at the direction of the eQTL effect in different cohorts, allows assessment of the consistency of the direction across multiple data sets.

Known genetic variants underlying for example Crohn's disease, IBD (inflammatory bowel disease) or asthma, are available from the literature [47]. In our study we demonstrated that using meta-analysis allows us to assign cell type labels to all eQTL detected. Combining these two sources of information allows us to test if the eQTLs showing association with a disease (such as Crohn's disease) are specific for a certain cell type. This information can be used in future research to find therapeutic targets for these diseases, or to understand why certain targets which might seem valid at first, turn out not to be effective.

Our study revealed that Crohn's disease indeed has a strong neutrophil component (`P < 0.0018`). While Crohn's disease showed a significant number of neutrophil specific eQTL, this was not the case for IBD. We concluded that derangements in different cell types are involved in IDB and Crohn's disease. This is surprising, because Crohn's disease is classified as a type of IBD [273]. Here we find supporting evidence that different pathophysiological processes are present in Crohn's disease, as compared to other IBD conditions such as Colitis ulcerosa.

The methodology for cell type specific eQTL also revealed some disadvantages. The selection of the covariate is very important. Here we used a continuous variable (cell abundance or cell percentage), in a comparable way to the mapping of sex specific eQTL, where the covariate is a dichotomous variable (male or female). As a consequence, in practice it is much harder to detect sex specific eQTL as compared to cell type-eQTL. We assume this is the reason why so few sex specific eQTLs are reported in the literature.

Another aspect is that large sample sizes are required for analysis. Methodological issues

concerning meta-analysis include the usage of multiple platforms (Illumina & Affymetrix) and even multiple versions from the same manufacturer (e.g. Illumina HT8, HT12). All this data needs to be integrated in a comprehensive and precise way. This creates a need for good infrastructure to keep track of the differences between data sets, and analytical tools which can handle missing or incomplete data [58, 173].

Several disadvantages arise in the context of experimental realities. Cell types are hard to purify and identify. In our opinion it is impossible to acquire a 100% pure fraction of, for example, CD4 T-Helper cells. Despite this, our method assumes that we can identify cell types without much bias. This might not be the case for some cell types which are present in many distinct sub cell types (such as T-cells), and consequently we cannot distinguish between them. Another example of a problematic situation refers to the changing expression patterns of harvested and purified neutrophils. These patterns may differ largely from the patterns when the neutrophils are still circulating in the blood.

6.8 xQTL workbench for infrastructural issues

The need for reliable and flexible infrastructure in systems genetics sparked our interest to develop the xQTL workbench. It combines the flexibility of the MOLGENIS generator suite [207] with the most frequently used computational and analytical tools for systems genetics (i.e. R/qtl [15, 16], Plink [232] and many other tools). Using R as Esperanto between the different languages creates a platform for systems geneticists to keep track of data generation and their analyses. Additionally, using R as computational back end allows researchers to contribute their methods and algorithms back to the scientific community by uploading them into the xQTL workbench. While the xQTL workbench system described in this thesis has the ability to absorb new analytical methods, we also considered the need for high flexibility. We expect it will be difficult to anticipate to new demands and requirements, or new developments that could offer better alternatives than databases for data and R for analysis [274]. Despite this, in our opinion the xQTL workbench is well suited to tackle such changing requirements by leveraging the strength of generators and the flexibility of the XGAP data model.

Infrastructure in systems genetics is created with generators, as it alleviates the need to handwrite most of the system's common codes [207]. This common code consists of basic database parts such as: *Create*, *Read*, *Update* and *Delete* (CRUD) functions for the underlying database system, generated web interfaces to enter and curate data, and application programming interfaces (APIs) [230] to transfer data from one language (e.g. Java) to another (e.g. R). We used the MOLGENIS generator together with the flexible XGAP data model [206] to set up a system capable of handling any genetical

genomics experiment 'at the push of a button' as the unofficial slogan states. All our collaborators reported gains on implementation time when using the xQTL workbench, as compared to a handwritten system in which all these components need to be (re) developed or invented. Parts of the system will always require fine tuning, but major parts of functionality do not change from one database system to the other, and thus allow automated generation of these parts.

The xQTL workbench allows R code to automagically use cluster and cloud computing for trivial parallelizable algorithms. An example of this is QTL mapping in which every phenotype is analysed independently. With a multi-core machine the distribution of computations to separate CPUs is easily achieved without much overhead (such as network latency). When using the xQTL workbench, researchers are able to write algorithms suited for a single phenotype/marker/individual, and test their analysis on their local desktop. When the algorithm is completed, the developer uploads the new algorithm into the xQTL workbench and applies it across rows/columns of the stored matrices. The user can then choose to use any available back end (either local host, cluster or cloud) to perform his/her computations. The xQTL workbench deals with distributing computations (if the analysis supports distributed computing) to multiple nodes when applicable.

An ever increasing amount of data and knowledge (in terms of the outcomes of computational analyses) are stored in large databases and are available to biologists for further study. This allows reuse of data by different biologists, decreases the need to regenerate data, and stimulates collaborations between research groups. Furthermore, translational databases (e.g. NCBI DB-GAP [224], KEGG [275]), which store and integrate data across many species, facilitate interactions between research communities such as Human and Worm genetics, by providing an opportunity to look across the borders of their own species [276].

At the same time, storage of raw data is still necessary in many cases. A good example is found in the next generation sequencing platforms. Although the costs for sequencing are being reduced in a high pace, additional funds are required to store and process the raw data coming from the machine. While at the one hand costs of sequencing are dropping, at the other hand sequencing produces more data as the technique is being performed at small hospitals and large sequencing facilities. It can be expected that when the price of sequencing becomes low enough, i.e. the costs of re-sequencing drop below the costs of storage, storage of raw data will probably be replaced by storage of samples (which are also stored now for re-analysis/validation). Such developments are likely to decrease the immediate need for storage space, but increase the computational burden of sequencing.

In this thesis we advocate R [274] as Esperanto language between storage and computational parts of our xQTL workbench platform. Our choice for R is motivated by several reasons including:

1. MOLGENIS provides a generated R API, that we used to plug in our analysis tools.

2. Many analytical tools are already available for R.

3. Biologists tend to be familiar with R, since the language is frequently used in data analysis. As a consequence, they can share their analysis software through the system.

4. R provides built in support for web browsing and secure connection (using Rcurl).

5. R provides an easy way to start/stop shell scripts.

Despite the realisation that other languages also fulfil (most of) these requirements, our familiarity and experience with R largely contributed to the decision to build the system with these components.

Application examples of the xQTL workbench system are the wormQTL [71] and the WormQTL-HD [276] databases. These databases contain a lot of worm data and provide the worm research community with tools for analysis. The WormQTL-HD database comes with the added advantage that it couples the worm phenotypes to human phenotypes. Consequently, geneticists working in the human domain can search for non-obvious (worm) models to test their hypothesis in worm. In other words, the database functions to bring these research communities closer together.

Several issues are relevant to consider in terms of disadvantages of the xQTL workbench. It is a complex system to set up for biologists, as its software is used as basis and requires other infrastructure on top. Therefore, programmers with knowledge of R, Java, and Unix are required for setting up the system and facilitate its maximal use.

Furthermore, the xQTL workbench system provides several QTL mapping routines that are available as proof of principle. Further adaptation of the system will increase the number of tools available in the workbench. While not a major disadvantage, it implies, for example, that MQM is not yet available to the users although it can easily be added and implemented into the system.

Using a generator requires regular updates, this introduces the risk to break the code in many (generated) places. A full and extensive test suite is then required to test the generated parts of the software for consistency and stability. Weighing the pros and cons, we consider this not really a drawback because a full test suit also raises confidence in the software, and makes it more reliable.

6.9 Applications in biology

The final research question of this thesis refers to the impact of our contributions and developments in the field of bioinformatics on advancements in methodology and knowledge in systems genetics. During the research project we regularly tested our tools and performed various experiments to validate our computational approaches.

In this thesis we show the application of Pheno2Geno on a large data set containing expression data from tilling arrays. Tilling arrays are a sensitive way to measure genome wide expression levels of RNA transcripts. They produce a large amount of data. Pheno2Geno uses these expression data to generate new high quality maps for QTL mapping, thereby exploiting the availability of QTL information in high-throughput phenotypes. We obtained raw expression data from our collaborators at Wageningen University (RAW 16.8 GB) and were able to analyse these data in less than 30 minutes on an average workstation. Genetic maps obtained were then used by the researchers to improve the resolution of their subsequent eQTL mapping. The observed increase in resolution allows the detection of leads in smaller regions of interest, and lowers costs for follow up analyses of the traits of interest.

Application of MQM has identified new genetic loci of interest in different studies. The benefits are thus more insight in the underlying genetics. This leads to an increase in knowledge about how to improve pathways for production of commercially interesting biological or therapeutic substances or products. Furthermore, by increasing the amount of loci detected using MQM (as well as cell type specific eQTL mapping), we are able to better understand the genetic basis of heritable traits such as crop yield, seed quality, and disease susceptibility, for example Crohn's disease. MQM improves our statistical model by adding genetic loci as explanatory variables, while cell type specific eQTL mapping does this by incorporating environmental effects. Although we only applied our cell type specific eQTL mapping to neutrophils and lymphocytes in whole blood as proof of principle. We still observed that SNPs associated with Crohn's disease are present more often than randomly expected, and show a neutrophil specific eQTL effect.

Our system xQTL workbench was used as a storage and computation platform in

wormQTL and WormQTL-HD databases. Furthermore an *Arabidopsis* database is presently being developed to contain similar features as the womQTL database. We demonstrated the need for translation analysis in WormQTL-HD and connected worm and human data. This shows (and confirms findings from other studies) that worms can be used as a model organism to test certain hypotheses derived from observations in humans. The advantages provided by the xQTL workbench allow other researchers to accomplish the same aim: storage of large amounts of data and performing reliable/ traceable analyses while avoiding duplication of efforts.

An illustrative example is the following. When I started my PHD research in 2009, my largest data set was 100mb (24 mice measured on 30,000 probe microarrays). This was easily manageable on a single computer and involved no storage or computational issues. Just four years later we cannot perform the same analysis because the size of the data obtained from the same experiment has increased so much, a database that can handle many gigabytes is required. Such an amount of data can only be stored using dedicated hardware, which also stores the complex relationships between the data. To allow integration of information from worms and humans in less than a minute, computations need to be optimised and distributed across many nodes of a cluster. The knowledge gain of this high-throughput data is that biologists are able to develop new hypotheses while browsing data collected by hundreds of people over a time scale of decades.

6.10 Future perspectives

As interesting and novel developments are created at the interface of the research fields of systems genetics and bioinformatics, this section first looks into the future of both fields separately, after which some ideas about the interdisciplinary domain are presented.

Genetics research is producing more data than ever before. This leads to an almost exponential increase in data similar to Moore's law of computing, which states that '*Every 12 to 24 months CPU power doubles*' [51, 54, 277]. Large scale sequencing facilities such as the Sanger Institute (USA) and the Beijing Genomics Institute (China) continue to produce an ever increasing stream of sequencing data for decreasing costs per analysis. This allows researchers to collect data on an unprecedented scale, but also triggers the massive re-use of previous data.

The field of epigenetics is also gaining more and more attention, in particular the field of DNA methylation. The creation of epiRILs [278] means that we have a genetic map to associate our phenotype with, but additionally we can look for epigenetic inheritance

of phenotypes using an epigenetic map. This seems trivial, but in fact it describes the easiest situation which already doubles the required amount of effort for mapping QTLs, because we need to map to a genetic map and an epigenetic map. Furthermore, epigenetic maps currently available only show the inheritance of stable epigenetic states in an experimental population without any sequence variation. Therefore, the situation is even more complex in the case of natural populations which have a mixture of genetic and epigenetic variations [279]. At present the stability of epigenetics is still uncertain and it may lead to a reconsideration of certain genetic theories, in particular those addressing the inheritance of DNA. This daunting challenge of first understanding how epigenetics play its role in inheritance, and then using this knowledge to improve our QTL mapping ability, will be a major and exciting challenge in the coming years.

Cell type specific eQTLs are just one of many possible interaction models to take into account during eQTL mapping. The current trend of performing data driven research in which we increase the amounts of data being collected and connect everything with everything is expected to remain in the future. The statistical downside is that it causes multiple testing issues [280], population stratification issues, and spurious associations by hidden factors [281] and/or batch effects.

The field of bioinformatics faces the computational challenges arising from the above trends and expected developments. The question is how we can contribute to solve these issues, demonstrate the added value of our applications, and present our legacy for the coming years.

Computational power has always increased exponentially over time. Recently, however, the Intel Company announced that Moore's law is not valid anymore [51, 52]. The last 3 years and the upcoming years will show a less then exponential growth in power per CPU [52]. It can be expected that the production of new and cheap computers will slightly offset this problem. Therefore, while separate CPUs will fail to show exponential increases in speed, another development is the emergence of new and improved dedicated machines, such as ARM cores, Field Programmable Gate Arrays, and ASICS miners for dedicated computation. Although these machines are less or equally powerful in general, they are designed to excel in just one specific task, compared to a CPU which does many different tasks. Our research showed that several 'smart' algorithms, of which some could be hard coded into an ARM core or FPGA, allows the same computations to be executed much faster compared to a CPU. Taken together, this will still allow scaling up of, for example, cell type specific eQTL analysis, to even larger data sizes.

We advocate using linear scaling of multi-core machines or large compute clusters. Although this is not the solution to our problems, using these systems in such a manner

that data are presented in a summarised way to biologists (as is done by wormQTL) will help to relieve some of the immediate need for collaborators. In our opinion the xQTL workbench will continuously be developed to suit emerging new technologies in the field of data storage and computation. Additionally, using R as Esparanto between multiple languages means we are less dependent on a single language, and are able to switch from one storage strategy to another, while the computational part of the system does not 'see' this change because it only sees the R glue layer.

The computational issues arising from high-throughput data can be partly solved with smarter algorithms such as Pheno2Geno, MQM, CTL and cell type specific QTL mapping. Some of these algorithms are already being used actively in research, and show that improved tools give either improved performance or improved power to detect results.

The xQTL workbench is a collaborative system in which bioinformaticians and biologists contribute to create a system which can serve both them and the communities to which they belong. Collaborators are curating data and adding new analysis routines to the wormQTL system. Pheno2Geno connects to R/qtl but we expect other QTL mapping packages without map creation facilities to be also interested to add Pheno2Geno to their pipeline. Additionally the whole high-throughput eQTL mapping pipeline is available for other researchers to use and learn from. We urge all people performing GWAS to perform meta-analysis (when possible) to improve their results and employ a strategy such as cell type specific eQTL mapping.

How and where do bioinformatics and systems genetics come together? High-throughput data is the connecting factor between both fields. Systems genetics cannot exist without the help of high-throughput methodologies such as whole genome sequencing and automated proteomics, because only with these technologies we are able to build a comprehensive picture of all the biomolecular levels that compose organisms. To generate biologically relevant knowledge from these large amounts of data, geneticists cannot do without bioinformaticians.

Here we have shown our contributions to the field of bioinformatics such as novel infrastructure, high-throughput computational methodologies and more optimised algorithms. These advances aid faster data processing, reduce redundancy during analysis, and create novel tools to get more out of existing data. New tools and improvements of existing tools, as presented in this thesis, combined with computational advances, creates the opportunity to do new things which were not possible before, such as large scale cell type specific eQTL mapping on 6,000+ human GWAS samples. Our tools will help systems geneticists to do their job better and in less time than before.

7 Additional for dissertation

This chapter contains additional sections required for a promotion at the University of Groningen, such as a Dutch summary, a list of abbreviations, and background information on the author.

7.1 Dutch summary / Nederlandse samenvatting

Systeem genetica is het interdisciplinaire werkveld dat zich bezig houdt met de gevolgen van genetische variatie op alle biomoleculaire niveaus van een biologisch systeem. Het effect van deze genetische variatie leidt tot verschillen in fenotypen, zoals opbrengst en resistentie. Het doel van systeem genetica is om grip te krijgen op biologische systemen door variatie in te delen in drie categorieën: genetische effecten, omgevingseffecten en (random) residuele error variantie, om vervolgens te kunnen verklaren hoe complexe fenotypen ontstaan uit een combinatie van deze effecten op alle biomoleculaire niveaus.

Tegenwoordig kan natuurlijk voorkomende genetische variatie (of genetische pertubatie) worden gebruikt om de genetische basis van fenotypische variatie te ontrafelen op verschillende biomoleculaire niveaus, zoals: genetica, transcriptomics, proteomics en metabolomics. Biomoleculaire metingen gecombineerd met pertubatie van verschillende omgevingsfactoren stellen ons in staat om de invloed van omgevingseffecten en de interactie tussen omgeving en genetica te onderzoeken. Experimenteel ontwerp en statistiek worden gebruikt om error variantie te minimaliseren. Om onderzoek te doen naar alle factoren die invloed uitoefenen op biologische systemen, is het noodzakelijk om grote hoeveelheden data te verzamelen van vele individuen, vele weefsels, op alle bekende biomoleculaire niveaus.

Moderne high-throughput technieken genereren grote hoeveelheden genomische, transcriptomic, proteomic en metabolomic data. De omvang van de data die verzameld wordt en de diversheid aan technieken die worden gebruikt zorgen voor een enorme uitdaging voor bioinformatici. Deze thesis zal ingaan op onze oplossingen voor deze ‘Big Data’ uitdaging binnen de systeem genetica. Hier wordt geadviseerd om gebruik te maken van geoptimaliseerde algoritmes, zoals: Pheno2Geno en Multiple QTL mapping. Ook wordt een collaboratieve aanpak voor dataopslag en gerelateerde computationele aspecten (xQTL workbench) gepromoot, om high-throughput data op te slaan en te analyseren binnen de systeem genetica.

Hoofdstuk 1 bevat een historische introductie in systeem genetica en bespreekt de uitdagingen die geleid hebben tot het schrijven van deze thesis, zoals de gigantische toename van data productie wat zorgt voor een toenemende complexiteit bij het diagnostiseren van patienten of het selecteren van gewassen voor optimale opbrengst.

Hoofdstuk 2 gaat over Pheno2Geno, een methode voor het construeren van genetische kaarten vanuit grootschalige omics data sets. De theorie achter het maken van genetische kaarten is ongeveer 100 jaar oud. De meeste software voor het maken van genetische kaarten in modelorganismes is geschreven in de jaren 80, en is vaak nog niet

aangepast aan nieuwe ontwikkelingen zoals multi-core computers en cluster computing. Pheno2Geno is software voor het maken van genetische kaarten uit 'Big Data' verkregen door experimenten die gebruik maken van tilling arrays of RNA sequence experimenten.

Hoofdstuk 3 omschrijft de implementatie van het Multiple QTL mapping (MQM) algoritme in R/qtl. Hiermee voegen we een nieuw algoritme toe aan de R/qtl toolset, een toolset speciaal ontwikkeld voor het mappen van QTLs in ingeteelde (muis)populaties. R/qtl vormt de basis voor meerdere tools die allemaal gebouwd zijn rond een gedeelde datastructuur. Deze structuur maakt het gemakkelijk om software aan te passen en nieuwe tools toe te voegen. Door de vele tools kunnen onderzoekers snel wisselen van methode wanneer dit vereist is, of verschillende methodes vergelijken zonder een nieuw software pakket te hoeven leren.

Huidige werkzaamheden over het gebruik van verschillen in correlatie om interactie netwerken te genereren en celtype specifieke eQTL effecten op te sporen worden beschreven in hoofdstuk 4. Correlated Traits Locus analyse (of CTL mapping) stelt onderzoekers in staat om genetische loci te vinden die geassocieerd zijn met correlatie verschillen tussen segregerende fenotypen. Een variant op deze methode is waardevol gebleken om celtype specifieke expressie QTL (eQTL) effecten te ontdekken. Deze effecten kunnen worden gebruikt om mengsels van cellen te ontwarren. We passen onze nieuwe methode toe op een celmengsel van geheel bloed genexpressie data en laten zien dat we in staat zijn om celltype specifieke eQTLs te detecteren.

In hoofdstuk 5 presenteren we onze ideeën voor een generiek opslag- en rekenplatform voor systeem genetica. Ons xQTL workbench systeem wordt momenteel gebruikt als een back end voor de WormQTL en WormQTL-HD databases. In xQTL workbench kunnen gebruikers hun gegevens opslaan en delen in een lokale of webomgeving, en analyses doen op data sets met behulp van de kracht van gedistribueerde computing. Het wordt standaard geleverd met QTL mapping tools, zoals: R/QTL, Plink en qtlbim maar het biedt ook een webinterface, data importers, API's en visualisaties.

Ik vertrouw erop dat je geniet van het lezen van dit proefschrift, zoals ik heb genoten van het maken ervan.

7.2 Abbreviations and acronyms

ABA:	Abscisic acid
ANOVA:	Analysis of variance
API:	Application Programming Interface
AR:	After ripened seeds
BC:	Backcross
bp:	Base pair(s)
cM:	centimorgan
CIM:	Composite Interval Mapping
CPNN:	Collaborative computing project for NMR
CSV:	Comma separated values
CTL:	Correlated traits locus
DE:	Differential expression (analysis)
designGG:	Experimental design of genetical genomics software
DRY:	Principle of don't repeat yourself
DSL:	Domain Specific Language
eQTL:	Expression quantitative trait locus
EBI:	European Bioinformatics Institute
FDR:	False discovery rate
FPGA:	Field-programmable gate array
FINDIS:	Finish disease database
GABA:	Gamma-Aminobutyric acid
GEMs:	Gene expression based genetic marker
GEN2PHEN:	EU project to unify human and model organism genetic variation databases
GGG:	Generalized genetical genomics
GMOD:	Generic model organism database project
GUI:	Graphical user interface
GWAS:	Genome Wide Association Study
GWL:	Genome Wide Linkage analysis
HIF:	Heterogeneous inbred family
HGVBaseG2P:	Human genome variation database of genotype-to-phenotype information
HTML:	Hypertext markup language
IDE:	Integrated Development Environment
JAR:	Java Software Archive
LGPL:	Lesser general public license
MAGE-TAB:	Microarray gene expression tab delimited file format
Mbp:	Mega base pairs = 1,000,000 bp

MOLGENIS:	Molecular genetics information systems toolkit
MPI:	Message Passing Interface
mQTL:	Metabolite abundance quantitative trait locus
MQM:	Multiple QTL mapping
NCBI:	National Center for Biotechnology Information
NMR:	nuclear magnetic resonance
NordicDB:	Nordic Control Cohort Database
OBF:	Open Bioinformatics Foundation
OntoCAT:	Ontology common API toolkit
PCA:	Principal component analysis
PD:	Primary dormant seeds
PEAA:	Patterns for enterprise application architecture
pQTL:	Protein abundance quantitative trait locus
QTL:	Quantitative trait locus
RDF:	Resource description format
REST:	Representative state transfer web services
RP:	Seeds at radicle protrusion
RIL:	Recombinant inbred line
SNOW:	Simple Network of Workstations
SNP:	Single Nucleotide Polymorphism
SOAP:	Simple Object Access Protocol
SOM:	Self organizing map
SQL:	Structured Query Language
SVM:	Support vector machines
UML:	Uniform data Modeling Language
vQTL:	Variance quantitative trait locus
WAR:	Web Application aRchive file
XML:	Extensible Markup Language
XGAP:	Extensible genotype and phenotype software platform.

7.3 Acknowledgements

"*Knowledge is in the end based on acknowledgement*"
- Ludwig Wittgenstein (1889 - 1951)

There are many people who need to be thanked and acknowledged, here are the people who made the cut:

Ritsert C. Jansen to whom I owe my life in science. Thank you for giving me the chance to work in your group and develop my skills in research and education. Thank you for giving me a stimulating place to work during my master and PhD, for all your help during all the different phases of this thesis, and for reminding me to use the **Klazien Offens** card when dealing with administrative and bureaucratic issues. I cannot thank you both enough for your help and support.

Anna Mulder, my girlfriend, for all the times I left you and **Oscar** (our cat) for my science endavours. Thank you for your trust in me and for giving me a home to come to. I would not have been able to finish this thesis in time, and in fashion, without your support and style.

My family; my loving and trusting father **Bert Arends** for everything you did (and still do) for all of us, and for keeping us together as a family. **Dinie Wakker**, for supporting my dad and us in all our crazy endavours. My older brother **Johan** and my younger sister **José** for being there when I was growing up, sharing many meals, discussions, and ofcourse the good times and bad ones. My grandmother **Oma ter Arkel**, who puts up with me not visiting her as often as I should. The rest of my family from both the **Ter Arkel** and the **Arends** side, and my parents-in-law **Greetje** and **Bart**. Thank you for your support.

Pjotr Prins Always Forgotten, Never Ignored.

I would like to thank all the people past and present at the **Groningen BioInformatics Centre** (GBIC) group for putting up with me. In particular my paranimphs: **Yang Li** for your mentorship and all the vibrant discussions we had at GBIC, but also for your help when co-supervising bachelor and/or master students. **Joeri van der Velde**, my coworker but mostly my friend for the good times we had hacking on Java and R in Haren and at the UMCG. Thank you for making traveling more enjoyable on the trips that took us to different far away places such as Boston and Prague.

Morris Swertz, my co-promotor, you were a great help when writing chapter 5 and you always made me feel at home at the **Genomics Coordination Centre** (GCC). Also all the guys from the GCC department for all the presentations, discussions with **Erik T. Roos**, coffee served by **Roan Kanninga** and 'end of sprint' beers with a great team.

I want to thank everyone involved in my promotion, but especially the reviewers (**Prof. dr. G. A. Brockmann**, **Prof. dr. E. O. de Brock** and **Prof. dr. M. H. Hofker**), for reading this thesis and giving me feedback on how to improve it.

Frank Johannes, **Gerald J. te Meerman** and **Martijn Dijkstra** for pre-reading this thesis and proving me with constructive feedback and suggestions. Additionally, I would like to thank everyone from **JohannesLab** such as **René Wardenaar** and **Pariya Behrouzi** for participating in tuesday morning seminars, journal clubs and coffee breaks.

Lude Franke and **Harm-Jan Westra** for the very informative collaboration on the human GWA study (chapter 4.2), I learned a LOT from you, and I am extremely thankful for your help with the final chapters of this thesis.

I have been very lucky to meet and work together with some of the cleverest minds in animal breeding and genetics. Being able to discuss my ideas with people such as **Karl W. Broman**, **Gary A. Churchill**, **Alan D. Attie**, **Rob Williams** and **Brian S. Yandell** made me push myself to the limits and beyond. I would like to thank all of you for the nice times I spent at the other side of the Pacific Ocean.

I was fortunate to have good collaboration in The Netherlands with Wageningen University, especially **Wilko Ligterink**, **Henk Hilhorst** and **Ronny Joosen** for doing the *Arabidopsis thaliana* experiments, and trusting me to analyze their data. I learned a lot from working together with you on the more 'biological' parts of this thesis. Also I would like to thank **Learning From Nature** (LFN) who invited me to Wageningen for lectures on using R/qtl on large data.

Furthermore I have had the pleasure to work with many dedicated students (now sometimes even colleagues) along the way: **Konrad Zych** for his dedication and help with Pheno2Geno (chapter 2). **Mark de Haan** for his work on the xQTL workbench system and the review work on Pheno2Geno. **Yalan Bi** for working on the *A. thaliana* gene expression data and alternative splicing work, which unfortunately was not finished in time for this thesis. **Adriaan van der Graaf** for showing me that even jocks can be turned into nerds. And naturally all the other students I have been privileged to work with during these last years.

I would like to thank the **European Union** for setting up the **SysGeNet** network, I have met many people and learned a lot about the mouse community in Europe. Especially I would like to thank **Klaus Schughart** for starting and maintaining the network, and **Leonard Schalkwyk** for inviting me to the UK for lectures at **King's College**.

Richard Stremme and **Evert van der Velde** for always being available after work for a beer and for providing entertainment and the needed distraction when I was fed up with writing this thesis.

The **University of Groningen** (RUG) for educating me and providing me with a place to do my research and feel at home. And finally the **University Medical Centre Groningen** (UMCG) for the pleasant working atmosphere and good lunches.

7.4 About the author

7.4.1 Curriculum Vitae

Danny Arends was born on the 15th of Juli 1983 in the city of Zwolle, located in the middle of the Netherlands. After moving twice, Danny grew up in Heiligerlee, a small village in the north of the Netherlands, where he attended elementary school.

The small school was a perfect match for young Danny during his childhood. Here, he quickly fell in love with mathematics and the wondrous world of numbers and patterns. With the advent of computers at the elementary school, the love for computers and their inner workings also began to develop.

Heiligerlee is located next to a small forest called 'De Hoogte'. This is were Danny and his friends went to build tree huts on a daily basis, throw snowballs in winter, and wage war with kids from the other school in Heiligerlee. The forest, combined with growing up amidst animals on a farm-like residence, sparked young Danny's interest in biology.

After elementary school, Danny went to the Ubbo Emmius lyceum in Stadskanaal (Groningen). The large scale VWO was a big change compared to the small elementary school. Long hours at school together with a long bus ride to and from school, made for long days. Fortunately new friends were made in the classroom and during the long bus rides. Danny finished high school after 6 years, taking mostly exact courses such as mathematics, physics and chemistry.

At age 17, obtaining a university degree was the next step in his career. Because of his interests in computers, Danny decided that computer science would be a good match. This turned out to be not so true. While deeply intrigued by the subject of computational machines, he was not satisfied with just studying the workings of a machine built by man.

After two years of computer science at the University of Groningen, Danny decided it was time for a change. Computer science was replaced by Life Science & Technology, a bachelor which was recently formed as a collaboration between the Biology faculty and Medical Sciences. He finished his bachelor in record tempo, partly due to the exemptions obtained from doing two years of computer science. A master in molecular

biology was quickly selected after being introduced to bioinformatics at GBIC during a previous bachelor project. The molecular biology master allowed for customization of the courses taken, and bioinformatics became the main theme in all of the master theses produced. The first thesis: “Machine learning to predict transcriptional regulation in prokaryotes” was produced in the lab of prof. dr. O. P. Kuipers. The second master thesis: “High performance computing of QTLs for experimental crosses” was done in the lab of prof. dr. R. C. Jansen under the supervision of Pjotr Prins. Parts of this second master project are found in this thesis (chapter 3).

After graduating his university master cum laude, Danny started a PHD project in the lab of prof. dr. R. C. Jansen at the Groningen Bioinformatics Centre, with a focus on the use of bioinformatic tools to handle current challenges in genetics and statistics. These four years of research at the GBIC have resulted in the thesis you are currently reading.

7.4.2 List of publications

Authored:

Danny Arends*, Pjotr Prins*, Ritsert C. Jansen and Karl W. Broman
R/qtl: high-throughput Multiple QTL mapping
Bioinformatics 26(23):2990-2 (2010)

Ronny V. L. Joosen*, **Danny Arends***, Leo Willems, Wilco Ligterink, Henk Hilhorst and Ritsert C. Jansen
Visualizing the genetic landscape of Arabidopsis seed performance
Plant Physiology 158(2):570-89 (2011)

Heorgy V. Byelas*, **Danny Arends***, Freerk van Dijk, K. Joeri van der Velde, Laurent Francioli, Martijn Dijkstra, Alexandros Kanterakis, Ishtiaq Ahmad, David van Enckvoort, Leon Mei, Peter Horvatovich, other members of BBMRI-NL, NBIC and Target, Morris A. Swertz
Large scale NGS pipelines using the MOLGENIS platform: processing the Genome of the Netherlands
Proceeding of: 12th Annual Bioinformatics Open Source Conference BOSC 2011

Danny Arends*, K. Joeri van der Velde*, Pjotr Prins, Karl W. Broman, Steffen Möller, Ritsert C. Jansen and Morris A. Swertz
xQTL workbench: a scalable web environment for multi-level QTL analysis
Bioinformatics 28(7):1042-4 (2012)

L. Basten Snoek*, K. Joeri Van der Velde*, **Danny Arends***, Yang Li*, Antje Beyer, Mark Elvin, Jasmin Fisher, Alex Hajnal, Michael O. Hengartner, Gino B. Poulin, Miriam Rodriguez, Tobias Schmid, Sabine Schrimpf, Feng Xue, Ritsert C. Jansen, Jan E. Kammenga and Morris A. Swertz
WormQTL: Public archive and analysis web portal for natural variation data in Caenorhabditis spp
Nucleic Acids Research 41(DB issue):D738-43 (2012)

Ronny V. L. Joosen*, **Danny Arends***, Yang Li*, Leo Willems, Joost J. B. Keurentjes, Wilco Ligterink, Ritsert C. Jansen and Henk Hilhorst
Identifying genotype-by-environment interactions in the metabolism of germinating Arabidopsis seeds using Generalized Genetical Genomics
Plant Physiology 162(2):553-66 (2013)

Harm-Jan Westra*, **Danny Arends***, ..., Ritsert C. Jansen and Lude Franke
Cell-type specific eQTL analysis without the need to sort cells
Submitted (2014)

Co-authored:

Morris A. Swertz, Martijn Dijkstra, Tomasz Adamusiak, K. Joeri van der Velde, Alexandros Kanterakis, Erik T. Roos, Joris Lops, Gudmundur A. Thorisson, **Danny Arends**, George Byelas, Juha Muilu, Anthony J. Brookes, Engbert O. de Brock, Ritsert C. Jansen and Helen E. Parkinson
The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button
BMC Bioinformatics 1 Suppl 1(Suppl 12):S12 (2010)

Morris A. Swertz, K. Joeri van der Velde, Bruno M. Tesson, Richard A. Scheltema, **Danny Arends**, Gonzalo Vera, Rudi Alberts, Martijn Dijkstra, Paul Schofield, Klaus Schughart, John M. Hancock, Damian Smedley, Katy Wolstencroft, Carole Goble, Engbert O. de Brock, Andrew R. Jones, Helen E. Parkinson and Ritsert C. Jansen
XGAP: *a uniform and extensible data model and software platform for genotype and phenotype experiments*
Genome Biology 11(3):R27 (2010)

Klaus Schughart, **Danny Arends**, P. Andreux, R. Balling, Pjotr Prins, *et al.*
SYSGENET: *a meeting report from a new European network for systems genetics*
Mammalian Genome 21(7-8):331-6 (2010)

Rudolf S. N. Fehrmann, Ritsert C. Jansen, Jan H. Veldink, Harm-Jan Westra, **Danny Arends**, Marc Jan Bonder, Jingyuan Fu, Patrick Deelen, Harry J. M. Groen, Asia Smolonska, Rinse K. Weersma, Robert M. W. Hofstra, Wim A. Buurman, ..., Lude Franke
Trans-eQTLs Reveal that Independent Genetic Variants Associated With a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA
Plos Genetics 7(8):e1002197 (2011)

Caroline Durrant, Morris A. Swertz, Rudi Alberts, **Danny Arends**, Steffen Möller, Richard Mott, Pjotr Prins, K. Joeri van der Velde, Ritsert C. Jansen and Klaus Schughart
Bioinformatics tools and database resources for systems genetics analysis in mice - a short review and an evaluation of future needs
Briefings in Bioinformatics 13(2):135-42 (2011)

K. Joeri van der Velde*, Mark de Haan, Konrad Zych, **Danny Arends**, L. Basten Snoek, Jan E. Kammenga, Ritsert C. Jansen, Morris A. Swertz and Yang Li
WormQTLHD - a web database for linking human disease to natural variation data in C. elegans
Nucleic Acids Research 42(1):D794-801 (2014)

Elisabeth M. Lodder, Brendon P. Scicluna, L. Beekman, **Danny Arends**, *et al.*
Multiple QTL mapping of cardiac collagen deposition in an F2 population of Scn5a mutant mice reveals interaction between Fgf1 and Pdlim3, Gpr158 & Itga6
European Heart Journal 34(suppl 1), 2602 (2013)

In preparation / under Review / in press:

Konrad Zych, K. Joeri van der Velde, Ronny V. L. Joosen, Wilco Ligterink, Ritsert C. Jansen and **Danny Arends**
Pheno2Geno - High-throughput generation of genetic markers and maps from molecular phenotypes
Submitted

Danny Arends, Pjotr Prins, Harm-Jan Westra, Yang Li, Lude Franke and Ritsert C. Jansen
Correlated Traits Locus mapping
Draft

Steffen Möller, René Schönfelder, Hajo Krabbenhöft, Benedikt Bauer, Yask Gupta, Pjotr Prins, **Danny Arends**, *et al.*
TiQS: web environment for expression QTL analysis
Submitted

Dunia P. del Carpio, Ram K. Basnet, **Danny Arends**, Ke Lin, Ric C. H. de Vos, Dorotha Muth, Jan Kodde, Kim Boutilier, Johan Bucher, Xiaowu Wang, Ritsert C. Jansen, Guusje Bonnema
Regulatory Network of Secondary Metabolism in Brassica rapa: An Insight In The Glucosinolate Pathway
Submitted

Acknowledged in:

Nino Demetrashvili, Edwin R. van den Heuvel and Ernst C. Wit
Probability genotype imputation method and integrated weighted lasso for QTL identification
BMC Genetics 14:125 (2013)

Yang Li, Morris A. Swertz, Gonzalo Vera, Jingyuan Fu, Rainer Breitling and Ritsert C. Jansen
designGG: an R-package and web tool for the optimal design of genetical genomics experiments
BMC Bioinformatics 10:188 (2009)

Yang Li, Rainer Breitling and Ritsert C. Jansen
Generalizing genetical genomics: getting added value from environmental perturbation
Trends in Genetics 24:518-524 (2008)

7.4.3 List of presentations

R/xqtl: High-throughput modeling, mapping and exploration of Big Data
SYSGENET meeting Braunschweig, April 2010

Introduction into QTL analysis
Dynamic Presentation University of Groningen, Aug 2010

MQM and HPC for R/qtl
CSBG meeting Wageningen University, Sept 2010

Introduction into QTL mapping - Bioinformatics I
University of Groningen June 2011

(Re)Construction of genetic maps from gene expression data
GBIC University of Groningen, July 2011

R/qtl for Big Data
MIT Department of Biology, invited by Jeroen PJ Saeij MIT, Boston (MA), May 2011

The Challenge of Big Data Genetical Genomics
NCSA, invited by Victor Jongeneel and Chris Fields NCSA, Urbana (IL), May 2011

Introduction into QTL mapping - Learning From nature (LFN)
Wageningen University, Feb 2012

Computer practical / tutorial - Learning From nature (LFN)
Wageningen University, Feb 2012

Teaching at Summer Course R, R/qtl and GeneNetwork
King's college (London, UK), Sept 2013

Overview R/qtl
Humbold University (Berlin, DE), Nov 2013

Teaching at Summer Course R, R/qtl and GeneNetwork
King's college (London, UK), June 2014

7.4.4 List of posters

User friendly cluster computing for QTL analysis
Danny Arends, K. Joeri van der Velde
NBIC Conference April 2010, & ISMB July 2010 Boston, USA

Multiple QTL mapping poster
Danny Arends, Pjotr Prins, Karl W. Broman and Ritsert C. Jansen
GBIC day September 2010 Groningen, The Netherlands

Pheno2Geno poster
Konrad Zych, **Danny Arends**, Ritsert C. Jansen
NBIC April 2012 Lunteren, The Netherlands

Pheno2Geno poster
Konrad Zych, **Danny Arends**, Ritsert C. Jansen
ECCB / ESCS Sept 2012 Basel, Swiss

GWAS in potato
Konrad Zych, **Danny Arends**, Ritsert C. Jansen
Kings College Sept 2013 London, UK

7.4.5 Awards

Best oral presentation at CTC 2014 in Berlin (Germany) (2014)

1st place Poster Award 'Pheno2Geno' at ESCS2012 Basel (Swiss) (2012)

Travel Grant by NBIC for ISMB in Boston (MA, USA) (2010)

2nd place Poster Award xQTL workbench NBIC Conference (Lunteren) (2010)

Travel Grant 'Short course on systems genetics' in Bar Harbor (MA, USA) (2009) by JAX laboratory

Bibliography

[1] E S Lander and D Botstein. *Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms.* **Proceedings of the National Academy of Sciences of the United States of America**, 83(19):7353–7, October 1986.

[2] J G Hacia, J B Fan, O Ryder, *et al. Determination of ancestral alleles for human singlenucleotide polymorphisms using high-density oligonucleotide arrays.* **Nature genetics**, 22(2):164–7, June 1999.

[3] E R Mardis. *Next-generation DNA sequencing methods.* **Annual review of genomics and human genetics**, 9:387–402, January 2008.

[4] H Hayatsu, K Negishi, M Shiraishi, K Tsuji, and K Moriyama. *Chemistry of bisulfite genomic sequencing; advances and issues.* **Nucleic acids symposium series**, 51:47–48, 2007.

[5] P Collas. *The current state of chromatin immunoprecipitation.* **Molecular biotechnology**, 45:87–100, 2010.

[6] P J Park. *ChIP-seq: advantages and challenges of a maturing technology.* **Nature Reviews Genetics**, 10:669–680, 2009.

[7] D A Lashkari, J L DeRisi, J H McCusker, *et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis.* **Proceedings of the National Academy of Sciences of the United States of America**, 94:13057–13062, 1997.

[8] T Lee and A C S Luk. *Tiling Arrays.* **Volume 1067 of Methods in Molecular Biology.** Humana Press, Totowa, NJ, 2013.

[9] Z Wang, M Gerstein, and M Snyder. *RNA-Seq: a revolutionary tool for transcriptomics.* **Nature Reviews Genetics**, 10:57–63, 2009.

[10] P H O'Farrell. *High resolution two-dimensional electrophoresis of proteins.* **The Journal of biological chemistry**, 250(10):4007–21, May 1975.

[11] R J Deshaies. *Charting the Protein Complexome in Yeast by Mass Spectrometry.* **Molecular & Cellular Proteomics**, 1(1):3–10, November 2001.

[12] J Fasolo and M Snyder. *Protein microarrays.* **Methods in molecular biology** (Clifton, N.J.), 548:209–222, 2009.

[13] R Aebersold and M Mann. *Mass spectrometry-based proteomics.* **Nature**, 422:198–207, 2003.

[14] R Espina, L Yu, J Wang, *et al. Nuclear magnetic resonance spectroscopy as a quantitative tool to determine the concentrations of biologically produced metabolites: implications in metabolites in safety testing.* **Chemical research in toxicology**, 22(2):299–310, February 2009.

[15] K W Broman, H Wu, S Sen, and G A Churchill. *R/qtl: QTL mapping in experimental crosses.* **Bioinformatics** (Oxford, England), 19(7):889–90, May 2003.

[16] D Arends, P Prins, R C Jansen, and K W Broman. *R/qtl: high-throughput multiple QTL mapping.* **Bioinformatics** (Oxford, England), 26(23):2990–2, December 2010.

[17] G J Mendel. *Experiments in plant hybridization (1865).* **Verhandlungen des naturforschenden vereines in Brünn**, (1865):3–47, 1866.

[18] I H Stamhuis, O G Meijer, and E J Zevenhuizen. *Hugo de Vries on heredity, 1889-1903. Statistics, Mendelian laws, pangenes, mutations.* **Isis; an international review devoted to the history of science and its cultural influences**, 90(2):238–67, June 1999.

[19] T H Morgan, A H Sturtevant, H J Muller, and C B Bridges. *The mechanism of mendelian heredity.* **H. Holt and company**, 1915.

[20] T H Morgan. *Sex limited inheritance in drosophila.* **Science** (New York, N.Y.), 32(812):120–2, July 1910.

[21] R Ming and J Wang. *Sex chromosomes in flowering plants.* **American Journal of Botany**, 94(2):141–150, 2007.

[22] R A Fisher. *The Correlation between Relatives on the Supposition of Mendelian Inheritance.* **Transactions of the Royal Society of Edinburgh**, 399–433, 1919.

[23] R A Fisher. *The Genetical Theory of Natural Selection*, volume 154. 1930.

[24] O T Avery, C M MacLeod, and M McCarty. *Studies on the chemical nature of the substance inducing transformation of pneumococcal types.* **The Journal of experimental medicine**, 79(2):137–158, January 1944.

[25] J D Watson and F H C Crick. *Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.* **Nature**, 171(4356):737–738, April 1953.

[26] S E Luria and M L Human. *A nonhereditary, host-induced variation of bacterial viruses.* **Journal of bacteriology**, 64:557–569, 1952.

[27] E Pettersson, J Lundeberg, and A Ahmadian. *Generations of sequencing technologies.* **Genomics**, 93(2):105–11, February 2009.

[28] E S Lander and D Botstein. *Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children.* **Science** (New York, N.Y.), 236:1567–1570, 1987.

[29] E S Lander and D Botstein. *Mapping mendelian factors underlying quantitative traits using RFLP linkage maps.* **Genetics**, 121(1):185–199, January 1989.

[30] U R Rosyara, J L Gonzalez-Hernandez, K D Glover, K R Gedye, and J M Stein. *Family-based mapping of quantitative trait loci in plant breeding populations with resistance to Fusarium head blight in wheat as an illustration.* **Theoretical and applied genetics**, 118(8):1617–31, May 2009.

[31] R C Jansen and J P Nap. *Genetical genomics: the added value from segregation.* **Trends in genetics**, 17(7):388–91, July 2001.

[32] D Mehta, K Heim, C Herder, *et al. Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood.* **European journal of human genetics**, 21(1):48–54, January 2013.

[33] The International HapMap Consortium. A *haplotype map of the human genome.* **Nature**, 437(7063):1299–320, October 2005.

[34] H Sinha, B P Nicholson, L M Steinmetz, and J H McCusker. *Complex genetic interactions in a quantitative trait locus.* **PLoS genetics**, 2(2):e13, February 2006.

[35] M A L West, K Kim, D J Kliebenstein, *et al. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis.* **Genetics**, 175(3):1441–50, March 2007.

[36] C S Haley and S A Knott. A *simple regression method for mapping quantitative trait loci in line crosses using flanking markers.* **Heredity**, 69(4):315–24, October 1992.

[37] Z B Zeng. *Precision mapping of quantitative trait loci.* **Genetics**, 136(4):1457–1468, April 1994.

[38] R C Jansen. *Interval mapping of multiple quantitative trait loci.* **Genetics**, 135(1):205– 211, September 1993.

[39] R C Jansen and P Stam. *High resolution of quantitative traits into multiple loci via interval mapping.* **Genetics**, 136(4):1447–1455, April 1994.

[40] D W Threadgill. *Meeting report for the 4th annual Complex Trait Consortium meeting: from QTLs to systems genetics.* **Mammalian genome**, 17(1):2–4, January 2006.

[41] J H Nadeau and A M Dudley. *Genetics. Systems genetics.* **Science** (New York, N.Y.), 331(6020):1015–6, February 2011.

[42] R B Brem, G Yvert, R Clinton, and L Kruglyak. *Genetic dissection of transcriptional regulation in budding yeast.* **Science** (New York, N.Y.), 296(5568):752–5, April 2002.

[43] G Yvert, R B Brem, J Whittle, *et al. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors.* **Nature genetics**, 35(1):57–64, September 2003.

[44] M Morley, C M Molony, T M Weber, *et al. Genetic analysis of genome-wide variation in human gene expression.* **Nature**, 430(7001):743–7, August 2004.

[45] R Development Core Team and R Core Team. *Computational Many-Particle Physics, volume 739 of Lecture Notes in Physics.* **Springer**, Berlin, Heidelberg, 2008.

[46] R Breitling, Y Li, B M Tesson, *et al. Genetical genomics: spotlight on* QTL *hotspots.* **PLoS genetics**, 4(10):e1000232, October 2008.

[47] L A Hindorff, P Sethupathy, H A Junkins, *et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* **Proceedings of the National Academy of Sciences of the United States of America**, 106(23):9362–7, June 2009.

[48] R S N Fehrmann, R C Jansen, J H Veldink, *et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA.* **PLoS genetics**, 7(8):e1002197, August 2011.

[49] T Zeller, P Wild, S Szymczak, *et al. Genetics and beyond–the transcriptome of human monocytes and disease susceptibility.* **PloS one**, 5(5):e10693, January 2010.

[50] J E Powell, A K Henders, A F McRae, *et al. The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics.* **PloS one**, 7(4):e35430, January 2012.

[51] Editorial. *Metagenomics versus Moore's law.* **Nature Methods**, 6(9):623–623, September 2009.

[52] A Shah. *Intel: Keeping up with moores law becoming a challenge*, 2013.

[53] E E Schadt, M D Linderman, J Sorenson, L Lee, and G P Nolan. *Computational solutions to large-scale data management and analysis.* **Nature Reviews Genetics**, 11(9):647–57, September 2010.

[54] G E Moore. *Cramming More Components Onto Integrated Circuits.* **Proceedings of the IEEE**, 86(1):82–85, January 1998.

[55] L M Silva and R Buyya. *Parallel programming models and paradigms.* **High Performance Cluster Computing: Architectures and Systems 2**, pages 4–27, 1999.

[56] J Qiu, J Ekanayake, T Gunarathne, *et al. Parallel Programming Models and Paradigms.* **BMC bioinformatics**, 11 Suppl 1(Suppl 12):S3, January 2010.

[57] I Foster, Y Zhao, I Raicu, and S Lu. *Cloud Computing and Grid Computing 360-Degree Compared.* **2008 Grid Computing Environments Workshop**, pages 1–10, November 2008.

[58] O Trelles, P Prins, M Snir, and R C Jansen. *Big data, but are we ready?* **Nature Reviews Genetics**, 12(3):224, March 2011.

[59] K Krampis, T Booth, B Chapman, *et al. Cloud BioLinux: pre-configured and ondemand bioinformatics computing for the genomics community.* **BMC bioinformatics**, 13(1):42, January 2012.

[60] M A L West, H van Leeuwen, A Kozik, *et al. High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis.* **Genome research**, 16(6):787–95, June 2006.

[61] M J Truco, H Ashrafi, A Kozik, *et al. An Ultra High-Density, Transcript-Based, Genetic Map of Lettuce.* **G3** (Bethesda, Md.), March 2013.

[62] G Gort and F A van Eeuwijk. *Codominant scoring of AFLP in association panels.* **Theoretical and applied genetics**, 121(2):337–51, July 2010.

[63] J Quackenbush. *Microarray data normalization and transformation.* **Nature genetics**, 32 Suppl:496–501, December 2002.

[64] T Benaglia, D Chauveau, D R Hunter, and D Young. *mixtools: An R Package for Analyzing Finite Mixture Models.* **Journal of Statistical Software**, 32(6):1–29, 2009.

[65] R C Jansen, H Geerlings, A J van Oeveren, and R C van Schaik. *A Comment on Codominant Scoring of AFLP Markers.* **Genetics**, 158(2):925–926, 2001.

[66] H Westra, R C Jansen, R S N Fehrmann, *et al. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects.* **Bioinformatics** (Oxford, England), 27(15):2104–11, August 2011.

[67] O Loudet, S Chaillou, C Camilleri, D Bouchez, and F Daniel-Vedele. *Bay-0 × Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis.* **Theoretical and applied genetics**, 104(6-7):1173–1184, May 2002.

[68] Y Li, M A Swertz, G Vera, *et al. designGG: an R-package and web tool for the optimal design of genetical genomics experiments.* **BMC bioinformatics**, 10:188, January 2009.

[69] R V L Joosen, D Arends, L A J Willems, *et al. Visualizing the genetic landscape of Arabidopsis seed performance.* **Plant physiology**, 158(2):570–89, February 2012.

[70] D Arends, K J van der Velde, P Prins, *et al. xQTL workbench: a scalable web environment for multi-level* QTL *analysis.* **Bioinformatics** (Oxford, England), 28(7):1042–4, April 2012.

[71] L B Snoek, K J van der Velde, D Arends, *et al. WormQTL–public archive and analysis web portal for natural variation data in Caenorhabditis spp.* **Nucleic acids research**, 41(Database issue):D738–43, January 2013.

[72] R Alberts, P Terpstra, L V Bystrykh, G de Haan, and R C Jansen. *A statistical multiprobe model for analyzing cis and trans genes in genetical genomics experiments with short-oligonucleotide arrays.* **Genetics**, 171(3):1437–9, November 2005.

[73] R Alberts, P Terpstra, Y Li, *et al. Sequence polymorphisms cause many false cis eQTLs.* **PloS one**, 2(7):e622, January 2007.

[74] K W Broman and S Sen. *A Guide to* QTL *Mapping with R/qtl. Statistics for Biology and Health.* **Springer**, New York, NY, 2009.

[75] B S Yandell, T Mehta, S Banerjee, *et al. R/qtlbim: QTL with Bayesian Interval Mapping in experimental crosses.* **Bioinformatics** (Oxford, England), 23(5):641–3, March 2007.

[76] S Banerjee, B S Yandell, and N Yi. *Bayesian quantitative trait loci mapping for multiple traits.* **Genetics**, 179(4):2275–89, August 2008.

[77] R C Jansen. *Quantitative Trait Loci in Inbred Lines.* **Handbook of Statistical Genetics**, (589–622). **John Wiley & Sons, Ltd.**, 2007.

[78] D J Balding, M Bishop, and C Cannings. *Handbook of Statistical Genetics*. **John Wiley & Sons, Ltd.**, 2007.

[79] R C Jansen. *Controlling the type I and type II errors in mapping quantitative trait loci.* **Genetics**, 138(3):871–881, November 1994.

[80] J W van Ooijen, M P Boer, R C Jansen, and C Maliepaard. *Map*QTL 4.0, *Software for the Calculation of* QTL *Position on Genetic Maps*, 2002.

[81] J G de Mooij-van Malsen, H A van Lith, H Oppelaar, B Olivier, and M J H Kas. *Evidence for epigenetic interactions for loci on mouse chromosome 1 regulating open field activity.* **Behavior genetics**, 39(2):176–82, March 2009.

[82] M J W Jeuken, N W Zhang, L K McHale, *et al. Rin4 causes hybrid necrosis and racespecific resistance in an interspecific lettuce hybrid.* **The Plant cell**, 21(10):3368–78, October 2009.

[83] J Kitano, J A Ross, S Mori, *et al. A role for a neo-sex chromosome in stickleback speciation.* **Nature**, 461(7267):1079–83, October 2009.

[84] J Fu, M A Swertz, J J B Keurentjes, and R C Jansen. *MetaNetwork: a computational protocol for the genetic study of metabolic networks.* **Nature protocols**, 2(3):685–94, January 2007.

[85] L Tierney, A J Rossini, and N Li. *Snow: a parallel computing framework for the R system.* **Int. J. Parallel Program**, 37(1):78–90, 2009.

[86] Y Li, O A Alvarez, E W Gutteling, *et al. Mapping determinants of gene expression plasticity by genetical genomics in C. elegans.* **PLoS genetics**, 2(12):e222, December 2006.

[87] X Huang, J Schmitt, L Dorn, *et al. The earliest stages of adaptation in an experimental plant population: strong selection on QTLS for seed dormancy.* **Molecular ecology**, 19(7):1335–51, April 2010.

[88] W E Finch-Savage and G Leubner-Metzger. *Seed dormancy and the control of germination.* **The New phytologist**, 171(3):501–23, January 2006.

[89] J M Dechaine, G Gardner, and C Weinig. *Phytochromes differentially regulate seed germination responses to light quality and temperature cues during seed maturation.* **Plant, cell & environment**, 32(10):1297–309, October 2009.

[90] A L Elwell, D S Gronwall, N D Miller, E P Spalding, and T L D Brooks. *Separating parental environment from seed size effects on next generation growth and development in Arabidopsis.* **Plant, cell & environment**, 34(2):291–301, February 2011.

[91] L Rivero-Lepinckas, D Crist, and R Scholl. *Growth of plants and preservation of seeds.* **Methods in molecular biology** (Clifton, N.J.), 323:3–12, January 2006.

[92] O Loudet, V Gaudon, A Trubuil, and F Daniel-Vedele. *Quantitative trait loci controlling root growth and architecture in Arabidopsis thaliana confirmed by heterogeneous inbred family.* **Theoretical and applied genetics**, 110(4):742–53, February 2005.

[93] M Reymond, S Svistoonoff, O Loudet, L Nussaume, and T Desnos. *Identification of QTL controlling root growth response to phosphate starvation in Arabidopsis thaliana.* **Plant, cell & environment**, 29(1):115–25, January 2006.

[94] O Loudet, S Chaillou, A Krapp, and F Daniel-Vedele. *Quantitative trait loci analysis of water and anion contents in interaction with nitrogen availability in Arabidopsis thaliana.* **Genetics**, 163(2):711–722, February 2003.

[95] O Loudet, S Chaillou, P Merigout, J Talbotec, and F Daniel-Vedele. *Quantitative trait loci analysis of nitrogen use efficiency in Arabidopsis.* **Plant physiology**, 131(1):345–58, January 2003.

[96] Y Barrière, V Méchin, B Lefevre, and S Maltese. *QTLs for agronomic and cell wall traits in a maize RIL progeny derived from a cross between an old Minnesota13 line and a modern Iodent line.* **Theoretical and applied genetics**, 125(3):531–49, August 2012.

[97] F Calenge, V Saliba-Colombani, S Mahieu, *et al. Natural variation for carbohydrate content in Arabidopsis. Interaction with complex traits dissected by quantitative genetics.* **Plant physiology**, 141(4):1630–43, August 2006.

[98] O Loudet, V Saliba-Colombani, C Camilleri, *et al. Natural variation for sulfate content in Arabidopsis thaliana is highly controlled by APR2.* **Nature genetics**, 39(7):896–900, July 2007.

[99] C Diaz, V Saliba-Colombani, O Loudet, *et al. Leaf yellowing and anthocyanin accumulation are two genetically independent strategies in response to nitrogen limitation in Arabidopsis thaliana.* **Plant & cell physiology**, 47(1):74–83, January 2006.

[100] O Loudet, T P Michael, B T Burger, *et al. A zinc knuckle protein that negatively controls morning-specific growth in Arabidopsis thaliana.* **Proceedings of the National Academy of Sciences of the United States of America**, 105(44):17193–8, November 2008.

[101] P Meng, A Macquet, O Loudet, A Marion-Poll, and H M North. Analysis of natural allelic variation controlling *Arabidopsis thaliana* seed germinability in response to cold and dark: identification of three major quantitative trait loci. **Molecular plant**, 1(1):145–54, January 2008.

[102] L Bentsink, J Hanson, C J Hanhart, *et al. Natural variation for seed dormancy in Arabidopsis is regulated by additive genetic and molecular pathways.* **Proceedings of the National Academy of Sciences of the United States of America**, 107(9):4264–9, March 2010.

[103] N Tamura, T Yoshida, A Tanaka, *et al. Isolation and characterization of high temperature-resistant germination mutants of Arabidopsis thaliana.* **Plant & cell physiology**, 47(8):1081–94, August 2006.

[104] N Galpaz and M Reymond. *Natural variation in Arabidopsis thaliana revealed a genetic network controlling germination under salt stress.* **PloS one**, 5(12):e15198, January 2010.

[105] V Chinnusamy, B Stevenson, B Lee, and J Zhu. *Screening for gene regulation mutants by bioluminescence imaging.* **Science's STKE: signal transduction knowledge environment**, 2002(140):pl10, July 2002.

[106] V Chinnusamy, K Schumaker, and J Zhu. *Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants.* **Journal of experimental botany**, 55(395):225–36, January 2004.

[107] R Finkelstein, S L Gampala, and C D Rock. *Abscisic acid signaling in seeds and seedlings.* **The Plant cell**, 14 Suppl:S15–45, January 2002.

[108] L Xiong, K S Schumaker, and J Zhu. *Cell signaling during cold, drought, and salt stress.* **The Plant cell**, 14 Suppl:S165–83, January 2002.

[109] A Linkies, K Müller, K Morris, *et al. Ethylene interacts with abscisic acid to regulate endosperm rupture during germination: a comparative approach using Lepidium sativum and Arabidopsis thaliana.* **The Plant cell**, 21(12):3803–22, December 2009.

[110] G C K Chiang, D Barua, E M Kramer, R M Amasino, and K Donohue. *Major flowering time gene, flowering locus C, regulates seed germination in Arabidopsis thaliana.* **Proceedings of the National Academy of Sciences of the United States of America**, 106(28):11661–6, July 2009.

[111] C H Orsi and S D Tanksley. *Natural variation in an ABC transporter gene associated with seed size evolution in tomato species.* **PLoS genetics**, 5(1):e1000347, January 2009.

[112] C Alonso-Blanco, M G M Aarts, L Bentsink, *et al. What has natural variation taught us about plant development, physiology, and adaptation?* **The Plant cell**, 21(7):1877– 96, July 2009.

[113] M Vargas, F A van Eeuwijk, J Crossa, and J Ribaut. *Mapping QTLs and QTL × environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods.* **Theoretical and applied genetics**, 112(6):1009–23, April 2006.

[114] L Moreau, A Charcosset, and A Gallais. *Use of trial clustering to study QTL × environment effects for grain yield and related traits in maize.* **Theoretical and applied genetics**, 110(1):92–105, December 2004.

[115] M P Boer, D Wright, L Feng, *et al.* *A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize.* **Genetics**, 177(3):1801–13, November 2007.

[116] A Macquet, M Ralet, O Loudet, *et al.* *A naturally occurring mutation in an Arabidopsis accession affects a beta-D-galactosidase that increases the hydrophilic potential of rhamnogalacturonan I in seed mucilage.* **The Plant cell**, 19(12):3990–4006, December 2007.

[117] V Quesada, S García-Martínez, P Piqueras, M R Ponce, and J L Micol. *Genetic architecture of NaCl tolerance in Arabidopsis.* **Plant physiology**, 130(2):951–63, October 2002.

[118] C Alonso-Blanco, L Bentsink, C J Hanhart, H Blankestijn-de Vries, and M Koornneef. *Analysis of natural allelic variation at seed dormancy loci of Arabidopsis thaliana.* **Genetics**, 164(2):711–729, June 2003.

[119] E J M Clerkx, M E El-Lithy, E Vierling, *et al.* *Analysis of natural allelic variation of Arabidopsis seed germination and seed longevity traits between the accessions Landsberg erecta and Shakdara, using a new recombinant inbred line population.* **Plant physiology**, 135(1):432–43, May 2004.

[120] M P Laserna, R A Sánchez, and J F Botto. *Light-related loci controlling seed germination in Ler × Cvi and Bay-0 × Sha recombinant inbred-line populations of Arabidopsis thaliana.* **Annals of botany**, 102(4):631–42, October 2008.

[121] A J Vallejo, M J Yanovsky, and J F Botto. *Germination variation in Arabidopsis thaliana accessions under moderate osmotic and salt stresses.* **Annals of botany**, 106(5):833–42, November 2010.

[122] R V L Joosen, J Kodde, L A J Willems, *et al.* *GERMINATOR: a software package for high-throughput scoring and curve fitting of Arabidopsis seed germination.* **The Plant journal: for cell and molecular biology**, 62(1):148–59, April 2010.

[123] S Penfield, R C Meissner, D A Shoue, N C Carpita, and M W Bevan. *MYB61 is required for mucilage deposition and extrusion in the Arabidopsis seed coat.* **The Plant cell**, 13(12):2777–91, December 2001.

[124] A A Arsovski, G W Haughn, and T L Western. *Seed coat mucilage cells of Arabidopsis thaliana as a model for plant cell wall research.* **Plant signaling & behavior**, 5(7):796– 801, July 2010.

[125] M Koornneef, C J Hanhart, H W M Hilhorst, and C M Karssen. *In Vivo Inhibition of Seed Development and Reserve Protein Accumulation in Recombinants of Abscisic Acid Biosynthesis and Responsiveness Mutants in Arabidopsis thaliana.* **Plant physiology**, 90(2):463–9, June 1989.

[126] Z Ren, Z Zheng, V Chinnusamy, *et al.* *RAS1, a quantitative trait locus for salt tolerance and ABA sensitivity in Arabidopsis.* **Proceedings of the National Academy of Sciences of the United States of America**, 107(12):5669–74, March 2010.

[127] J Argyris, P Dahal, E Hayashi, D W Still, and K J Bradford. *Genetic variation for lettuce seed thermoinhibition is associated with temperature-sensitive expression of abscisic Acid, gibberellin, and ethylene biosynthesis, metabolism, and response genes.* **Plant physiology**, 148(2):926–47, October 2008.

[128] S Lee, J Kang, H Park, *et al. DREB2C interacts with ABF2, a bZIP protein regulating abscisic acid-responsive gene expression, and its overexpression affects abscisic acid sensitivity.* **Plant physiology**, 153(2):716–27, June 2010.

[129] L Hunt, M J Holdsworth, and J E Gray. *Nicotinamidase activity is important for germination.* **The Plant journal: for cell and molecular biology**, 51(3):341–51, August 2007.

[130] T Mitchell-Olds and J Schmitt. *Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis.* **Nature**, 441(7096):947–952, June 2006.

[131] J J B Keurentjes, J Fu, I R Terpstra, *et al. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci.* **Proceedings of the National Academy of Sciences of the United States of America**, 104(5):1708–13, January 2007.

[132] D J Kliebenstein, M A L West, H van Leeuwen, *et al. Identification of QTLs controlling gene expression networks defined a priori.* **BMC Bioinformatics**, 7:308, 2006.

[133] H C Rowe, B G Hansen, B A Halkier, and D J Kliebenstein. *Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome.* **The Plant cell**, 20(5):1199–216, May 2008.

[134] R Kooke and J J B Keurentjes. *Multi-dimensional regulation of metabolic networks shaping plant development and performance.* **Journal of experimental botany**, 63(9):3353–65, May 2012.

[135] E N Smith and L Kruglyak. *Gene-environment interaction in yeast gene expression.* **PLoS biology**, 6(4):e83, April 2008.

[136] A Gerrits, Y Li, B M Tesson, *et al. Expression quantitative trait loci are highly sensitive to cellular differentiation state.* **PLoS genetics**, 5(10):e1000692, October 2009.

[137] J Zhu, P Sova, Q Xu, *et al. Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation.* **PLoS biology**, 10(4):e1001301, January 2012.

[138] Y Li, R Breitling, and R C Jansen. *Generalizing genetical genomics: getting added value from environmental perturbation.* **Trends in genetics**, 24(10):518–24, October 2008.

[139] G A Churchill. *Fundamentals of experimental design for cDNA microarrays.* **Nature genetics**, 32 Suppl:490–5, December 2002.

[140] J M Peregrín-Alvarez, C Sanford, and J Parkinson. *The conservation and evolutionary modularity of metabolism.* **Genome biology**, 10(6):R63, January 2009.

[141] O Fiehn, J Kopka, P Dörmann, *et al. Metabolite profiling for plant functional genomics.* **Nature biotechnology**, 18(11):1157–61, November 2000.

[142] D J Kliebenstein, V M Lambrix, M Reichelt, J Gershenzon, and T Mitchell-Olds. *Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutaratedependent dioxygenases control glucosinolate biosynthesis in Arabidopsis.* **The Plant cell**, 13(3):681–93, March 2001.

[143] J J B Keurentjes, J Fu, R C H de Vos, *et al. The genetics of plant metabolism.* **Nature genetics**, 38(7):842–9, July 2006.

[144] E K F Chan, H C Rowe, J A Corwin, B Joseph, and D J Kliebenstein. *Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana.* **PLoS biology**, 9(8):e1001125, August 2011.

[145] C G de Oliveira Dal'Molin, L Quek, R W Palfreyman, S M Brumbley, and L K Nielsen. *AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis.* **Plant physiology**, 152(2):579–89, February 2010.

[146] U Roessner, C Wagner, J Kopka, R N Trethewey, and L Willmitzer. *Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry.* **The Plant journal: for cell and molecular biology**, 23(1):131–42, July 2000.

[147] J Lisec, R C Meyer, M Steinfath, *et al. Identification of metabolic and biomass QTL in Arabidopsis thaliana in a parallel analysis of RIL and IL populations.* **The Plant journal: for cell and molecular biology**, 53(6):960–72, March 2008.

[148] J D Bewley. *Seed Germination and Dormancy.* **The Plant cell**, 9(7):1055–1066, July 1997.

[149] X Shu, T Frank, Q Shu, and K Engel. *Metabolite profiling of germinating rice seeds.* **Journal of agricultural and food chemistry**, 56(24):11612–20, December 2008.

[150] J Dowdle, T Ishikawa, S Gatzek, S Rolinski, and N Smirnoff. *Two genes in Arabidopsis thaliana encoding GDP-L-galactose phosphorylase are required for ascorbate biosynthesis and seedling viability.* **The Plant journal: for cell and molecular biology**, 52(4):673–89, November 2007.

[151] A M Kinnersley and F J Turano. *Gamma Aminobutyric Acid (GABA) and Plant Responses to Stress.* **Critical Reviews in Plant Sciences**, 19(6):479–509, 2000.

[152] N Bouché and H Fromm. *GABA in plants: just a metabolite?* **Trends in plant science**, 9(3):110–5, March 2004.

[153] A A Kelly, A Quettier, E Shaw, and P J Eastmond. Seed storage oil mobilization is important but not essential for germination or seedling establishment in *Arabidopsis*. **Plant physiology**, 157(2):866–75, October 2011.

[154] A Fait, R Angelovici, H Less, *et al. Arabidopsis seed development and germination is associated with temporally distinct metabolic switches.* **Plant physiology**, 142(3):839–54, November 2006.

[155] R Angelovici, G Galili, A R Fernie, and A Fait. *Seed desiccation: a bridge between maturation and germination.* **Trends in plant science**, 15(4):211–8, April 2010.

[156] R C Meyer, M Steinfath, J Lisec, *et al. The metabolic signature related to high plant growth rate in Arabidopsis thaliana.* **Proceedings of the National Academy of Sciences of the United States of America**, 104(11):4759–64, March 2007.

[157] Y Barriere, A Laperche, L Barrot, *et al. QTL analysis of lignification and cell wall digestibility in the Bay-0 × Shahdara RIL progeny of Arabidopsis thaliana as a model system for forage plant.* **Plant science: an international journal of experimental plant biology**, 168(5):1235–1245, May 2005.

[158] A M Wentzell, H C Rowe, B G Hansen, *et al. Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways.* **PLoS genetics**, 3(9):1687–701, September 2007.

[159] D H Hobbs, J E Flintham, and M J Hills. *Genetic control of storage oil synthesis in seeds of Arabidopsis.* **Plant physiology**, 136(2):3341–9, October 2004.

[160] K Nakabayashi, M Okamoto, T Koshiba, Y Kamiya, and E Nambara. *Genome-wide profiling of stored mRNA in Arabidopsis thaliana seed germination: epigenetic and genetic regulation of transcription in seed.* **The Plant journal: for cell and molecular biology**, 41(5):697–709, March 2005.

[161] K A Howell, R Narsai, A Carroll, *et al. Mapping metabolic and transcript temporal switches during germination in rice highlights specific transcription factors and the role of RNA instability in the germination process.* **Plant physiology**, 149(2):961–80, March 2009.

[162] Y M Tikunov, S Laptenok, R D Hall, A Bovy, and R C H de Vos. *MSClust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data.* **Metabolomics: Official journal of the Metabolomic Society**, 8(4):714–718, August 2012.

[163] N Schauer, D Steinhauser, S Strelkov, *et al. GC-MS libraries for the rapid identification of metabolites in complex biological samples.* **FEBS letters**, 579(6):1332–7, February 2005.

[164] K Sheppard, J Yuan, M J Hohn, *et al. From one amino acid to another: tRNAdependent amino acid biosynthesis.* **Nucleic acids research**, 36(6):1813–25, April 2008.

[165] S Binder. *Branched-Chain Amino Acid Metabolism in Arabidopsis thaliana.* **The *Arabidopsis* book / American Society of Plant Biologists**, 8:e0137, January 2010.

[166] Y Fujiki, T Sato, M Ito, and A Watanabe. *Isolation and characterization of cDNA clones for the e1beta and E2 subunits of the branched-chain alpha-ketoacid dehydrogenase complex in Arabidopsis.* **The Journal of biological chemistry**, 275(8):6007–13, March 2000.

[167] C Alonso-Blanco, H Blankestijn-de Vries, C J Hanhart, and M Koornneef. *Natural allelic variation at seed size loci in relation to other life history traits of Arabidopsis thaliana.* **Proceedings of the National Academy of Sciences of the United States of America**, 96(8):4710–7, April 1999.

[168] R Palanivelu, L Brass, A F Edlund, and D Preuss. *Pollen tube growth and guidance is regulated by POP2, an Arabidopsis gene that controls GABA levels.* **Cell**, 114(1):47– 59, July 2003.

[169] A Fait, A N Nesi, R Angelovici, *et al. Targeted enhancement of glutamate-to-gamma-aminobutyrate conversion in Arabidopsis seeds affects carbon-nitrogen balance and storage reserves in a development-dependent manner.* **Plant physiology**, 157(3):1026– 42, November 2011.

[170] M R Tuinstra, G Ejeta, and P B Goldsbrough. *Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci.* **Theoretical and Applied Genetics**, 95(5-6):1005–1011, 1997.

[171] J J B Keurentjes, L Bentsink, C Alonso-Blanco, *et al. Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population.* **Genetics**, 175(2):891–905, February 2007.

[172] R C Jansen, B M Tesson, J Fu, Y Yang, and L M McIntyre. *Defining gene and QTL networks.* **Current opinion in plant biology**, 12(2):241–6, April 2009.

[173] P Prins, G Smant, and R C Jansen. *Genetical genomics for evolutionary studies.* **Methods in molecular biology** (Clifton, N.J.), 856:469–85, January 2012.

[174] Y Li, R Breitling, L B Snoek, *et al. Global genetic robustness of the alternative splicing machinery in Caenorhabditis elegans.* **Genetics**, 186(1):405–10, September 2010.

[175] R S Hageman, M S Leduc, R Korstanje, B Paigen, and G A Churchill. *A Bayesian framework for inference of the genotype-phenotype map for segregating populations.* **Genetics**, 187(4):1163–1170, April 2011.

[176] A Bureau, J Dupuis, B Hayward, K Falls, and P van Eerdewegh. *Mapping complex traits using Random Forests.* **BMC genetics**, 4 Suppl 1:S64, January 2003.

[177] L Rönnegård and W Valdar. *Detecting major genetic loci controlling phenotypic variability in experimental crosses.* **Genetics**, 188(2):435–47, June 2011.

[178] H Gilbert and P Le Roy. *Comparison of three multitrait methods for QTL detection.* **Genetics Selection Evolution**, 35(3):281–304, 2003.

[179] A de la Fuente. *From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases.* **Trends in genetics**, 26(7):326– 33, July 2010.

[180] P Langfelder and S Horvath. *WGCNA: an R package for weighted correlation network analysis.* **BMC Bioinformatics**, 9:559, 2008.

[181] B M Tesson, R Breitling, and R C Jansen. *DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules.* **BMC bioinformatics**, 11(1):497, January 2010.

[182] A Fukushima. *DiffCorr: an R package to analyze and visualize differential correlations in biological networks.* **Gene**, 518(1):209–14, April 2013.

[183] R R Sokal and J F Rohlf. *Significance tests in correlation.* **W. H. Freeman and Company**, 1995.

[184] E C Neto, M P Keller, A F Broman, *et al. Quantile-based permutation thresholds for quantitative trait loci hotspots.* **Genetics**, 191(4):1355–1365, August 2012.

[185] T A Knijnenburg, L F A Wessels, M J T Reinders, and I Shmulevich. *Fewer permutations, more accurate P-values.* **Bioinformatics** (Oxford, England), 25(12):i161–8, June 2009.

[186] J Zhu, M C Wiener, C Zhang, *et al. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations.* **PLoS computational biology**, 3(4):e69, April 2007.

[187] M Ashburner, C A Ball, J A Blake, *et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* **Nature genetics**, 25(1):25–9, May 2000.

[188] E E Schadt, J Lamb, X Yang, *et al. An integrative genomics approach to infer causal associations between gene expression and disease.* **Nature genetics**, 37(7):710–7, July 2005.

[189] M E Smoot, K Ono, J Ruscheinski, P Wang, and T Ideker. *Cytoscape 2.8: new features for data integration and network visualization.* **Bioinformatics** (Oxford, England), 27(3):431–2, February 2011.

[190] P Shannon, A Markiel, O Ozier, *et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks.* **Genome research**, 13(11):2498– 504, November 2003.

[191] S Kullback. *Information Theory and Statistics.* **Wiley**, 1959.

[192] I Csiszár and P C Shields. *Information Theory and Statistics: A Tutorial.* **Foundations and Trends® in Communications and Information Theory**, 1(4):417–528, December 2004.

[193] C D Brown, L M Mangravite, and B E Engelhardt. *Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs.* **PLoS genetics**, 9(8):e1003649, August 2013.

[194] B P Fairfax, S Makino, J Radhakrishnan, *et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles.* **Nature genetics**, 44(5):502–10, May 2012.

[195] J Fu, M G M Wolfs, P Deelen, *et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression.* **PLoS genetics**, 8(1):e1002431, January 2012.

[196] A Metspalu. *The Estonian Genome Project.* **Drug Development Research**, 62(2):97–101, June 2004.

[197] T Tanaka, J Shen, G R Abecasis, *et al. Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study.* **PLoS genetics**, 5(1):e1000338, January 2009.

[198] A Hofman, C M van Duijn, O H Franco, *et al. The Rotterdam Study: 2012 objectives and design update.* **European journal of epidemiology**, 26(8):657–86, August 2011.

[199] A Teumer, R Rawal, G Homuth, *et al. Genome-wide association study identifies four genetic loci associated with thyroid volume and goiter risk.* **American journal of human genetics**, 88(5):664–73, May 2011.

[200] M Inouye, K Silander, E Hamalainen, *et al. An immune response network associated with blood lipid levels.* **PLoS genetics**, 6(9):e1001113, September 2010.

[201] H Westra, M J Peters, T Esko, *et al. Systematic identification of trans eQTLs as putative drivers of known disease associations.* **Nature genetics**, September 2013.

[202] M C Whitlock. *Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach.* **Journal of evolutionary biology**, 18(5):1368– 73, September 2005.

[203] P C A Dubois, G Trynka, L Franke, *et al. Multiple common variants for celiac disease influencing immune gene expression.* **Nature genetics**, 42(4):295–302, April 2010.

[204] D V Zhernakova, E de Klerk, H Westra, *et al. DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts.* **PLoS genetics**, 9(6):e1003594, June 2013.

[205] T Lappalainen, M Sammeth, M R Friedländer, *et al. Transcriptome and genome sequencing uncovers functional variation in humans.* **Nature**, 501(7468):506–11, September 2013.

[206] M A Swertz, K J van der Velde, B M Tesson, R A Scheltema, D Arends, *et al.* XGAP: *a uniform and extensible data model and software platform for genotype and phenotype experiments.* **Genome biology**, 11(3):R27, January 2010.

[207] M A Swertz, E O de Brock, S A F T van Hijum, *et al. Molecular Genetics Information System (*MOLGENIS*): alternatives in developing local experimental genomics databases.* **Bioinformatics** (Oxford, England), 20(13):2075–83, September 2004.

[208] L D Stein. *Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges.* **Nature Reviews Genetics**, 9(9):678–88, September 2008.

[209] G A Thorisson, J Muilu, and A J Brookes. *Genotype-phenotype databases: challenges and solutions for the post-genomic era.* **Nature Reviews Genetics**, 10(1):9–18, January 2009.

[210] H H H Göring, J E Curran, M P Johnson, *et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes.* **Nature genetics**, 39(10):1208–16, October 2007.

[211] G A R Heap, G Trynka, R C Jansen, *et al. Complex nature of SNP genotype effects on gene expression in primary human leucocytes.* **BMC medical genomics**, 2:1, January 2009.

[212] R B Brem, J D Storey, J Whittle, and L Kruglyak. *Genetic interactions between polymorphisms that affect gene expression in yeast.* **Nature**, 436(7051):701–3, August 2005.

[213] L V Bystrykh, E Weersing, B Dontje, *et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'.* **Nature genetics**, 37(3):225–32, March 2005.

[214] A Ramos, M P Moisan, F Chaouloff, C Mormède, and P Mormède. *Identification of female-specific QTLs affecting an emotionality-related behavior in rats.* **Molecular psychiatry**, 4(5):453–62, September 1999.

[215] J Fu, J J B Keurentjes, H Bouwmeester, *et al. System-wide molecular evidence for phenotypic buffering in Arabidopsis.* **Nature genetics**, 41(2):166–7, February 2009.

[216] D S Fay. *Classical genetics goes high-tech.* **Nature methods**, 5(10):863–4, October 2008.

[217] R Ihaka and R Gentleman. *R: A language for data analysis and graphics.* **Journal of computational and graphical statistics**, 5(3), 299-314, 1996.

[218] R Alberts, G Vera, and R C Jansen. *affy*GG: *computational protocols for genetical genomics with Affymetrix arrays.* **Bioinformatics** (Oxford, England), 24(3):433–4, February 2008.

[219] D Smedley, M A Swertz, K Wolstencroft, *et al. Solutions for data integration in functional genomics: a critical assessment and case study.* **Briefings in bioinformatics**, 9(6):532–44, November 2008.

[220] C J Mungall and D B Emmert. *A Chado case study: an ontology-based modular schema for representing genome-associated biological information.* **Bioinformatics** (Oxford, England), 23(13):i337–46, July 2007.

[221] B D O'Connor, A Day, S Cain, *et al. GMODWeb: a web framework for the Generic Model Organism Database.* **Genome biology**, 9(6):R102, January 2008.

[222] L D Stein, C Mungall, S Shu, *et al. The generic genome browser: a building block for a model organism system database.* **Genome research**, 12(10):1599–610, October 2002.

[223] R C Gentleman, V J Carey, D M Bates, *et al. Bioconductor: open software development for computational biology and bioinformatics.* **Genome biology**, 5(10):R80, January 2004.

[224] M D Mailman, M Feolo, Y Jin, *et al. The NCBI dbGaP database of genotypes and phenotypes.* **Nature genetics**, 39(10):1181–6, October 2007.

[225] C Wu, H Huang, H Juan, and S Chen. *GeneNetwork: an interactive tool for reconstruction of genetic networks using microarray data.* **Bioinformatics** (Oxford, England), 20(18):3691–3, December 2004.

[226] H Zeng, L Luo, W Zhang, *et al. PlantQTL-GE: a database system for identifying candidate genes in rice and Arabidopsis by gene expression and QTL information.* **Nucleic acids research**, 35(Database issue):D879–82, January 2007.

[227] Z Hu, E R Fritz, and J M Reecy. *AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond.* **Nucleic acids research**, 35(Database issue):D604–9, January 2007.

[228] E J Chesler, L Lu, J Wang, R W Williams, and K F Manly. *WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior.* **Nature neuroscience**, 7(5):485–6, May 2004.

[229] J Yang, C Hu, H Hu, *et al. QTLNetwork: mapping and visualizing genetic architecture of complex traits in experimental populations.* **Bioinformatics** (Oxford, England), 24(5):721–3, March 2008.

[230] M A Swertz, M Dijkstra, T Adamusiak, ..., D Arends *et al. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button.* **BMC Bioinformatics**, 11 Suppl 1(Suppl 12):S12, 2010.

[231] J Wang, R W Williams, and K F Manly. *WebQTL: web-based complex trait analysis.* **Neuroinformatics**, 1(4):299–308, January 2003.

[232] S Purcell, B Neale, K Todd-Brown, *et al. PLINK: a tool set for whole-genome association and population-based linkage analyses.* **American journal of human genetics**, 81(3):559–75, September 2007.

[233] J Goecks, A Nekrutenko, and J Taylor. *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* **Genome biology**, 11(8):R86, January 2010.

[234] *Open Bioinformatics Foundation.* **Open Bioinformatics Foundation**, June 2010.

[235] M Leu, K Humphreys, I Surakka, *et al. NordicDB: a Nordic pool and portal for genome-wide control data.* **European journal of human genetics**, 18(12):1322– 6, December 2010.

[236] A Deursen and P Klint. *Little languages: little maintenance?* **Journal of Software Maintenance: Research and Practice**, 10(2):75–92, March 1998.

[237] M Fowler. *Patterns of Enterprise Application Architecture.* **Source**, 48:560, 2002.

[238] D Fredman, M Siegfried, Y P Yuan, *et al. HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources.* **Nucleic acids research**, 30(1):387–91, January 2002.

[239] A R Jones, M Miller, R Aebersold, *et al. The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics.* **Nature biotechnology**, 25(10):1127–33, October 2007.

[240] B Smith, M Ashburner, C Rosse, *et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* **Nature biotechnology**, 25(11):1251– 5, November 2007.

[241] A Brazma, M Krestyaninova, and U Sarkans. *Standards for systems biology.* **Nature Reviews Genetics**, 7(8):593–605, August 2006.

[242] S D M Brown, P Chambon, and M H de Angelis. *EMPReSS: standardized phenotype screens for functional annotation of the mouse genome.* **Nature genetics**, 37(11):1155, November 2005.

[243] V J Carey, M Morgan, S Falcon, R Lazarus, and R C Gentleman. *GGtools: analysis of genetics of gene expression in bioconductor.* **Bioinformatics** (Oxford, England), 23(4):522–3, February 2007.

[244] S V Bhave, C Hornbaker, T L Phang, *et al. The PhenoGen informatics website: tools for analyses of complex traits.* **BMC genetics**, 8:59, January 2007.

[245] C F Taylor, D Field, and S A Sansone. *Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project.* **Nature biotechnology**, 26(8):889–96, August 2008.

[246] T Oinn, M Addis, J Ferris, *et al. Taverna: a tool for the composition and enactment of bioinformatics workflows.* **Bioinformatics** (Oxford, England), 20(17):3045–54, November 2004.

[247] B E Gaertner and P C Phillips. *Caenorhabditis elegans as a platform for molecular quantitative genetics and the systems biology of natural variation.* **Genetics research**, 92(5-6):331–48, December 2010.

[248] J E Kammenga, P C Phillips, M de Bono, and A Doroszuk. *Beyond induced mutants: using worms to study natural variation in genetic pathways.* **Trends in genetics**, 24(4):178–85, April 2008.

[249] M F Palopoli, M V Rockman, A TinMaung, *et al. Molecular basis of the copulatory plug polymorphism in Caenorhabditis elegans.* **Nature**, 454(7207):1019–1022, 2008.

[250] J E Kammenga, A Doroszuk, J A G Riksen, *et al.* A *Caenorhabditis elegans wild type defies the temperature-size rule owing to a single nucleotide polymorphism in tra-3.* **PLoS genetics**, 3(3):e34, March 2007.

[251] M V Rockman, S S Skrovanek, and L Kruglyak. *Selection at linked sites shapes heritable phenotypic variation in C. elegans.* **Science** (New York, N.Y.), 330(6002):372– 6, October 2010.

[252] P T McGrath, M V Rockman, M Zimmer, *et al. Quantitative mapping of a digenic behavioral trait implicates globin variation in C. elegans sensory behaviors.* **Neuron**, 61(5):692–9, March 2009.

[253] K C Reddy, E C Andersen, L Kruglyak, and D H Kim. A *polymorphism in npr-1 is a behavioral determinant of pathogen susceptibility in C. elegans.* **Science** (New York, N.Y.), 323(5912):382–4, January 2009.

[254] A Doroszuk, L B Snoek, E Fradin, J A G Riksen, and J E Kammenga. A *genomewide library of CB4856/N2 introgression lines of Caenorhabditis elegans.* **Nucleic acids research**, 37(16):e110, September 2009.

[255] E W Gutteling, A Doroszuk, J A G Riksen, *et al. Environmental influence on the genetic correlations between life-history traits in Caenorhabditis elegans.* **Heredity**, 98(4):206–13, April 2007.

[256] A Viñuela, L B Snoek, J A G Riksen, and J E Kammenga. *Genome-wide gene expression regulation as a function of genotype and age in C. elegans.* **Genome research**, 20(7):929–37, July 2010.

[257] E C Andersen, J P Gerke, J A Shapiro, *et al. Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity.* **Nature genetics**, 44(3):285–90, March 2012.

[258] J A Ross, D C Koboldt, J E Staisch, *et al. Caenorhabditis briggsae recombinant inbred line genotypes reveal inter-strain incompatibility and the evolution of recombination.* **PLoS genetics**, 7(7):e1002174, July 2011.

[259] M A Swertz and R C Jansen. *Beyond standardization: dynamic software infrastructures for systems biology. Nature reviews.* **Genetics**, 8(3):235–43, March 2007.

[260] M Elvin, L B Snoek, M Frejno, *et al. A fitness assay for comparing RNAi effects across multiple C. elegans genotypes.* **BMC Genomics**, 12:510, 2011.

[261] A Viñuela, L B Snoek, J A G Riksen, and J E Kammenga. *Aging Uncouples Heritability and Expression-QTL in Caenorhabditis elegans.* **G3** (Bethesda, Md.), 2(5):597–605, May 2012.

[262] C Durrant, M A Swertz, R Alberts, D Arends, *et al. Bioinformatics tools and database resources for systems genetics analysis in mice–a short review and an evaluation of future needs.* **Briefings in bioinformatics**, 13(2):135–42, March 2012.

[263] B E Stranger, M S Forrest, and M Dunning. *Relative impact of nucleotide and copy number variation on gene expression phenotypes.* **Science**, 315(5813):848–853, February 2007.

[264] J S Bailey, L Grabowski-Boase, B M Steffy, *et al. Identification of quantitative trait loci for locomotor activation and anxiety using closely related inbred strains.* **Genes, Brain and Behavior**, 7(7):761–9, October 2008.

[265] W G Beamer, K L Shultz, G A Churchill, *et al. Quantitative trait loci for bone density in C57BL/6J and CAST/EiJ inbred mice.* **Mammalian genome: official journal of the International Mammalian Genome Society**, 10(11):1043–9, November 1999.

[266] R E Voorrips, G Gort, and B Vosman. *Genotype calling in tetraploid species from bi-allelic marker data using mixture models.* **BMC bioinformatics**, 12(1):172, January 2011.

[267] M M Ristić. *R programming for bioinformatics*, 2009.

[268] G Ostrouchov, W C Chen, D Schmidt, and P. Patel. *Programming with Big Data in R*, 2012.

[269] A Alexandrescu. *The D Programming Language*. **Addison-Wesley**, 2011.

[270] D Arends. *Using CTFE in D to speed up Sine and Cosine*, **www.dannyarends.nl**, 2012.

[271] N Derome, B Bougas, S M Rogers, *et al. Pervasive sex-linked effects on transcription regulation as revealed by expression quantitative trait loci mapping in lake whitefish species pairs (Coregonus sp., Salmonidae).* **Genetics**, 179(4):1903–17, August 2008.

[272] A van Nas, L Ingram-Drake, J S Sinsheimer, *et al. Expression quantitative trait loci: replication, tissue- and sex-specificity in mice.* **Genetics**, 185(3):1059–68, July 2010.

[273] D C Baumgart and W J Sandborn. *Crohn's disease.* **Lancet**, 380(9853):1590– 605, November 2012.

[274] RDC Team. *R: A language and environment for statistical computing.* **R foundation for Statistical Computing**, 1, 2005.

[275] M Kanehisa and S Goto. *KEGG: kyoto encyclopedia of genes and genomes.* **Nucleic acids research**, 28(1):27–30, January 2000.

[276] K J van der Velde, M de Haan, K Zych, D Arends *et al. WormQTLHD–a web database for linking human disease to natural variation data in C. elegans.* **Nucleic acids research**, 42(1):D794–801, January 2014.

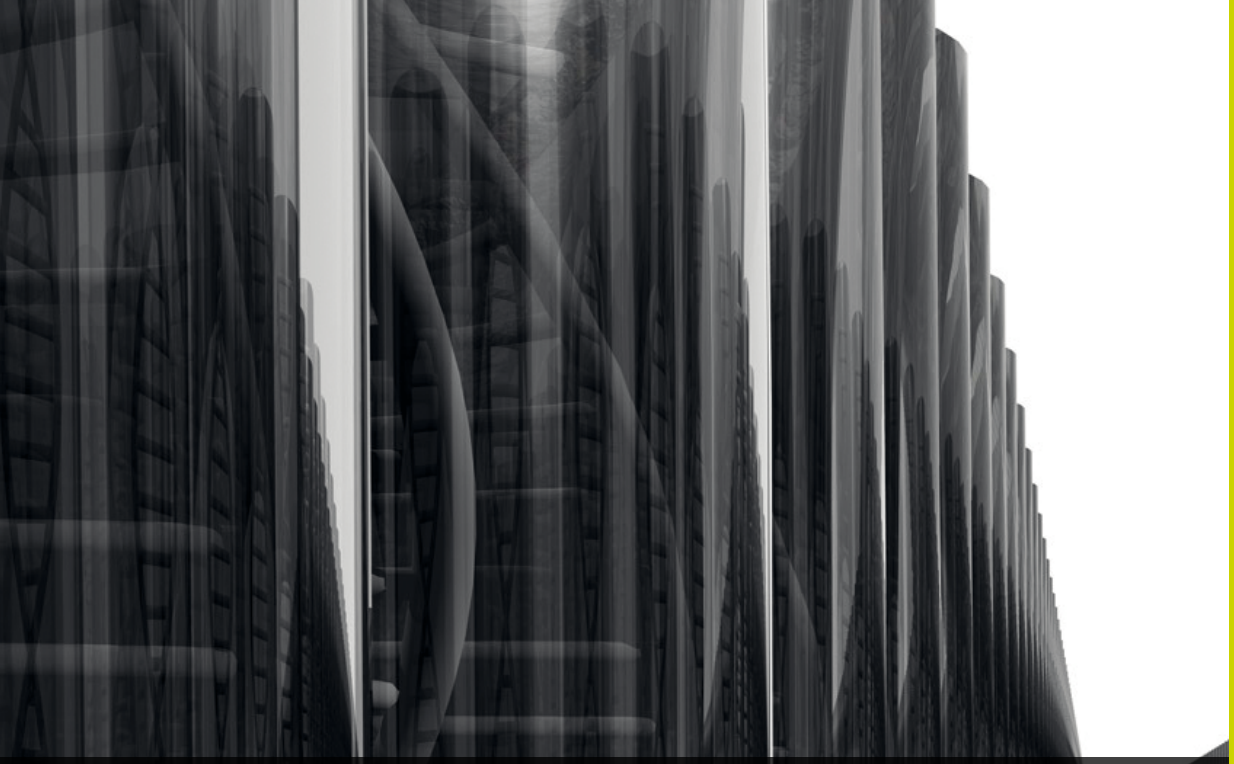
[277] G E Moore. *Cramming more components onto integrated circuits*, 1965.

[278] F Johannes and M Colomé-Tatché. *Quantitative epigenetics through epigenomic perturbation of isogenic lines*. **Genetics**, 188(1):215–27, May 2011.

[279] S Cortijo, R Wardenaar, M Colomé-Tatché, *et al. Mapping the Epigenetic Basis of Complex Traits*. **Science**, 343(6175), 1145-1148.

[280] Y Benjamini. *Simultaneous and selective inference: Current successes and future challenges*. **Biometrical journal**, 52(6):708–21, December 2010.

[281] D Huff. *How to Lie with Statistics [Paperback]*. **W. W. Norton & Company**; Reissue edition, 1993.



High-throughput computational methods and software for quantitative trait locus (QTL) mapping

In recent years many new technologies such as tiling arrays and high-throughput sequencing have come to play an important role in systems genetics research. For researchers it is of the utmost importance to understand how this affects their research. This work describes possible solutions to this 'Big Data' avalanche which has hit systems genetics.

This thesis describes the work carried out during the author's 4 year PHD project at the Groningen Bioinformatics Centre to develop smarter and more optimized algorithms such as Pheno2Geno and MQM, and to use a collaborative approach such as xQTL workbench to store and analyse high-throughput systems genetics data.

ISBN: 978-90-367-7209-9