

# **Generalized Genetical Genomics**

Advanced methods and applications

The work described in this thesis was carried out at the Groningen Bioinformatics Centre, University of Groningen, The Netherlands. The research was financially supported by Netherlands Organization for Scientific Research (NWO-86504001) and EU FP7 PANACEA 222936.

Printed by Drukkerij van Denderen B.V. Groningen, The Netherlands

ISBN 978-90-367-4594-9

RIJKSUNIVERSITEIT GRONINGEN

# **Generalized Genetical Genomics**

Advanced Methods and Applications

## **Proefschrift**

ter verkrijging van het doctoraat in de  
Wiskunde en Natuurwetenschappen  
aan de Rijksuniversiteit Groningen  
op gezag van de  
Rector Magnificus, dr. F. Zwarts,  
in het openbaar te verdedigen op  
maandag 8 november 2010  
om 14.45 uur

door

**Yang Li**

geboren op 4 oktober 1974  
te Hubei, China

Promotores

Prof. dr. R.C. Jansen  
Prof. dr. R. Breitling

Beoordelingscommissie

Prof. dr. G.A. Churchill  
Prof. dr. F.A. van Eeuwijk  
Prof. dr. E.C. Wit

*To my parents  
To Chengjian and Yuanxin*



---

# Contents

<b>Abstract</b>	<b>1</b>
<b>1 Introduction: Generalized Genetical Genomics</b>	<b>3</b>
1.1 Introduction . . . . .	3
1.2 Concepts of generalized genetical genomics . . . . .	3
1.2.1 Multifactorial experimentation . . . . .	3
1.2.2 Environments that matter . . . . .	5
1.2.3 Controlled environmental perturbation . . . . .	8
1.2.4 Challenges: controlling the uncontrolled . . . . .	13
1.2.5 Concluding remarks and future perspectives . . . . .	14
1.3 Thesis contribution and organization . . . . .	16
<b>2 designGG: a tool for designing genetical genomics experiments</b>	<b>19</b>
2.1 Introduction . . . . .	19
2.2 Implementation . . . . .	20
2.3 Results . . . . .	23
2.3.1 Web tool . . . . .	23
2.3.2 R package . . . . .	24
2.4 Expected results . . . . .	27
2.5 Conclusions . . . . .	28
2.6 Availability and requirements . . . . .	28
2.7 Author contributions . . . . .	28
2.8 Acknowledgments . . . . .	28

<b>3</b>	<b>Causal inference in genome-wide association and linkage studies</b>	<b>29</b>
3.1	Causal inference from genetic data . . . . .	29
3.2	Concerns about causal inference . . . . .	32
3.3	Restoring the potential of causal inference . . . . .	39
3.4	Concluding remarks . . . . .	40
3.5	Appendix . . . . .	41
3.6	Acknowledgments . . . . .	42
<b>4</b>	<b>Genetical genomics: spotlight on QTL hotspots</b>	<b>43</b>
4.1	Introduction . . . . .	43
4.2	Results/Discussion . . . . .	43
<b>5</b>	<b>Mapping determinants of gene expression plasticity by genetical genomics</b>	<b>49</b>
5.1	Introduction . . . . .	49
5.2	Results/Discussion . . . . .	50
5.2.1	Genome-wide detection of expression and plasticity QTLs . . . . .	51
5.2.2	Test for genetic assimilation . . . . .	52
5.2.3	Functional assessment of transband genes . . . . .	53
5.2.4	Estimating the rate of false-positives in <i>cis</i> -QTL effects . . . . .	55
5.2.5	Power analysis for plasticity QTLs . . . . .	56
5.3	Conclusion . . . . .	56
5.4	Materials and methods . . . . .	58
5.4.1	Genetical genomics experiment . . . . .	58
5.4.2	Data analysis . . . . .	62
5.5	Acknowledgments . . . . .	66
<b>6</b>	<b>eQTLs are Highly Sensitive to Cellular Differentiation State</b>	<b>67</b>
6.1	Introduction . . . . .	67
6.2	Results . . . . .	68
6.2.1	Genetic regulation of gene expression . . . . .	68
6.2.2	Cell-type-independent static eQTLs . . . . .	70
6.2.3	Cell-type-dependent dynamic eQTLs . . . . .	73
6.2.4	Detailed analysis of static and dynamic eQTLs . . . . .	74
6.3	Discussion . . . . .	76
6.4	Materials and methods . . . . .	77
6.4.1	Ethics statement . . . . .	77
6.4.2	Recombinant inbred mice . . . . .	77
6.4.3	Cell purification . . . . .	77



6.4.4	RNA isolation and Illumina microarrays . . . . .	78
6.4.5	Clustering of genes . . . . .	78
6.4.6	Full ANOVA model for eQTL mapping . . . . .	78
6.4.7	<i>Local</i> and <i>distant</i> eQTLs . . . . .	79
6.4.8	Classification of eQTLs . . . . .	79
6.4.9	Estimating the FDR for the main QTL effect . . . . .	79
6.4.10	Estimating the FDR for interaction QTL effect . . . . .	80
6.4.11	Detection of swapping eQTLs . . . . .	80
6.4.12	URLs . . . . .	81
6.5	Acknowledgments . . . . .	81
<b>7</b>	<b>Global genetic robustness of <i>C. elegans</i> alternative splicing machinery</b>	<b>83</b>
7.1	Introduction . . . . .	83
7.2	Results . . . . .	84
7.3	Conclusions . . . . .	87
7.4	Materials and methods . . . . .	89
7.4.1	Worm samples, genotyping and Affymetrix GeneChips . . . . .	89
7.4.2	Data analysis . . . . .	89
7.4.3	Classification of eQTLs pattern . . . . .	91
7.4.4	Permutation . . . . .	92
7.4.5	Deleted genes . . . . .	92
7.4.6	Comparison with previous experiment . . . . .	93
7.4.7	Power to detect quantitative changes in alternative splicing . . . . .	93
7.4.8	Supporting information . . . . .	93
7.5	Acknowledgments . . . . .	93
<b>8</b>	<b>Summarizing discussion</b>	<b>95</b>
8.1	Introduction . . . . .	95
8.2	Designing a genetic experiment for thousands of phenotypes . . . . .	97
8.2.1	Population . . . . .	98
8.2.2	Combining studies . . . . .	99
8.2.3	Sample assignment for molecular profiling . . . . .	100
8.3	Significance thresholds for eQTL detection . . . . .	101
8.4	Defining gene and QTL networks . . . . .	102
8.4.1	Correlation-based networks . . . . .	102
8.4.2	QTL-based networks . . . . .	104
8.4.3	Hotspots . . . . .	105
8.4.4	Non-genetic variation . . . . .	106

8.5 Conclusions . . . . .	109
<b>Bibliography</b>	<b>130</b>
<b>Samenvatting</b>	<b>131</b>
<b>Acknowledgements</b>	<b>133</b>
<b>Publications and Awards</b>	<b>135</b>
<b>Curriculum Vitae</b>	<b>139</b>

---

## Abstract

Generalized genetical genomics (GGG) is a systems genetics approach that combines the analysis of genetic variation with population-wide assessment of variation in molecular traits in multiple environments to identify genotype-by-environment interactions.

This thesis starts by introducing the generalized genetical genomics strategy (Chapter 1). Then, we present a newly developed software, designGG for designing optimal GGG experiments (Chapter 2).

Next, two important statistical issues relevant to GGG studies were addressed. We discussed the critical concerns on causal inference with genetic data. In addition, we examined the permutation method used for determining the significance of quantitative trait loci (QTL) hotspots in linkage and association studies (Chapter 3–4).

Furthermore, we applied the GGG strategy to three pilot studies: In the first of these, we showed that heritable differences in the plastic responses of gene expression are largely regulated in “trans”. In the second pilot study, we demonstrated that heritable differences in transcript abundance are highly sensitive to cellular differentiation stage. In the third study, we found that the alternative splicing machinery exhibits a general genetic robustness in *Caenorhabditis elegans* and that only a minor fraction of genes shows heritable variation in splicing forms and relative abundance. (Chapter 5–7).

Finally, we conclude by discussing various fundamental issues involved in data preprocessing, QTL mapping, result interpretation and network reconstruction and suggesting future directions yet to be explored in order to expand the reach of systems genetics (Chapter 8).



## Chapter 1

# Introduction: Generalized Genetical Genomics

---

## 1.1 Introduction

Genetical genomics (Jansen and Nap 2001) is a useful approach for studying the effect of genetic perturbations on biological systems at the molecular level. However, molecular networks depend on the environmental conditions and, thus, a comprehensive understanding of biological systems requires studying them across multiple environments. We propose a generalization of genetical genomics, which combines genetic and sensibly chosen environmental perturbations, to study the plasticity of molecular networks. This strategy forms a crucial step toward understanding why individuals respond differently to drugs, toxins, pathogens, nutrients and other environmental influences.

## 1.2 Concepts of generalized genetical genomics

### 1.2.1 Multifactorial experimentation

Many genetic and environmental factors can influence the functioning of a biological system. Understanding the interplay between these factors is essential for making progress in personalized medicine, epidemiology, environmental toxicology, breeding and many other fields where genetic and environmental variation matter. For example, the patient's response to drug treatment can depend strongly on his/her genotype, and gene regulatory networks that control important phenotypes, such as cellular proliferation rate, will depend not only on the genotype but also on the tissue or cell type under study.

This has important consequences for genetic strategies for studying molecular networks, including genetical genomics or expression genetics (Jansen and Nap 2001, Jansen 2003, Jansen and Nap 2004, Rockman and Kruglyak 2006, Haley and

de Koning 2007). Genetical genomics measures molecular phenotypes, such as gene expression, protein abundance or metabolite levels, in many genetically diverse individuals and uses classic quantitative trait mapping to identify the underlying regulatory influences (see Box 1.1 for a brief outline of this concept). However, the resulting molecular network will be specific for a single experimental condition (e.g. one species, one tissue type or one physical condition).

A generalized genetical genomics approach would study genetic and controlled environmental perturbations in combination. Like genetical genomics, such a generalized strategy will enable the mapping of quantitative trait loci (QTLs) underlying molecular traits of interest. Furthermore, it will also detect how QTL effects differ across multiple environments of interest and how the genotype influences the response to environmental changes (Figure 1.1a). This means that heritable differences in environmental plasticity can be explored on a genome-wide scale (Gibson and Weir 2005). Such experiments require careful experimental design, however, partly because many of the current studies that examine a single environment seem to operate at the limits of statistical feasibility.

**Box 1.1: Genetical genomics: a combination of genetic variation with genomic profiling to reconstruct molecular networks.**

In general, the strategy of genetical genomics contains the following steps.

(i) Select or create a population of genetically different individuals showing a relevant phenotypic variation in the environment of interest. Experimental populations (e.g. backcrosses, F2 populations, recombinant inbred lines, doubled haploids) and natural populations (germplasm collections, cell lines, pedigrees, case-controls, trios, twins) can be used.

(ii) Use molecular markers to genotype the individuals throughout the genome.

(iii) Determine the molecular profile, such as transcript, protein or metabolite abundance, of each individual in the population. A variety of molecular levels can be studied, as detailed in Table 1.1

We recommend not only studying large numbers of genetically different individuals (more is always better), but also using their marker genotype data to intelligently select and distribute individuals within and across environments. This should maximize the power and resolution of QTL mapping for one or more regions of special interest, such as a previously detected phenotypic QTL or across the entire genome.

**Table 1.1:** Molecular profiling technologies

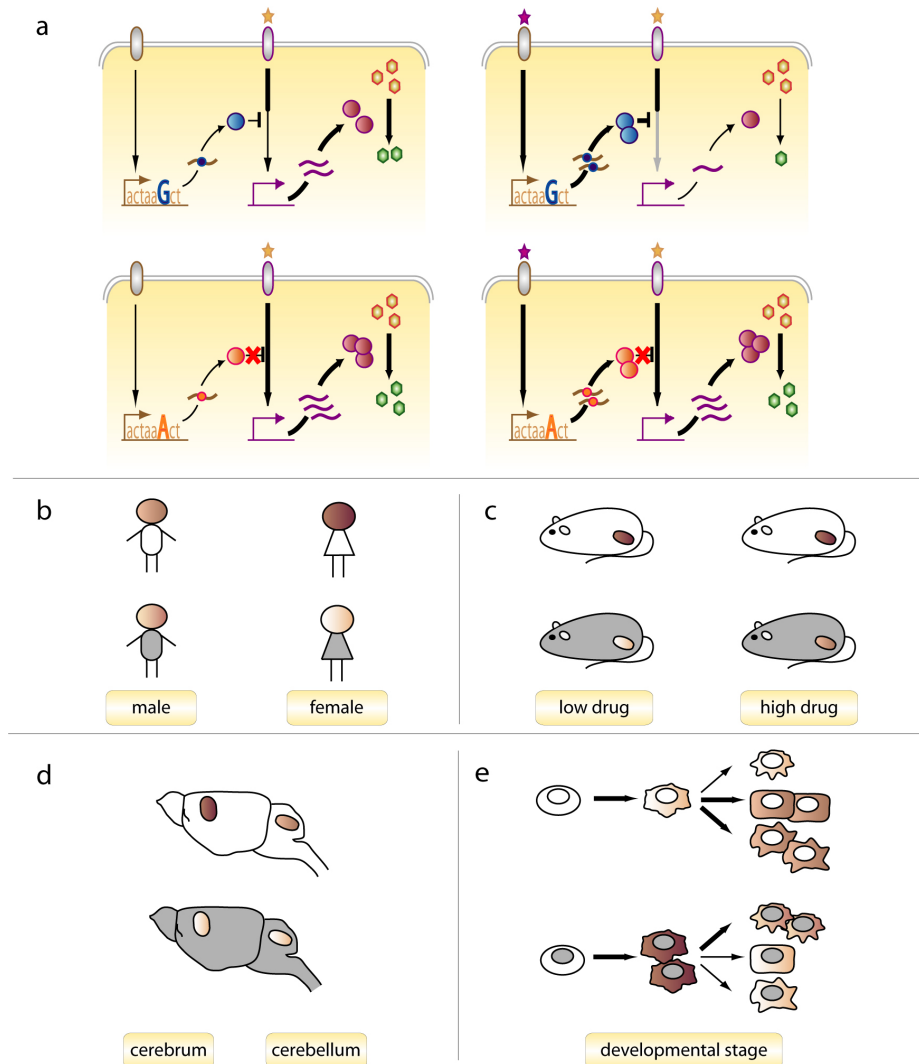
Genome	Microarray-based fingerprinting using thousands to millions of molecular markers (Blow 2007)
Transcriptome	Microarray-based profiling of transcript abundance using tens of thousands of probes (Hoheisel 2006)
Proteome	Gel or mass spectrometry based profiling of protein abundance of thousands of proteins (Cox and Mann 2007)
Metabolome	Untargeted mass spectrometry based profiling of metabolite of thousands of metabolites (Breitling <i>et al.</i> 2008b)
Kinome	Microarray-based profiling of phosphorylation for hundreds of kinase enzymes (Ptacek <i>et al.</i> 2005)
Epigenome	Chromatin immunoprecipitation (ChIP-chip) assays based profiling of thousands of DNA methylation and chromatin modification pattern (Buck and Lieb 2004)

### 1.2.2 Environments that matter

Environmental factors of interest for genetical genomics range from levels of drugs or toxic compounds, to differences in the social, agricultural or ecological setting, and to cell and tissue types, but they can also include sex differences or genetic background. All of these factors can be varied to reveal the sensitivity of the molecular network or to demonstrate its stability and robustness in the face of internal and external perturbation.

A generalized approach to genetical genomics can therefore come in many different flavors. For example, the presence or absence of the Y chromosome can be considered an (internal) environmental perturbation. When performing a molecular profiling study in humans (Figure 1.1b), one might wonder whether it is wise to include only one sex to reduce unnecessary biological variation or to split the experiment equally across the sexes (as doubling the study size is usually not an option). If the ultimate aim is a better understanding of a sex-specific trait, the relevant population should be studied. If, however, general conclusions about human biology are aimed for, both sexes need to be considered; otherwise one runs the risk of missing important trait-by-sex interactions. Sex can be included in the analysis of variance as an additional factor, at no statistical cost, and this is standard practice in mammalian QTL studies (Solberg *et al.* 2004). The effects that are shared between sexes will be detected, as will those that are different, potentially indicating a need for further in-depth study.

Drug treatment is another example of an (external) environmental perturbation. Different genotypes respond differently to drugs or toxins, and understanding the molecular details of these differences (pharmacogenomics) is becoming in-



**Figure 1.1:** Examples of quantitative trait loci-by-environment (QTL×E) interaction. (a) One of many possible molecular mechanisms underlying QTL×E and (b–e) different QTL×E cases. The causal gene underlying the QTL has sequence variants *actaaGct* and *actaaAct* in (a) and a white and grey variant in (b–e). All panels show a similar picture of a QTL that consistently appears in multiple environments, while its effect is modulated by environment. The different levels of molecular trait are indicated by different shades of brown in (b–e).

(a) QTL×E in a molecular circuitry. A single nucleotide polymorphism (SNP) in the coding region of a gene can change its function: for example, the *actaaGct* allele encodes a functional inhibitor protein, whereas the other *actaaAct* allele encodes a nonfunctional protein. Therefore, individuals carrying the *actaaGct* allele have lower abundance for downstream transcripts, enzymes and metabolites than individuals carrying the *actaaAct* allele: the QTL can be traced back to the SNP. Only environmental up- or downregulation of the *actaaGct* allele has functional consequences. As a result, the QTL effect is stronger when the environmental signal is present than when it is absent: the QTL interacts with environment.



**Figure 1.1: (continued)**(b)  $QTL \times E$  in molecular profiling. Even when they share a large part of their genetic make-up, male and female organisms can differ in their heritable molecular variation. Here, the QTL effect is larger in females than in males. If the studied trait is a biomarker for disease susceptibility and is used to inform medical treatment strategies, studying both sexes becomes essential.

(c)  $QTL \times E$  in a drug dosage experiment in mice. Biological systems are sensitive to external triggers such as drugs or toxins. One genotype has a molecular circuitry that makes it responsive to a drug, so that it already responds at a low drug dosage, whereas the other is less sensitive and requires high levels of the drug (but might be less sensitive to adverse effects). Detecting this kind of  $QTL \times E$  interactions is one of the major aims of pharmacogenomics.

(d)  $QTL \times E$  across tissues. Different tissues, even within a single organ such as the brain, carry out completely different tasks and will have different molecular circuitries. Studying them for  $QTL \times E$  interactions in a single experiment is usually not advisable because many genes will only be expressed in particular tissues. However, sometimes, as in this example, it might be particularly informative to identify those QTLs that are shared across several tissues that together contribute to a phenotype of interest, for example, a behavioral trait controlled by the integrated action of different brain areas.

(e)  $QTL \times E$  during the proliferation and differentiation of cells. This panel shows gene expression dynamics during development. The cells proliferate and differentiate to specialized cells, which provide different internal environments. Accordingly, their molecular circuitry has to change, which requires a tightly regulated interaction between genes and development stage. In many biological studies, it is not initially known if a QTL of interest is modulated across stages or restricted to a particular stage of differentiation, so collecting samples from different related stages will provide valuable insights. During the intermediate differentiation stage, there is a strong heritable difference in the expression of a master regulator gene, which leads to a redirection toward different final cell fates. At earlier and later stages, the expression QTL is much less prominent. In this example, the QTL affects the gene expression and the relative numbers of cells.

creasingly important for providing the basis for personalized drug development (Figure 1.1c). More generally, genotype-by-environment interactions with complex nongenetic factors, such as lifestyle, have to be considered in any human genetical genomics study.

However, the cellular environment also changes in a more subtle way without external intervention. For instance, different tissues have widely different functions and molecular profiles. One would expect the underlying network structures to be distinct. Each tissue will have its own susceptibility to genetic polymorphisms. A mutation in an oncogene could, for example, lead to upregulation of cell proliferation genes in one tissue but not in another. A generalized genetical genomics approach can determine how variable such heritable differences are across several tissues (Figure 1.1d).

Furthermore, molecular networks will also change along differentiation trajectories, and different genotypes will show different dynamics of molecular traits during development (Figure 1.1e). One QTL might control stem cell genes during initial lineage commitment, whereas another QTL might control the same set of genes during terminal differentiation.

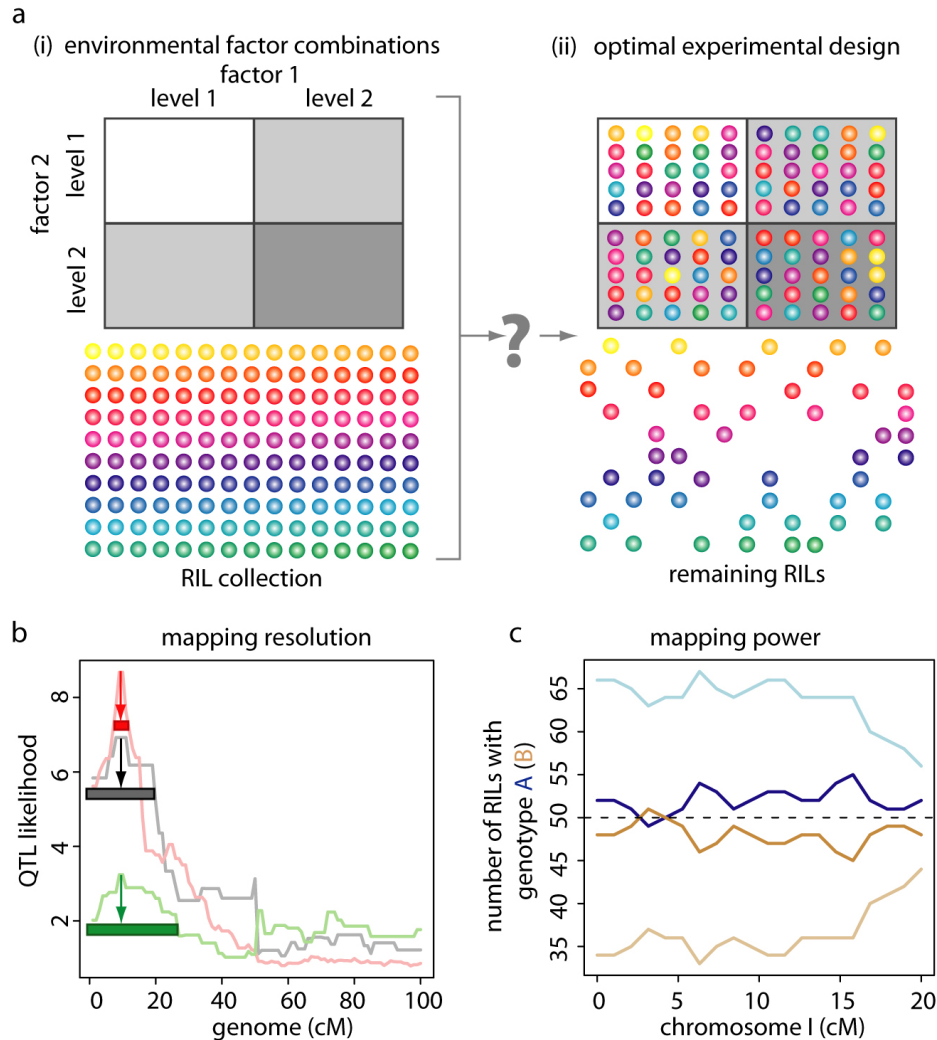
In all above studies, a wide variety of molecular mechanisms can cause genotype-by-environment interactions that can be studied using genetical genomics.

### 1.2.3 Controlled environmental perturbation

What would be the best strategy to design a multifactorial perturbation experiment for exploring the interaction between genetic and environmental factors? In some cases, it is possible to replicate genotypes across environments (e.g. using different mice from the same recombinant inbred strain); in other cases, this is not an option (e.g. exposing a human subject to several different environments at the same time). Traditionally, understanding the effect of QTLs across environments has relied on the first option. This approach was used in some of the first genetical genomics studies that examined environmental variation of gene expression QTLs (Li *et al.* 2006b, Smith and Kruglyak 2008). However, as we will show, there are important advantages to be gained by not replicating the same genotypes, but rather increasing the genetic diversity of the sample.

For example, suppose that we want to measure gene expression across four environments (as shown in Box 1.2 and Figure 1.2a ). We have access to 150 different recombinant inbred lines (RILs), but we can only afford to perform 100 microarrays. An intuitive way to perform the generalized experiment would be to study only 25 different RILs for all four environments, leaving 125 RILs unused. The alternative would be to allocate 100 different RILs to the four environments evenly, measuring 25 different RILs in each condition, keeping the microarray cost unchanged and leaving 50 RILs unused. But which design will produce the best outcome? A few considerations from classical experimental design theory show that, in most cases, the latter design is to be recommended. Moreover, we can do much better than simply choosing and allocating RILs at random.

A careful experimental design is particularly important if the resulting data are to be maximally informative (Fu and Jansen 2006, Churchill 2002, Yang and Speed 2002). What is the best strategy to obtain an optimal allocation of genetically different samples to different environments (and dyes and arrays)? The aim is to achieve the most accurate estimate of the quantitative trait loci (QTL) effects and QTL-by-environment interaction effects of interest either in one or more regions of



**Figure 1.2:** Designing a genetical genomics experiment with multiple environmental factors. (a) The optimal allocation of individuals with different genotypes to environmental conditions is a challenge in genetical genomic experiments. (i) We show 150 recombinant inbred lines (RILs) as balls, with the colors referring (in a simplified way) to genotypes: the more similar the colors the more similar the genetic fingerprint of two RILs. Each cell in the table represents a combination of different levels of two environmental factors (indicated by different levels of grey). (ii) The proposed strategy is to allocate 100 different RILs into four environments evenly, measuring 25 different RILs in each condition, keeping the microarray effort unchanged and leaving 50 RILs unused. (b) Simulation results comparing QTL resolution for three different strategies: using 25 lines in a single condition (green); using the same 25 lines in each of the four conditions (black) and using a total of 100 lines, 25 different ones in each condition (red). A single QTL at the 10th marker was simulated. The support intervals (1.5 drop-off) are indicated by the bars, showing that using 100 different lines dramatically improves the mapping resolution.

**Figure 1.2: (continued)**(c) Genotype comparison for lines based on random selection (depicted by the lighter shades of blue and brown) and our proposed design selection (shown by the darker colors). The brown and blue colors represent the two different genotypes. One hundred fifty individuals were simulated with segregation distortion (genotype A:genotype B = 2:1) caused by a locus at 1 cM on chromosome I, leading to a strongly unbalanced genotype difference that impairs QTL mapping in this region. Computer-assisted optimal selection of individuals focusing on the distorted region removes the genotype imbalance within and across environments, leading to improved mapping power.

special interest, such as a previously detected phenotypic QTL, or across the entire genome. We note that by minimizing the sum of the variances of the parameter estimates of interest (A-optimality (Kerr and Churchill 2001, Wit *et al.* 2005a)) using an optimization algorithm, such as simulated annealing (Wit and McClure 2004, Kirkpatrick *et al.* 1983), an optimal allocation can be found (Fu and Jansen 2006). In the optimization, the experimenter can, of course, give less weight to parameters of lesser interest, which will then be estimated with lower accuracy. For example, if the emphasis is on one or more genome regions of special interest, parameters for the markers in these regions can be given full weight in the optimization algorithm, whereas parameters for other markers can be given lesser or even zero weight. As a result, mapping resolution can improve (Figure 1.2b) and the power for finding QTLs can be increased (Figure 1.2c).

We suggest starting with a random initial allocation of samples to environments that can then be improved step-by-step by re-allocating samples (or sample pairs) from one environment to another or by replacing a sample by an unused sample (e.g. choosing 100 RILs to be profiled from a sample size of 150 RILs). A prototype web tool implementing the optimization algorithm for a wide variety of experimental situations is available online at <http://gbic.biol.rug.nl/designGG> to highlight the design issues.

First, the resolution and power of QTL mapping depends on the number of genetically different samples in linkage and association studies. From this point of view, it would be wise to include as many genetically different samples as possible and not to replicate them across environments, because more recombination events will be observed and rare alleles are more likely to be present in the samples (Darvasi and Soller 1997). Figure 1.2b compares the resolution available from different design strategies: it is obvious that maximizing the genetic diversity leads to the sharpest QTL peak with tightest support interval and thus implicitly the most specific list of candidate regulators.

**Box 1.2: Generate your own generalized genetical genomics experiment.**

Let us assume there are 150 recombinant inbred lines (RILs) available and 100 single-color arrays can be used to measure the genome-wide expression level (Figure 1.2a), and that there are two different environmental factors, such as drug treatment (factor 1) and pathogen exposure (factor 2). Each factor has two different levels: different amounts of drug and low versus high pathogen exposure.

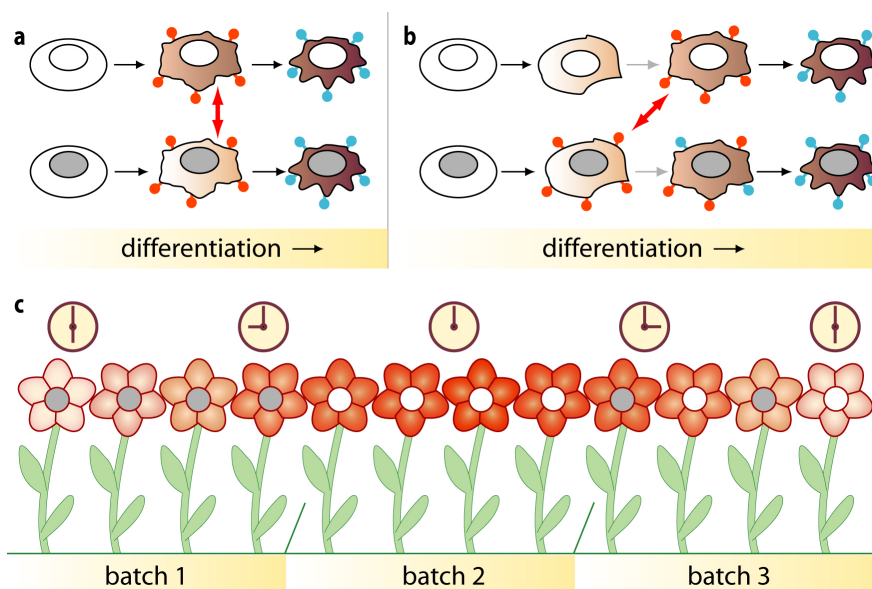
A careful experimental design is particularly important if the resulting data are to be maximally informative (Fu and Jansen 2006, Churchill 2002, Yang and Speed 2002). What is the best strategy to obtain an optimal allocation of genetically different samples to different environments (and dyes and arrays)? The aim is to achieve the most accurate estimate of the quantitative trait loci (QTL) effects and QTL-by-environment interaction effects of interest either in one or more regions of special interest, such as a previously detected phenotypic QTL, or across the entire genome. We note that by minimizing the sum of the variances of the parameter estimates of interest (A-optimality (Kerr and Churchill 2001, Wit *et al.* 2005a)) using an optimization algorithm, such as simulated annealing (Wit and McClure 2004, Kirkpatrick *et al.* 1983), an optimal allocation can be found (Fu and Jansen 2006). In the optimization, the experimenter can, of course, give less weight to parameters of lesser interest, which will then be estimated with lower accuracy. For example, if the emphasis is on one or more genome regions of special interest, parameters for the markers in these regions can be given full weight in the optimization algorithm, whereas parameters for other markers can be given lesser or even zero weight. As a result, mapping resolution can improve (Figure 1.2b) and the power for finding QTLs can be increased (Figure 1.2c).

We suggest starting with a random initial allocation of samples to environments that can then be improved step-by-step by re-allocating samples (or sample pairs) from one environment to another or by replacing a sample by an unused sample (e.g. choosing 100 RILs to be profiled from a sample size of 150 RILs). A prototype web tool implementing the optimization algorithm for a wide variety of experimental situations is available online at <http://gbic.biol.rug.nl/designGG> to highlight the design issues.

Second, if more genetically different individuals are used, it is less probable that two genetic unlinked loci will be confounded by chance.

Third, a proper statistical analysis would allow QTLs to be detected with almost the same power in an experiment with 100 RILs in a constant environment as in an experiment with 100 RILs across four environments, if QTLs are modestly modulated across the environments; the statistical model should include QTL-by-environment terms to account for the modulation of QTL effect across environments (Boer *et al.* 2007). Of course, the conditions to be studied will have to be chosen in a prudent fashion, because there is little to be gained from extreme perturbations.

Finally, in a population of RILs, the QTL effect at a particular genome location can be estimated most accurately if 50% of the profiled individuals are homozygous for one allele and the other 50% are homozygous for another allele at that locus. This suggests that genotype information should be used to select 100 RILs from the pool of 150 RILs and to allocate them across treatments. Then it will be possible to



**Figure 1.3:** Two examples of confounding environmental factors in generalized genetical genomics experiments. (a,b) The challenge of synchronization. In this hypothetical example of a multi-condition genetical genomics experiment, stem cells (white) differentiate into daughter cells continuously during life. Specific surface proteins (red and blue lollipop shapes) can be used to purify each of three subsequent differentiation stages (starting, intermediate and final). At each stage, gene expression levels are measured in a genetically diverse population (two alleles indicated by grey and white nuclei) and used for expression quantitative trait loci (eQTL) mapping to investigate genetic differences in the differentiation process. According to the expressed surface marker, a group of differentiation-related genes shows differential expression in the intermediate stage (as indicated by different shades of brown in the cytosol) in a set of genetically different lines and the variation can be mapped to a certain eQTL (grey versus white allele).

(a) A first and exciting explanation would be that this eQTL is a master regulator of the differentiation process. In this case, the two genotypes reach the same final state but through different differentiation trajectories.

(b) A second, and probably less exciting, interpretation would be that this eQTL only affects the expression timing of the surface marker used for cell staging. In this case, the two genotypes not only reach the same final state but also follow the same differentiation trajectory, and the observed differential expression between two genotypes actually results from sampling at two different stages (as the red double-headed arrow indicates). Thus, the eQTL does not influence the differentiation process at all and is clearly not a master regulator.

(c) The challenge of batching. In this hypothetical example, a group of crossed plant lines with different genotypes (as shown by grey versus white flower centers) are collected, and transcript abundance is measured (as indicated by the gradient of red flowers petals). It is clear that genotype correlates with gene expression level, and this can be mapped as an eQTL, indicating that a specific genetic locus plays an important role in deciding the gene expression

**Figure 1.3: (continued) level.** However, if on closer examination, the samples were collected at different times of the day and, accidentally, samples with one genotype at a particular locus were more common at early and very late time points, this unintentional correlation of genotype with an uncontrolled external factor (such as daylight) would lead to a spurious mapping result for those genes that vary in expression in a diurnal fashion. (The observed expression difference between genotypes is confounded by the effect of daytime on expression). In many real life cases, the uncontrolled factor is much less obvious than in this example and will therefore be very hard to detect with sometimes dramatic consequences (see Ref. (Alberts *et al.* 2005) for an example of how spurious linkage between batch and certain genome regions can lead to ghost regulators being detected). Therefore, careful randomization and statistical consideration of batch effects is essential (Akey *et al.* 2007).

maximize the number of informative genotype differences at one or more regions of particular interest, such as a previously detected phenotypic QTL, or to maximize the number of informative genotype contrasts across the entire genome. Figure 1.2c compares random and optimal selection. One can see that it is possible to achieve an almost balanced genotype ratio even in unfavorable circumstances.

However, optimally allocating the different samples (or distant pairs/triples/ $x$ -tuples of samples) to different environments is not straightforward. Allocating the samples at the same time to multiple dyes and arrays might further complicate the task. Figure 1.2 presents a computational tool to overcome these challenges.

The concept of allocating samples to optimize the power of an experiment is related to the 'distant pair' design (Fu and Jansen 2006). They allocated genetically distant individuals to two-color microarrays to maximize the number of informative genetic contrasts, in a standard genetical genomics experiment without environmental perturbation. Consequently this method will not result in a powerful design for generalized genetical genomics. This is evident in a study on gene expression plasticity in *Caenorhabditis elegans* (Li *et al.* 2006b), where 80 available RILs were assigned to arrays by distant pairing, and the same set of genotyped lines was reused in two environmental conditions. A new experiment using 160 selected RILs would have better mapping power and resolution and make full use of the total population of not,  $\sim 1000$  recombinant inbred worm lines that are now available.

#### 1.2.4 Challenges: controlling the uncontrolled

Even if an optimal experiment design has been achieved, uncontrolled factors, such as the time and the stage at which the biological sample is collected, will still in-

fluence the interpretation of generalized genetical genomics studies. For example, if samples are collected at the same developmental stage using a morphological or molecular selection marker, any unanticipated interaction between genotype and the selection marker can lead to spurious results (Figure 1.3a,b). To overcome this, one can include time or stage of sampling in the experimental design as an additional factor to check for potential interaction patterns. Other hidden factors can influence the molecular profiling itself. For example, if the samples were profiled environment-by-environment, any unanticipated interaction between the genotype and other uncontrolled factors changing during the profiling could also lead to spurious results (Figure 1.3c). Here one might want to include the order of profiling in the experimental design as an additional factor at multiple levels and use complete or incomplete block designs to eliminate any imbalance between factors of interest and the blocks (e.g. based on block-wise profiling of ten samples). In addition, methods such as surrogate variable analysis (Leek and Storey 2007) can be applied for detecting such uncontrolled environmental or genetic factors in an ANOVA model. As in the case for controlled environmental perturbations, including uncontrolled factors in the QTL model will not lead to a significant loss of QTL mapping power but will result in more robust interpretations.

### 1.2.5 Concluding remarks and future perspectives

Many of the most important properties of a living system depend on the interplay between genetic and environmental variation. Generalized genetical genomics is a powerful strategy to elucidate that interplay through modeling of quantitative trait loci (QTL) and QTL-by-environment interactions underlying biomolecular trait variation. However, there are some caveats: first, if a gene of interest is expressed only in one environment, all the samples in other environments will not be informative for QTL mapping, and distributing samples across multiple environments would reduce resolution and power because effective sample sizes for the gene(s) of interest would drop to unacceptably small numbers. Such studies have, however, been valuable as an initial proof of principle, showing that some QTLs are surprisingly robust, for example, those in different tissues (Bystrykh *et al.* 2005), whereas others change even in slightly different environments (Li *et al.* 2006b). Second, our illustrations in Figure 1.1 and Figure 1.2 are two-category or  $2 \times 2$  tables, and enthusiastic experimentalists might be tempted to try  $6 \times 6$  comparisons or even more complex studies. However, we would not recommend these approaches because the large number of QTL-by-environment parameters would diffuse the information on the QTL. Third, it is still a point of contention whether the standing genetic



depth in our study populations is large enough to cause variation in the traits of interest: mutations affecting essential molecular traits might only occur at extremely low population frequencies and are therefore likely to be absent in the study population. Finally, it remains to be proven whether detectable heritable variation in molecular traits (particularly in gene expression) actually has biological relevance for the major physiological properties of an organism.

Given these concerns, one might claim that one should not complicate QTL mapping (and genetical genomics) by including more environments, but the counterpart of this assertion is that, in the presence of a QTL-by-environment interaction, any inference about a QTL's main effect in a single-condition experiment will be confounded with the interaction effects. This can lead to serious mistakes in drug dosage and toxicity evaluation, for instance. To overcome the current limitations of QTL mapping, it will be necessary to use larger populations in a suitable experimental design, combine measurements at diverse biomolecular levels as discussed in Box 1.1, and integrate measurements from multiple populations of the same or different type (Li *et al.* 2005); all of these approaches essentially correspond to an experiment with controlled and uncontrolled environmental perturbations. The minimum number of samples required differs between traits (some traits might require a more in-depth study than others), types of population (natural versus experimental), between organisms (experiments on yeast require hundreds of samples, whereas studies on humans can need thousands) and between types of molecular data (transcript data perhaps being more noisy than metabolite data, or vice versa). Applying classical experimental design theory as outlined in Figure 1.2 will help to obtain the maximum amount of information within realistic constraints on study size. Here we have argued that genotyped individuals can be "intelligently" distributed across multiple environments and that a large population of genotyped individuals can be a useful resource from which to select a subset of genetically most dissimilar samples. The same concepts can also be applied to experiments in natural populations, which are a particularly appealing target for studying gene-by-environment interaction (Gibson and Weir 2005). For example, rather than replicating individuals across environments, it would be interesting to use independent sets of individuals in each environment to increase the probability that rare alleles are present in the experiment. The effect of rare alleles of particular interest can be studied most sensitively when carriers of these interesting rare alleles are oversampled before the molecular phenotyping is performed, so that the phenotyped groups contain a more balanced representation of these individuals than the initial population.

We look forward to others enriching their genetical genomics experiments by

including sensibly chosen environmental variation. Doing so will open up a rich new area for studying the norms of reaction at the molecular level.

### 1.3 Thesis contribution and organization

Studying the effect of genetic perturbations on biological systems at the molecular level has been a hot research area since the completion of the first major genome projects and the introduction of the genetical genomics strategy (Jansen and Nap 2001). In this thesis, a generalization of genetical genomics which combines genetic and sensibly chosen environmental perturbations, as described in the **Chapter 1**, was developed to study the plasticity of molecular networks. This strategy forms a crucial step toward gaining a broader picture of the genomic responses to environmental perturbations, which are of interest to biomedical, agricultural, and evolutionary geneticists.

Doing genetical genomics is expensive. In **Chapter 2**, we show that current (generalized) genetical genomics studies can be improved significantly by applying a new experimental design. We developed a designGG R package and web tool for selecting samples from a population and intelligently allocating them to different conditions. DesignGG, which allocates individuals with dissimilar genomes to the same condition, gives more weight to factors of major interest and known regulatory loci if desired, and thereby maximizes the power for decomposing expression variation.

The ambitious goal of generalized genetical genomics experiments is to provide insight into the structure of regulatory networks underlying complex traits. In **Chapter 3** we discuss several important statistical issues involved in causal network inference for genome-wide linkage/association studies, including the effects of population size, QTL effect size, allele frequency, sensitivity and positive predictive value. In addition, one of the most interesting observations in genetical genomics studies are “hotspots”, i.e. genomic regions that influence a large number of molecular traits. The biological implications and statistical issues involved in detecting hotspots are discussed in detail in **Chapter 4**.

Generalized genetical genomics has already been applied in a number of biological studies in *C. elegans*, yeast (Smith and Kruglyak 2008) and mouse. **Chapter 5** describes the first genome-wide genetic study of gene expression plasticity by generalized genetical genomics and investigates whether environment-induced plastic responses of gene expression show heritable differences. We used recombinant inbred lines of the nematode worm *C. elegans* that were derived from parental lines

originally collected in Bristol (UK) and Hawaii (USA), and measured genome-wide gene expression at two different temperatures. Quantitative trait locus mapping uncovered genes with genetically determined differences in their plastic response to temperature changes, and a majority of them were found to be regulated by genes at a different genome position (regulated in *trans*).

**Chapter 6** describes a second application of generalized genetical genomics, this time in mouse. We analyzed gene expression across four developmentally closely related blood cell types collected from a large number of genetically different but related mouse strains. The results show that a large number of eQTLs exhibited a “dynamic” behavior across cell types, and the sensitivity of eQTLs to cell stage is largely associated with gene expression changes in target genes. These results stress the importance of studying gene expression variation in well-defined cell populations. Only such studies will be able to reveal the important differences in gene regulation between different cell types.

**Chapter 7** describes the first study of genetic variation controlling alternative splicing patterns (i.e., QTLs affecting the differential expression of transcript isoforms) in a large recombinant inbred population of *C. elegans*, using a new generation of whole-genome very-high-density oligonucleotide microarrays. Our findings suggest that the regulatory mechanism of alternative splicing in *C. elegans* is robust towards genetic variation at the genome-wide scale. This is in striking contrast to earlier observations in humans, which showed much less genetic robustness.

**Chapter 8** contains the summarizing discussion of this thesis, including the hard lessons learnt and perspectives for future research.



## Chapter 2

---

# designGG: an R-package and web tool for the optimal design of genetical genomics experiments

### ABSTRACT

*High-dimensional biomolecular profiling of genetically different individuals in one or more environmental conditions is an increasingly popular strategy for exploring the functioning of complex biological systems. The optimal design of such genetical genomics experiments in a cost-efficient and effective way is not trivial. This chapter presents designGG, an R package for designing optimal genetical genomics experiments. A web implementation for designGG is available at <http://gbic.biol.rug.nl/designGG>. All software, including source code and documentation, is freely available. DesignGG allows users to intelligently select and allocate individuals to experimental units and conditions such as drug treatment. The user can maximize the power and resolution of detecting genetic, environmental and interaction effects in a genome-wide or local mode by giving more weight to genome regions of special interest, such as previously detected phenotypic quantitative trait loci. This will help to achieve high power and more accurate estimates of the effects of interesting factors, and thus yield a more reliable biological interpretation of data. DesignGG is applicable to linkage analysis of experimental crosses, e.g. recombinant inbred lines, as well as to association analysis of natural populations.*

## 2.1 Introduction

Genetical genomics (Jansen and Nap 2001) has become a popular strategy for studying complex biological systems using a combination of classical genetics, biomolecular profiling and bioinformatics (Bystrykh *et al.* 2005, Schadt *et al.* 2005, Chen *et al.* 2008, Brem and Kruglyak 2005). By measuring molecular variation, using transcriptomics, proteomics, metabolomics and related emerging technologies, in genetically different individuals, genetical genomics has the potential to identify the functional consequences of natural and induced genetic variation. Recently, genetical genomics has been generalized to achieve a comprehensive understanding of

the dynamics of molecular networks by combining environmental and genetic perturbation (Li *et al.* 2008, Li *et al.* 2006b). This type of large scale ‘omics’ study leads to a better understanding of why individuals of the same species respond differently to drugs, pathogens, and other environmental factors.

However, most molecular profiling experiments are very costly, and as a consequence most genetical genomics studies are performed at the verge of statistical feasibility. Therefore, experimental design needs careful consideration to achieve maximum power from limited resources, such as microarrays and experimental animals (Churchill 2002, Fu and Jansen 2006). But, even in standard scenarios this requires sophisticated application of statistical concepts to intelligently select genetically different individuals from a population and allocate them to different conditions and experimental units. This topic has motivated classical statistical research since a long time (Fisher 1947). More recently, the concepts developed there have been adapted to the high dimensional data sets of post-genomics research (Churchill 2002, Kerr and Churchill 2001, Yang and Speed 2002, Fournier *et al.* 2007), and useful simplified design strategies have been suggested (Kerr and Churchill 2001, Wit *et al.* 2005a). However, to transfer these statistical ideas to the even more complex context of genetical genomics (Fu and Jansen 2006, Lam *et al.* 2008, Rosa *et al.* 2006) still requires considerable expertise in statistics.

Here we present an online web tool to make these selections and allocations easy for biologists with little/no statistical training. The program will find the best experimental design to produce the most accurate estimates of the most relevant biological parameters, given the number of experimental factors to be varied, the genotype information on the population, the profiling technology used, and the constraints on the number of individuals that can be profiled. Advanced users can download the underlying methods as an R package to adapt the program for a more tailored design. Without loss of generality, we will illustrate the method using microarrays, while they apply equally well to other profiling technologies, such as mass spectrometry. Also, we will only discuss molecular technologies that profile samples individually (e.g., single color microarrays) or in pairs (e.g., dual color microarrays), but an extension of the R scripts to more advanced multiplex technologies would be straightforward (Woo *et al.* 2005).


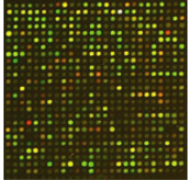
## 2.2 Implementation

The objective of *designGG* is to find an optimal allocation of genetically different samples to different conditions and experimental units (arrays) favoring a precise

## Optimize your Genetical Genomics Experiment

**1. Define analysis platform** ?

Single channel
  Dual channel

**2. Define individual genotypes** ?

Upload tab delimited file, e.g.

```

RILs
Markers "str1" "str2" "str3" "str4" "str5" "str6"
"C181" 1 0 0 0 1
"C182" 1 0 0 0 1
"C183" 1 0 0 0 1
"C184" 1 0 0 0 1
"C185" 1 0 0 1 1
"C186" 1 0 0 1
"C187" 1 0
"C188" 1
"C189" 1
  
```

Browse... Example file show advanced options

**3. Define experimental factors** ?

Factors	Levels				
<input checked="" type="checkbox"/> TempCelsius	15	24	29	+	-
<input checked="" type="checkbox"/> Tissue	Brain	Liver	Kidney	+	-
<input type="checkbox"/> Factor 3	1	2	3	+	-

show advanced options

**4. Set constraints** ?

Total number of slides

Number of strains per level

Optimize Experiment Design Test

Figure 2.1: Screenshot of the designGG web interface.

estimate of interesting parameters, such as main genetic effects and interaction effects between genotype and drug treatment. A simple case with one environmental factor can be expressed as  $y = \mu + G \times E + e$ , where  $y$  is the measurement vector,  $e$  is the error term, and  $G \times E$  denotes main effect and interaction effects of genotype and environment. In matrix notation, a model with one or more genotype factors (quantitative trait loci; QTL) and one or more environmental factors can be written as:  $Y = X\beta + E$ , where  $X$  is the design matrix of samples by parameters and  $\beta$  is the effect of genotype and environmental factors. The least squares estimate of  $\beta$  is  $b = (X^T X)^{-1} X^T Y$  with  $\text{var}(b) = \sigma^2 (X^T X)^{-1}$ . The optimal experiment design is defined as the one that minimizes the double sum of the variances of  $b$  firstly summed over all parameters and then summed over all genotypic markers. We use an optimization algorithm (simulated annealing (Wit *et al.* 2005b)) to search the experimental design space of all possible allocations to produce an optimal design matrix  $X$ . During the optimization, the algorithm utilizes the available marker information from the individuals to optimize the allocation of individuals to microarrays and conditions.

In the optimization, the experimenter can, of course, give more weight to parameters of higher interest, which will then be estimated with higher accuracy. Particularly, prior knowledge about expected effect sizes of interesting factors can be incorporated as weight parameters for the algorithm and the weight is inversely proportional to the expected effect size of the corresponding factors. In addition, it is also possible to specify the genome regions that are of major interest in a particular experiment, by specifying a region parameter. For example, if the relevant phenotype is known to map to certain genome regions, parameters for the markers in these regions can be given full weight in the optimization algorithm, whereas parameters for other markers can be given lesser or even zero weight. Thus, mapping resolution can improve and the power for finding QTLs in focal regions can be increased.

DesignGG is a package entirely written in the R language (The R Project for Statistical Computing, <http://www.r-project.org>). Every function of the designGG library is available as a stand-alone R tool and detailed help is available according to the standard format of R documentation.



## 2.3 Results

### 2.3.1 Web tool

Users can apply this method using a web interface (Figure 2.1) that we have generated using MOLGENIS (Swertz *et al.* 2004, Swertz and Jansen 2007):

1. Choose the platform. Select the single- or dual-channel option for one-color or two-color gene expression microarrays (the dual-channel option is also used for any other technology profiling pairs of samples).
2. Upload a tab separated value (TXT) file containing the genotype data matrix (individuals markers). Each cell contains a genotype label (e.g. A or B for the parental alleles, H for heterozygous loci; NA for missing data).
3. Set parameters. Specify the number of environmental factors, their number of levels, and the possible values of these levels. Specify either the total number of slides (assays) or the number of samples allocated within each condition.
4. Use advanced options if only one or a few genome regions or particular factors are of major interest. It is possible to optimize the experimental design by focusing on certain regions (e.g. the first 20 markers on chromosome I). Prior knowledge about expected effect sizes of interesting factors can also be incorporated as weight parameters for the algorithm.
5. Start the optimization algorithm by clicking on the button Optimize Experimental Design (Figure 2.1).
6. Get results. After the optimization is finished, the optimal experimental design will be displayed online (in table format), and will be available as text files for download.

**Table 2.1:** Example table of genotype data.\*

	Strain 1	Strain 2	Strain 3	Strain 4	Strain 5	...
C1M1	A	B	B	B	A	...
C1M2	A	H	A	B	A	...
C1M3	A	A	B	H	A	...
...	...	...	...	...	...	...

\* Heterozygous loci are indicated by an H.

### 2.3.2 R package

Here we illustrate how to apply the `designGG` R package using an example: suppose we are studying the effect of genetic factors (Q), temperature (F1), drug treatment (F2) and their interaction on gene expression using two-colour microarrays. There are 100 microarray slides available for this experiment, and we plan to study two different levels for each environment, which are 16°C and 24°C for F1 (temperature), and 5  $\mu$ M and 10  $\mu$ M for F2 (drug treatment). Then the R package can also be used in command line form as follows:

1. Prepare the input file specifying the genotype of each individual at each marker position. The file should be formatted as tab separated values (TXT), as illustrated in Table 2.1.
2. Load the `designGG` package by starting the R application and typing the command:

```
> library(designGG)
```

Specify the input arguments (Steps 3–5 correspond to steps 2–4 of using the web tool. The order of the following commands in steps 3-5 does not matter).

3. Choose the platform of the experiment. In this example, we use two-color microarray, thus:

```
> bTwoColorArray <- T # if paired; F otherwise
```

4. Load the marker data and specify the following required arguments (number of environmental factors, number of levels per factor, the values of each level, and the number of available slides):

```
> data(genotype)
# an example data attached with the designGG package
# The command below can be used to read TXT data
# genotype <- read.table( genotype.txt )
> nEnvFactors <- 2
> nLevels     <- c( 2, 2 )
> Level       <- list( c(16, 24), c(5, 10) )
> nSlides     <- 100; nTuple <- NULL
```

Table 2.2: The description and possible values of designGG arguments

Arguments	Description	Possible value(s)
TwoColorArray <sup>a</sup>	The type of platform	T(RUE) or F(ALSE) for the dual- or single-channel option, respectively. For example, F for one-color and T for two-color gene expression microarrays (the dual-channel option is also used for any other technology profiling pairs of samples)
genotype <sup>a</sup>	Genotype information	A matrix of marker genotypes for each marker and each strain. The values can be numeric: '1' and '0' for two homozygous genotypes, respectively (optionally, '0.5' for heterozygous allele). They can also be characters: 'A', 'B' or 'H' and 'H' is for heterozygous allele; NA for missing data. The column names are strain names, such as 'Strain 1', 'Strain 2', etc. The row names are marker names, such as 'C1M1', 'C2M2', etc.
nEnvFactors <sup>a</sup>	Number of environmental factors in the study	A numeric integer value between 1 and 3 which indicates the number of environmental factors to be studied. Experiments with more than three environmental factor are not recommended here since the power to estimate the high-order interactions is very limited for a realistic number of samples (several hundreds).
nLevels <sup>a</sup>	Number of levels for each environmental factor	A numeric integer vector. For example, there are two different levels for two environmental factors under study, then we use $nLevels < -c(2, 2)$
Level <sup>b</sup>	Level values for each environmental factor	A list which specifies the levels for each factor in the experiment. The element is a vector describing all levels of the environmental factor. In the given example, temperature levels are 16 and 24 and drug treatment levels are 5 and 10. The we use: $Level < -list(c(16, 24), c(5, 10))$
nSlides <sup>c</sup>	Total number of slides available for the experiment.	A numeric integer value
nTuple <sup>c</sup>	Average number of strains to be assigned onto each condition	A numeric value which is larger than 1
region <sup>b</sup>	Genome region of biological interest	A numeric integer vector which indicates the markers of biological interest, for example those previously detected for phenotypic quantitative trait loci. The value is the marker index (i.e., the row number in the <i>genotype</i> data table), <i>not</i> the marker name.
weight <sup>b</sup>	The weights for estimating genetic and environmental factors, and their interaction terms	A numeric vector which indicates the parameters of biological interest. Higher weights correspond to higher interest, and the optimization is adjusted in such a way as to result in a higher accuracy of the estimate for the parameters with higher weight. Prior knowledge about expected effect sizes of interesting factors can also be incorporated as <i>weight</i> parameters for the algorithm. The weight is inversely proportional to the expected effect size of the corresponding parameter, if the same relative accuracy is intended. <sup>d</sup>
nIterations <sup>b</sup>	Number of iterations of the simulated annealing method	A numeric integer value larger than 1. Default = 3000
directory <sup>b</sup>	Output file directory	The path where output files will be saved.
fileName <sup>b</sup>	Output file names	The name for output tables in CSV format to be produced.

<sup>a</sup>Required input arguments from users

<sup>b</sup>Optional input arguments

<sup>c</sup>Alternative arguments: either of them is required

<sup>d</sup>When there is no environmental perturbation, *weights* is 1, as there is only one parameter of interest (genotype); When  $nEnvFactor = 1$ ,  $weight = c(w_Q, w_{F1}, w_{QF1})$ ; When  $nEnvFactor = 2$ ,  $weight = c(w_Q, w_{F1}, w_{F2}, w_{QF1}, w_{QF2}, w_{F1F2}, w_{QF1F2})$ ; When  $nEnvFactor = 3$ ,  $weight = c(w_Q, w_{F1}, w_{F2}, w_{F3}, w_{QF1}, w_{QF2}, w_{QF3}, w_{F1F2}, w_{F1F3}, w_{F2F3}, w_{QF1F2}, w_{QF1F3}, w_{QF2F3}, w_{QF1F2F3})$ . Here  $w_Q$  represents the weight for genotype effect,  $w_{F1}$  represents the weight for environmental factor F1 effect and  $w_{QF1}$  represents the weight for interaction between genotype and F1 effect, etc.

An alternative to specifying *nSlides* is to specify *nTuple*, the number of strains to be allocated onto each condition. For example,

```
> nTuple <- 25 ; nSlides <- NULL
```

5. In addition to the required arguments specified in step 4, there are some optional ones for a tailored experimental design: e.g., we might be especially interested in the genome region between 1st marker and 20th marker, where a known phenotypic QTL from previous study locates. They can then specify that the optimization algorithm should only take genotypes at markers 1 to 20 into account:

```
> region <- seq( 1, 20, by=1 )
```

Additionally, if we want that the estimates of all interaction effects are twice as accurate as the estimates of the main effects (genotype, temperature and drug treatment), then we specify weights for the estimates:

```
> weight <- c( 0.5,0.5,0.5,1,1,1,1 )
```

Here the order of elements in the weight vector is such that first the main effects are listed, starting with the genotype, followed by the two environmental factors in the order used for *nLevels* and *Level*, then the one-way interactions, in the same order, and finally the two-way interaction between all three factors.

6. The following commands specify the directory where the resulting optimal design tables are to be stored and the name of the output files (design tables):

```
> directory <- C:\myproject\design  
> fileName <- myDesign
```

A detailed explanation of the above arguments can also be found in Table 2.2.

7. Run designGG to obtain your optimal design:

```
> myOutput <- designGG(genotype, nSlides, nTuple,  
  nEnvFactors, nLevels, Level, region=region,  
  weight=weight, nIterations=10)
```

It should be noted that the number of iteration of the simulated annealing method (*nIterations*) is set to 10 here for testing purposes. The default value (*nIterations* = 3000) is recommended, but it will result in a longer computing time.

8. Output can be found in the directory or retrieved with:

```
> optimalArrayDesign <- myOutput$arrayDesign
> optimalCondDesign <- myOutput$conditionDesign
```

Example output tables for allocation of strains on arrays and different conditions are shown in Table 2.3 and 2.4, respectively.

9. In addition, users can check the curve of optimization score recorded as the algorithm iterates using:

```
> plotAllScores ( myOutput$plot.obj )
```

Details of default settings such as *method* (SA: simulated annealing) or *nSearch* (equals 2) can be found in the designGG manual or the online help. Example genotype data and output tables are also provided along with the package. The R package can be found in Additional file 1 and most up-to-date version of the software can be downloaded at <http://gbic.biol.rug.nl/designGG>.

**Table 2.3:** Example table of the allocation of strains to arrays.\*

	Channel 1	Channel 2
<b>array 1</b>	Strain 28	Strain 92
<b>array 2</b>	Strain 70	Strain 47
<b>array 3</b>	Strain 22	Strain 89
...	...	...

\* This is applicable for technologies that profile samples in pairs, e.g. two-color microarrays.

## 2.4 Expected results

Two tables summarize the optimal design: The table pair design is only used for two-channel experiments and describes how samples are paired together in one assay e.g., a two-color microarray chip (Table 2.3). The table environment design lists how samples are assigned to environments/experimental factors (Table 2.4).

**Table 2.4:** Example table of the allocation of strains to experimental conditions.\*

	Temperature	Drug	Selected Strains				
<b>condition 1</b>	16	5	Strain 28	Strain 81	Strain 18	Strain 61	...
<b>condition 2</b>	24	5	Strain 70	Strain 40	Strain 83	Strain 92	...
<b>condition 3</b>	24	10	Strain 14	Strain 3	Strain 89	Strain 22	...
...	...	...	...	...	...	...	...

\* If the number of strains is smaller than the number of combinations of factors, the same strain can be used multiple times.

## 2.5 Conclusions

DesignGG, a freely-available R package and web tool presented in this work, represents a novel tool for the researcher interested in system genetics. Based on the careful experimental design provided by designGG, limited resources, such as arrays and samples, are maximally exploited, and more accurate estimates of parameters of interest can be achieved.

## 2.6 Availability and requirements

Project name: designGG R package and web tool

Project home page: <http://gbic.biol.rug.nl/designGG>

Programming language: R

Requirement: R statistical software available at <http://www.r-project.org/> for the stand-alone version.

## 2.7 Author contributions

YL developed designGG. RCJ and RB directed the project. MAS, GV and JF helped to implement the web tool. All authors wrote the manuscript, and read and approved the final version.

## 2.8 Acknowledgments

This work was supported by the Netherlands Organization for Scientific Research, NWO-86504001. We thank Danny Arends for help in implementing the web tool.

## Chapter 3

---

# Critical reasoning on causal inference in genome-wide linkage and association studies

### ABSTRACT

*Genome-wide linkage and association studies of tens of thousands of clinical and molecular traits are currently under way, offering rich data for inferring causality between traits and genetic variation. However, the inference process is based on discovering subtle patterns in the correlation between traits and is therefore challenging and could create a flood of untrustworthy causal inferences. Here we introduce the concerns and show they are valid already in simple scenarios of two traits linked or associated to the same genomic region. We argue that more comprehensive analysis and Bayesian reasoning are needed and can overcome some of these pitfalls, although not in every conceivable case. We conclude that causal inference methods may still be of use in the iterative process of mathematical modeling and biological validation.*

### 3.1 Causal inference from genetic data

Understanding how genes, proteins, metabolites and phenotypes connect in networks is a key objective in biology. Genes are transcribed and translated into proteins that can act as enzymes to convert precursor metabolites into product metabolites. These relationships are often depicted informally using graphs with arrows pointing in the assumed direction of causality, for example, from genes to proteins to metabolites to classical phenotypes. These diagrams reflect our assumptions about causality in biological systems and in many cases have been painstakingly validated in controlled experimental settings. Today, more than ever before, we are faced with large-scale “post-genomics” data that have the potential to reveal a multitude of yet unknown but potentially causal relationships.

Methods for causal inference have been introduced as early as the 1920s (Wright 1921) and have been further developed and applied since then in genetic epidemiology and other fields (Duffy and Martin 1994, Pearl 2000, Spirtes *et al.* 1993). Causal inference is a formal statistical procedure that aims to establish predictive models.

For example, if a reduction in the level of critical metabolite is the cause of a disease, then an intervention that increases the metabolite level should alleviate the disease. By contrast, if the reduced metabolite is a consequence of the disease, then intervention will not have the desired effect. Causal reasoning is thus critical to the process of target discovery in pharmaceutical research.

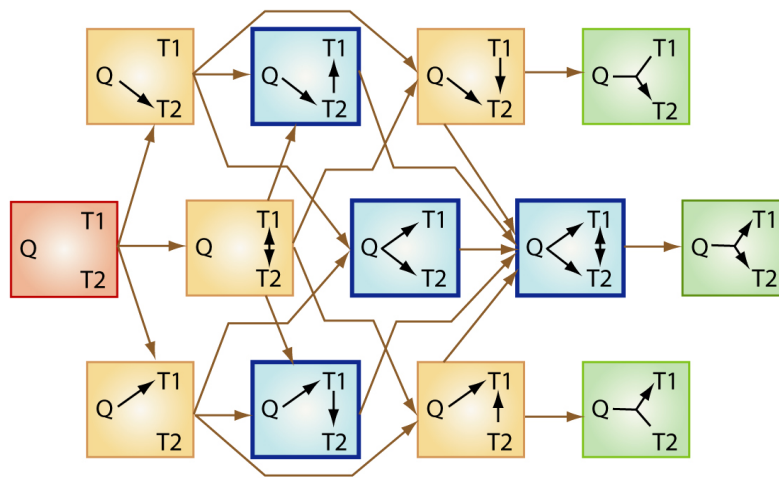
Recent genome-wide linkage studies (GWLS) on model organisms (Chen *et al.* 2008, Zhu *et al.* 2008, Schadt *et al.* 2005) and genome-wide association studies (GWAS) on humans (Emilsson *et al.* 2008) have successfully connected molecular and classical traits into networks with arrows indicating inferred causal relationships (Chen *et al.* 2007, Aten *et al.* 2008, Millstein *et al.* 2009, Chaibub Neto *et al.* 2008, Rockman and Kruglyak 2008, Zhu *et al.* 2004, Bing and Hoeschele 2005, Li *et al.* 2005, Kulp and Jagalur 2006). Causality cannot be established from data alone. Some assumptions about the causal relationships among the variables being modeled are needed. Once these are established, causal inference can be propagated to additional variables. In GWLS and GWAS settings it is typical to assume that genomic variation (quantitative trait locus; QTL) acts as a causal anchor from which all arrows are directed outward. Although this assumption seems quite natural, caution is warranted when the sample is not random, as in case-control studies.

There are many possible causal networks even in a simple system consisting of a genomic locus (QTL) and two traits, T1 and T2 (Figure 3.1). Causal inference in GWLS and GWAS involves, in its simplest form, the identification of pairs of traits with a common QTL (QTL-trait-trait triads) and determining whether the QTL directly affects each of two traits (independent), or if the QTL affects only one trait which in turn affects the other trait (causal or reactive). If none of these situations apply we assume that the causation is more complex (undecided).

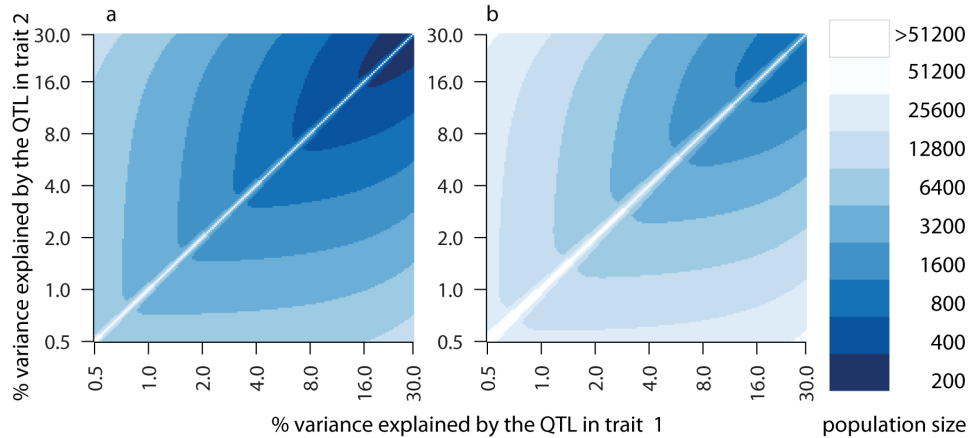
Biological variation in the two traits beyond that induced by the common QTL is key to distinguishing between the independent and causal scenarios. If there is a causal link, the biological and QTL variation from T1 will propagate to T2. If the variation propagates in an approximately linear fashion, we can, with simple linear regression (Box 3.1), subtract the biological and QTL variation in T1 from T2 and are left with the additional or 'residual' variation in T2 unrelated to the QTL. If we attempt the reciprocal analysis, the additional variation in T2 may make the linear regression fail to subtract all of the QTL variation from T1. As a result the residual variation in T1 will still relate to the QTL. This reasoning suggests a simple approach to distinguish among the independent and causal models on the basis of the outcome of two reciprocal statistical tests: does the residual variation in T1 still relate to the QTL, and does the residual variation in T2 still relate to the QTL.



Traits are declared independent (yes, yes), causal (yes, no), reactive (no, yes), or more complex (no, no) in which case no decision is made (see Box 3.1 for the statistical details). While the apparent simplicity of this approach is seductive, here we highlight some possible pitfalls illustrated by three simple but realistic scenarios, and discuss avenues to restoring the potential of causal inference.



**Figure 3.1: Triad models.** Many different causal relationships are possible among a triad of two traits ( $T1$  and  $T2$ ) and a QTL ( $Q$ ). The simplest case (red box) to the left shows no causality, in which case the QTL and the two traits do not influence each other. In the next set of models (yellow), at least one trait is not associated with the QTL. All these models are excluded from consideration based on the assumption that the QTL mapping step has correctly inferred the QTL-trait associations. The models that remain to be discriminated are highlighted in blue and green: the procedure to decide in favor of one of the blue causal topologies is outlined in the text. The three models furthest to the right (green) are extensions of the causal model that include additional interaction terms, e.g. the QTL may modulate the causal effect of  $T1$  on  $T2$ . Equivalently, these models may be seen as relaxing the assumption of equal covariance across genotype classes. An extreme scenario is the Simpson's paradox model in which the traits show opposite correlations for different genotypes at the QTL. Such complexities are usually not considered, but may form an important part of actual biological networks. The brown arrows indicate which of the models are nested and can thus be directly compared by statistical testing.



**Figure 3.2:** Population size required for reliable causal inference. Here we show the required population size in (a) genome-wide linkage studies (GWLS) and (b) genome-wide association studies. Each color represents a different population size; the scale is shown in the right panel. These numbers have been calculated from the equations in Table 3.1 by using a 10% significance threshold for the  $t$ -tests, 90% positive predictive value and 50% sensitivity. We assume that there is only biological variation and no measurement error. The  $x$  (or  $y$ ) axis indicates the percentage of variance explained by a QTL in trait T1 or T2, respectively on a logarithmic scale ranging from 0.5% to 30%. Allele frequencies of the biallelic QTL are set equal in GWLS, and 10% and 90% in GWAS. Furthermore we use Bayesian reasoning (Box 3.2): we assume a priori that only 1% (20%) of the QTL-trait-trait connections is truly causal in GWLS (GWAS).

### 3.2 Concerns about causal inference

It is compelling to explore how this causal inference method for QTL-trait-trait triads performs, particularly in GWAS where the majority of QTL identified explain much less than 5% of the total variance (Visscher *et al.* 2008). The method will declare certain triads to be independent and others to be causal, but such inferences are not without error. Of all triads that are truly causal, what proportion can be correctly identified as such? This proportion is referred in statistics as the 'sensitivity' of the method. It is good for a method to be sensitive, but not sufficient to make it of practical use. Triads with truly independent traits may also have a chance to be identified, incorrectly, as causal by the method. As a consequence, the potential number of false causal links arising from, say, 80% independent trait-trait pairs can overwhelm the number of true causal links arising from the 20% causal trait-trait pairs. The proportion of true causal links amongst those identified as causal is referred to in statistics

as the 'positive predictive value'. A good method combines a high positive predictive value, say 90%, with an acceptable sensitivity, say 10% or higher (see Box 3.1 for the statistical details). A QTL is a genomic region that can contain multiple candidate genes and polymorphisms. Without prior knowledge that two traits sharing a common QTL are biologically or biochemically related, they are more likely to be regulated by different genes or polymorphisms within the QTL region. In which case we would say the traits are independent and that their apparent relationship is explained by linkage disequilibrium and not by a shared biological pathway. Different types of prior knowledge about the (unknown) number of true causal and true independent relationships can be incorporated into the causal inference (Box 3.2).

**Box 3.1: Causal inference with triads.**

(A) Decision procedure

The triad analysis is a statistical decision procedure consisting of the following steps:

Step 1. Establish that two traits are linked to the same locus. This rules out the red and yellow models (Figure 3.1). We are ignoring the green models. So we are now reduced to the four blue models (independent, causal, reactive, undecided).

Step 2. Regress T2 on T1 and T1 on T2 to obtain residuals of each trait adjusted for the other. Denote residuals by R2 and R1, respectively.

Step 3. Compute a bivariate t-test for association between the residuals (R1 and R2) and the QTL. Note that R2 is 100% adjusted for both QTL effect under the causal model only (zero expected value; Table 3.1). We note that in other implementations of triad analysis one would compute univariate t-tests of R1 against QTL and R2 against QTL. This ignores the correlation between these two tests and we have amended it here.

Step 4. Choose a model based on outcomes of the bivariate t-tests using a p-value of, e.g., 10%: independent if (yes, yes), causal if (yes, no), reactive if (no, yes). If none of these apply we default to the "undecided" case.

(B) Properties of procedure

We describe two statistical measures and derive implications for population size:

**Sensitivity.** The sensitivity of the method is the probability of correctly detecting a true causal relationship. This probability is obtained from the non-central bivariate t-distribution (QTL effect of residuals determine the non-centrality; Table 3.1).

**Positive predictive value.** The positive predictive value is probability of a declared causal connection being true. We incorporate prior knowledge (Box 3.2): P1 is the product of the prior probability of a link to be causal times the probability to correctly identify a causal link as such; P2 is the product of the prior probability of a link to be independent times the probability to incorrectly identify an independent link as causal. Then the positive predictive value is  $P1 / (P1+P2)$ .

**Required population size.** The above process is repeated for all combinations of QTL variance in the two traits, and for sample size ranging from 200 to 51,200. The minimum sample size to achieve both 50% sensitivity and 90% positive predictive value is plotted (Figure 3.2).

We present three different scenarios to illustrate the properties of the method. In

**Table 3.1:** Equations for regression parameters in the basic independent and causal model (first scenario in the main text)<sup>a,b</sup>

		Independent model	Causal model
		$T1 = QTL + e1$	$T1 = QTL + e1$
		$T2 = QTL + e2$	$T2 = T1 + e2$
Regress T1 on T2	Slope	$1 - v_2/v_{t2}$	$1 - v_2/v_{t2}$
Regress residual R1 on QTL	QTL effect	$2v_2/v_{t2}$	$2v_2/v_{t2}$
	Variance <sup>c</sup>	$v_1 + v_2(v_2/v_{t2} - 1)^2$	$v_2(v_2/v_{t2} - 1)^2 + v_1(v_2/v_{t2})^2$
Regress T2 on T1	Slope	$1 - v_1/v_{t1}$	1
Regress residual R2 on QTL	QTL effect	$2v_1/v_{t1}$	0
	Variance <sup>c</sup>	$v_2 + v_1(v_1/v_{t1} - 1)^2$	$v_2$
Covariation of QTL effects	Covariance <sup>c</sup>	$v_1(v_1/v_{t1} - 1) + v_2(v_2/v_{t2} - 1)$	$v_2(v_2/v_{t2} - 1)$

<sup>a</sup> T1 and T2 have mean zero and equal QTL effect; this can always be achieved by subtracting the means and re-scaling.

<sup>b</sup> Here, e1 and e2 represent variance in the biological process, not measurement errors;  $v_1$  and  $v_2$  denote the variances of e1 and e2; and  $v_{t1}$  and  $v_{t2}$  denote the total variance which is sum of the QTL and the biological variances. The ratio  $v_1/v_{t1}$  is the proportion of total variance that is not explained by the QTL.

<sup>c</sup> Multiply by  $(1/n_A + 1/n_B)$  in case of two genotypes where  $n_A(n_B)$  is the number of samples with genotype A (B); multiply by  $4n/(n(n_A + n_B) - (n_A - n_B)^2)$  in case of three genotypes where  $n = n_A + n_H + n_B$  is the total number of samples. Note that  $4n/(n(n_A + n_B) - (n_A - n_B)^2) = 1/n_A + 1/n_B$  if  $n_H=0$ .

the first scenario T1 is causal for T2, all QTL and biological variation in T1 is propagated to T2 and, on top of this variation, T2 shows additional variation. This additional variation may originate from an independent perturbation such as another QTL affecting T2 but not T1, or an environmental perturbation affecting T2 but not T1. The correlation between T1 and T2 is resulting fully from the causal relationship between the two traits. Exact analytical equations can be used to compute the required population size to attain desired levels of sensitivity and positive predictive value (Box 3.1). It requires specifying the size of the QTL effect, the frequency in the population of the major QTL allele, and the prior believe that the triad is causal rather than independent. A population size of approximately 200-6,000 (GWLS) to 800-25,000 (GWAS) provides 50% sensitivity and 90% positive predictive value for causal inference with QTL explaining from 30% down to 0.5% of total variance (Figure 3.2, with parameters as specified in the legend). Lowering the sensitivity to 10% would reduce the required population size, but this effect is visible only in the area

close to the diagonal (Figure 3.2). In this area traits are too tightly correlated and there is little additional variation in T2, making it difficult to infer the correct causal direction, i.e. sensitivity is low.

**Box 3.2: Bayesian Reasoning.**

Bayes rule (Stephens and Balding 2009) is a probability property that allows one to combine evidence from data with existing knowledge and expertise through the inclusion of priors in an inference process. The definition of the prior in a causal inference on a QTL-trait-trait triad is the result of a partly subjective process that can be guided by the following considerations:

**QTL confidence interval size.** The larger the confidence intervals of the QTL are, the more likely it is that distinct polymorphisms control the traits. In GWLS, linkage disequilibrium is pervasive leading to large confidence intervals.

**SNP density in the QTL region within the population.** The more polymorphic the QTL region is, the more likely it is that the traits are actually controlled by distinct polymorphisms. In GWAS, populations are heterogeneous leading to a lot of allelic diversity along the genome.

**Gene density within the confidence interval.** Polymorphisms that lie within gene coding regions are more likely to propagate variation at phenotypic level than polymorphisms in non-coding regions. The fewer the number of genes within the QTL confidence interval, the more likely that the two traits are affected by the same polymorphism.

**Local or distant eQTL.** If a gene expression trait is locally regulated by an eQTL and the other trait is distantly regulated by the eQTL, then the gene with the local eQTL is more likely to be causal for the other trait than the other way around (Zhu *et al.* 2004).

**Additional shared QTL.** The sharing of multiple additional QTL between the two traits may be taken as additional evidence that they are connected in the network (Jansen and Nap 2001). It is more likely that these QTL affect the traits through the same polymorphisms than it is that locations of multiple distinct polymorphisms coincide by chance.

**QTL hotspot.** Regions of the genome, known as QTL hotspots, have been reported that harbor QTL for large numbers of traits. These could be the result of a single major polymorphism or of many polymorphisms in linkage disequilibrium and each affecting different traits independently. Further investigation and experience in understanding this phenomenon is needed to determine which is more likely.

**Independent biological knowledge.** Biological knowledge about the two traits (for example if the two genes belong to a same KEGG pathway) can be used as a priori evidence that the traits are related.

In the second scenario one or more shared hidden factors cause additional correlation between the traits. One can think of undetected QTL with pleiotropic effects on the traits, structural chromosomal variation leading to co-expression of genes in a particular region, physiological variation related to daily circadian rhythms, or environmental variation due to features of the experimental implementation. In a causal

model, the effect of the hidden factor acts on T2 in two ways: indirectly through T1, but also directly. For increasing values of hidden factor correlation (while keeping QTL and total variance constant), the linear regression will tend to subtract the effect of the hidden factor and not that of the QTL. As a consequence the causal links will look more like independent (yes, yes); increasing sample size will not help to attain the desired levels of sensitivity and positive predictive value. In an independent model, the effect of the hidden factor acts on T1 and T2 directly, and not indirectly. As with the causal model, for increasing values of hidden factor correlation (while keeping QTL and total variance constant), the linear regression will typically tend to subtract the effect of the hidden factor and not that of the QTL. However, in the special case of equal slopes for hidden factor and QTL, the linear regression will be able to subtract hidden factor and QTL effects. A true independent model then tends to change from correct identification (yes, yes) via either causal (yes, no) or reactive (no, yes) to undecided (no, no). Increasing sample size will help only when slopes are still slightly different, not if they are equal. Note that equal slopes cannot occur in the causal model, because the hidden factor acts directly and indirectly on T2. Sample size shown in Figure 3.2 is still approximately adequate if the hidden factor variance is small, i.e. equals at most the QTL variance.

In the third scenario, measurement error comes into play, which is realistic for most technologies for scoring molecular and classical traits. Note that the use of surrogate variables, such as RNA expression as a proxy for the causal protein levels, may also introduce a kind of measurement error. Measurement variation is never 'biologically' propagated from one trait to another trait, yet it will change (reduce or increase) the correlation between the two traits, and thus the causal inference will be affected. Correlated measurement errors are analogous to the hidden factor scenario described above with one exception. The special case of equal slopes for hidden factor and QTL can now occur also in the causal model: slopes for correlated measurement error and QTL can be equal. In this case, a true causal model can change from correct identification (yes, no) to undecided (no, no). Independent measurement errors will cause the linear regression to fail to subtract the QTL variation in both reciprocal analyses; therefore the causal model will tend to look more like independent (yes, yes) if measurement variance increases. However, an actual causal link from one trait measured with large measurement error to a downstream trait measured with small measurement error can be reported as reactive (Rockman and Kruglyak 2008). Again, increasing sample size will not be helpful to attain the desired levels of sensitivity and positive predictive value.

## Glossary

### Allele Frequencies

At a given polymorphic locus, the different alleles may have different predominance within the studied population. In GWLS using a cross originating from two inbred founders, the QTL has two alleles in equal frequencies in the population under study. By contrast, in GWAS due to a combination of random segregation, drift and selection, allele frequencies can be markedly different from equal. Imbalanced allele frequencies are less optimal for QTL detection

### Causal anchor

Causal anchors are causal relationships that are provided by knowledge external to the data. Because meiotic recombination is a random process that predates the establishment of phenotypes, correlation between DNA variation (QTL) and a trait implies causation of the DNA variation on the trait variation in experimental populations: QTL can therefore be used as causal anchors. The assumption should be carefully evaluated in natural populations, which may have hidden structure, or in case-control studies where sampling may indirectly alter allelic associations.

### Causal inference

A process of determining whether variation observed in a trait is a cause or a consequence of variation observed in another trait. Here we adopt the definition used in (Pearl 2000) that causality is defined by the effects of intervention in a system. If X is a cause of Y, then we can predict that an intervention that alters the level of X will result in a change in Y.

### Correlation

Correlation is a statistical measure of how much two variables change together. Correlation best captures linear relationships between variables (on original scale or after a transformation).

### Distant eQTL

A distant (or *trans*) eQTL is an eQTL which is located far from the gene it controls (for example on a different chromosome).

### eQTL

An expression Quantitative Trait Locus is a region in the genome at which allelic variation correlates with the mRNA expression level variation of a certain gene.

### Genome-wide association studies (GWAS)

A genome wide association study is an experiment in which the genomes of unrelated individuals is screened for genetic markers (typically millions of single nucleotide polymorphisms) at which allelic variation correlates with variation in studied traits.

### Genome-wide linkage studies (GWLS)

A genome wide association study is an experiment in which the genomes of related individuals is screened for genetic markers (typically a few hundreds or thousands of single nucleotide polymorphisms) at which allelic variation correlates with variation in studied traits. Examples of GWLS include experimental crosses such as recombinant inbred panels, intercrosses and backcrosses.

**Glossary(continued)****Local eQTL**

A local (or cis) eQTL is an eQTL which is located nearby the gene it controls in the genome. Often a local eQTL will be caused by allelic variation in the regulatory region of the gene or within the gene itself.

**mQTL**

A metabolite Quantitative Trait Locus is a region in the genome at which allelic variation correlates with the abundance variation of a certain metabolite.

**pQTL**

A protein Quantitative Trait Locus is a region in the genome at which allelic variation correlates with the abundance variation of a certain protein. Just like eQTL, pQTL can be local or distant according to the genomic position of the gene encoding for the protein relative to the QTL.

**Prior**

A prior (or prior probability) reflects the initial belief in a given proposition (such as “Trait T1 is causal for trait T2”) before observing the data. The application of Bayes’ rule combines the evidence provided by observed data with the prior to provide a measure of evidence of the proposition that accounts for previous experience or external knowledge.

**QTL confidence interval**

QTL mapping identifies regions of the genome in which allelic variation is linked or associated with a certain trait. The sample size, the density of available genotyped markers and the extent of recombination in the QTL region within the studied population are among the factors that influence the size of the confidence interval. Confidence intervals can extend from only a few hundred kilo base pairs to several mega base pairs complicating the identification of the actual polymorphism behind the QTL.

**QTL mapping** A genomic region is said to be a Quantitative Trait Locus for a trait if allelic variation in this region correlates with trait variation. QTL can be mapped through GWAS or GWLS.

**QTL-trait-trait triads**

A set constituted by a QTL and two traits mapping to that QTL. Since a QTL can affect directly a trait, or indirectly through another intermediary trait, multiple causal scenarios can explain this triad as illustrated in particular by the blue models in Figure 3.1. This article discusses our ability to discriminate between those different scenarios.

**Regression**

Regression is a statistical procedure which evaluates the dependence between a variable (e.g. a trait) and one or multiple other variables (e.g. another trait, or QTL genotypes).

**Residuals**

In a regression, residuals are the differences between the observed values and the values fitted by the regression.

**Variance**

Variance is a statistical parameter that quantifies the spread in the distribution of a variable. For phenotypic traits variance originates from both genetic and non-genetic sources and we can estimate the proportion of trait variance that is contributed by a given QTL.



### 3.3 Restoring the potential of causal inference

We have explored causal inference in the simple context of QTL-trait-trait triads using a statistical decision procedure (Box 3.1) to possibly reject the undecided model in favor of one of the nested causal, reactive and independent models. This procedure is similar to other implementations of triad analysis (Chen *et al.* 2008, Schadt *et al.* 2005, Chen *et al.* 2007) which, although not identical, lead to comparable results (Millstein *et al.* 2009). Other computational methods for causal inference such as structural equation modeling (Li *et al.* 2006a, Liu *et al.* 2008) or Bayesian network analysis (Zhu *et al.* 2007) can operate on larger numbers of traits and QTL. These methods also rely on the correlation structure in the data and will therefore suffer from some of the same problems as triad analysis: they require large population size, and can be confounded by hidden factors or measurement noise. This calls for several recommendations to restore the potential of causal inference.

Our first recommendation is to use Bayesian reasoning in the causal inference procedure. Prior belief or knowledge about the number of true causal and true independent links that might be expected in a typical QTL, depending on the study design, should be considered to safeguard against high false positive rates (low positive predictive values). In studies that involve mapping gene expression (eQTL), protein (pQTL) or metabolite (mQTL) traits, information about co-localization of QTL and genes that are functionally linked to the trait provides information about the likelihood of causal links. Lastly, biological annotations such as Gene Ontology (Ashburner *et al.* 2000) or Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000) pathways should also be considered when weighing evidence for causal links. The use of more informative priors (Box 3.2) provides better prioritizing and filtering of the large numbers of possible triads, and may reduce the required population size for reliable causal inference to more realistic numbers.

Our second recommendation is to identify and eliminate or account for experimental factors that can induce spurious correlation. It is not usually possible to measure all relevant factors, yet even some of the most obvious factors such as age or sex of study subjects are often not taken into account. Any variation in diet, time since last feeding or time of sample collection, the size of plant seeds or the size of litter, temperature and light cycles, location in the greenhouse or field, can have profound effects. Such factors can be easily included in the model, but only when they are recorded (Li *et al.* 2008, Akey *et al.* 2007). While it may not be necessary in inbred line cross studies, it is critical to consider the impact of population structure in almost every other setting where genetic variation is present. Methods are available to estimate kinship and the corresponding structure of the correlation. Combining

these methods with causal inference can minimize the effects of spurious genetic correlation (Kang *et al.* 2008). The effects of hidden factors affecting larger numbers of traits can be detected and corrected for by dimension reduction methods (Kang *et al.* 2008, Dubois *et al.* 2010, Fehrmann *et al.* 2008, Leek and Storey 2007, Stegle *et al.* 2010). Causal inference can then be applied to the residual data. However, these multivariate analysis methods also have the potential to remove signals relevant for causal inference from data and their application should be considered carefully.

Our third and final recommendation is to consider a richer set of possible models than the four blue models in Figure 3.1. For example, fitting a model like the top right yellow model in Figure 3.1 could provide a powerful case for the causal signal in the data (Kulp and Jagalur 2006, Li *et al.* 2006a, Liu *et al.* 2008). The green models in Figure 3.1 with more complex correlation structure can also be informative and have been explored (Kulp and Jagalur 2006). If two traits have multiple QTL in common, then this may be taken as additional evidence that the two traits are connected in the network (Jansen and Nap 2001). This allows for the possibility to generalize the triad analysis to a multiple QTL-trait-trait analysis. A test of the effects of all QTL that propagate from one trait to another can be obtained by modifying step 3 in the decision procedure (Box 3.1) to assess the combined effect (Sargon 1958).

### 3.4 Concluding remarks

Many in the scientific community share a healthy skepticism of causal inference and for good reasons as we have shown. Nevertheless we conclude that causal inference in linkage or association analysis may soon become a feasible strategy given the rapidly growing prior knowledge of biological networks, the increasing population sizes, the advent of cheaper and more accurate measurement techniques, and the possibility of coupling causal inference methods with Bayesian reasoning. Further development of methods that consider the simultaneous effects of multiple traits and multiple QTL is needed, as well development of techniques that address the effects of experimental factors, study design and population structure. Reasonable caution remains warranted and statistical methods of causal inference should be viewed as a necessary step in an era of high throughput data generation and discovery.

### 3.5 Appendix

This section contains two tables with equations for the hidden factor and measurement error cases. The procedures and definition of parameters in the tables are: we regress T1 on T2, and T2 on T1; i.e. we fit the model  $T1 = \beta_1 \times T2 + R1$ , where R1 denotes the residual, and the model  $T2 = \beta_2 \times T1 + R2$ , with residual R2.  $b_1$  and  $b_2$  are the least squares estimates of  $\beta_1$  and  $\beta_2$ , respectively. They are used to derive equations for the residuals R1 and R2.

The observed residual QTL effect for R1 can be shown to have expectation  $\mu_1$  and variance  $(1/n_A + 1/n_B)\sigma_1^2$  with  $n_A$  ( $n_B$ ) being the number of samples with genotype A (B). The estimates of  $\mu_1$  and  $\sigma_1^2$  are denoted by  $m_1$  and  $s_1^2$ , respectively, and specified in tables below.

Here  $v_1$  and  $v_2$  denote the variances of e1 and e2, and  $v_{t1}$  and  $v_{t2}$  the total variance (sum of the QTL variance and the error variances).

Table I: Equations for parameters in the independent and causal model with additional covariance by hidden factors (the hidden factors and the QTL have the different slopes in the regression models)\*.

	Independent	Causal T1 → T2
Model	T1 = QTL + e1 + e0 T2 = QTL + e2 + k × e0	T1 = QTL + e1 + e0 T2 = QTL + e1 + e2 + k × e0
$v_{t1}$	$1 + v_1 + d^2$	$1 + v_1 + d^2$
$v_{t2}$	$1 + v_2 + k^2 d^2$	$1 + v_1 + v_2 + k^2 d^2$
$b_1$	$1 - (v_2 + k^2 d^2 - kd^2)/v_{t2}$	$1 - (v_2 + k^2 d^2 - kd^2)/v_{t2}$
$m_1$	$2(v_2 + k^2 d^2 - kd^2)/v_{t2}$	$2(v_2 + k^2 d^2 - kd^2)/v_{t2}$
$s_1^2$	$v_1 + v_2(X - k)^2 + d^2(1 + X - k)^2$	$v_1 X^2 + v_2(X - 1)^2 + d^2(X + 1 - k)^2$
$b_2$	$1 - (v_1 + d^2 - kd^2)/v_{t1}$	$1 - d^2(1 - k)/v_{t1}$
$m_2$	$2(v_1 + d^2 - kd^2)/v_{t1}$	$2d^2(1 - k)/v_{t1}$ <sup>NOTE1</sup>
$s_2^2$	$v_2 + v_1(Y - 1)^2 + d^2(k + Y - 1)^2$	$v_1 Z^2 + v_2 + d^2(Z + k - 1)^2$
$s_{12}$	$v_1(Y - 1) + v_2(X - 1) + d^2(kX - k + 1)(Y - 1 + k)$	$v_1 X Z + v_2(X - 1) + d^2(kX + 1 - k)(Z + k - 1)$

<sup>NOTE1</sup> The residual of regression from downstream T2 to upstream trait T1 will now correlate to the QTL for  $k$  unequal to 1.

\* In the table,  $X = (v_2 + k^2 d^2 - kd^2)/v_{t2}$ ,  $Y = (v_1 + d^2 - kd^2)/v_{t1}$  and  $Z = d^2(1 - k)/v_{t1}$ .

Table II. Equations for parameters in the independent and causal model with uncorrelated measurement errors.

	Independent	Causal T1 → T2
Model	T1 = QTL + e1 + e01 T2 = QTL + e2 + e02	T1 = QTL + e1 + e01 T2 = QTL + e1 + e2 + e02
$v_{t1}$	$1 + v_1 + d^2$	$1 + v_1 + d^2$
$v_{t2}$	$1 + v_2 + d^2$	$1 + v_1 + v_2 + d^2$
$b_1$	$1 - (v_2 + d^2)/v_{t2}$	$1 - (v_2 + d^2)/v_{t2}$
$m_1$	$2(v_2 + d^2)/v_{t2}$	$2(v_2 + d^2)/v_{t2}$
$s_1^2$	$v_1 + d^2 + ((v_2 + d^2)/v_{t2} - 1)^2 d^2 +$ $((v_2 + d^2)/v_{t2} - 1)^2 v_2$	$d^2 + ((v_2 + d^2)/v_{t2})^2 v_1 + (((v_2 + d^2)/v_{t2}) - 1)^2 d^2$ $+ (((v_2 + d^2)/v_{t2}) - 1)^2 v_2$
$b_2$	$1 - (v_1 + d^2)/v_{t1}$	$1 - d^2/v_{t1}$
$m_2$	$2(v_1 + d^2)/v_{t1}$	$2d^2/v_{t1}$ <small>NOTE2</small>
$s_2^2$	$v_2 + d^2 + ((v_1 + d^2)/v_{t1} - 1)^2 d^2 +$ $((v_1 + d^2)/v_{t1} - 1)^2 v_1$	$d^2 + v_2 + (d^2/v_{t1} - 1)^2$ $d^2 + (d^2/v_{t1})^2 v_1$
$s_{12}$	$((v_1 + d^2)/v_{t1} - 1)(v_1 + d^2) +$ $((v_2 + d^2)/v_{t2} - 1)(v_2 + d^2)$	$(v_2 + d^2)d^2/(v_{t1}v_{t2})v_1 + (d^2/v_{t1} - 1)d^2 +$ $(v_2 + d^2)/v_{t2} - 1)d^2 + ((v_2 + d^2)/v_{t2} - 1)v_2$

NOTE2 The residual of regression from downstream T2 to upstream trait T1 will now correlate to the QTL for  $d > 0$ .

### 3.6 Acknowledgments

This work was funded by EU 7th Framework Programme under the Research Project PANACEA, Contract No. 222936 to YL, and by the BioRange programme from the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI) to BMT.

## Chapter 4

---

# Genetical genomics: spotlight on QTL hotspots

### ABSTRACT

*Genetical genomics studies the heritable variation in molecular traits, ranging from gene expression to metabolite levels. One of its main aims is the identification of “hotspots”, i.e. genomic regions that influence a large number of molecular traits. Such hotspots have been found to be surprisingly rare, indicating that most genetic variants are buffered and do not result in system-wide changes of molecule abundances. A recent study that reports a much larger number of hotspots turns out to be statistically flawed. The rarity of hotspots could be due to fundamental constraints on the function of biological systems in variable environments. This would have important implications for the study of complex common disorders using a genetical genomics strategy.*

## 4.1 Introduction

Genetical genomics aims at identifying quantitative trait loci (QTLs) for molecular traits such as gene expression or protein levels (eQTL and pQTL, respectively). One of the central concepts in genetical genomics is the existence of hotspots (Schadt *et al.* 2003), where a single polymorphism leads to widespread downstream changes in the expression of distant genes, which are all mapping to the same genomic locus. Several groups have hypothesized that many genetic polymorphisms—e.g., in major regulators or transcription factors—would lead to large and consistent biological effects that would be visible as eQTL hotspots.

## 4.2 Results/Discussion

Rather surprisingly, however, there have been only very few verified hotspots in published genetical genomics studies to date. In contrast to local eQTLs, which

Table 4.1: eQTL hotspots reported in selected genetical genomics studies.<sup>a</sup>

Paper	Organism	Population size	Number of		Threshold for eQTLs	Number of hotspots
			local eQTLs	trans eQTLs		
Brem et al., Science, (2002)	yeast	40	185	385	$p < 5 \times 10^{-5}$	8
Yvert et al., Nat Genet, (2003)	yeast	86	578	1,716	$p < 3.4 \times 10^{-5}$	13
Schadt et al., Nature, (2003)	mouse	111	1,022	1,985	LOD > 4.3	7
Kirst et al., Plant Physiol, (2004)	eucalyptus	91	1	8	experiment-wise $\alpha = 0.10$	2
Monks et al., AJHG, (2004)	human	15 CEPH families (167)	13	20	$p < 5 \times 10^{-5}$	0
Morley et al., Nature, (2004)	human	14 CEPH families	29	118	$p < 4.3 \times 10^{-7}$	2
Cheung et al., Nature, (2005)	human	57	65	0	$p < 0.001$	0
Stranger et al., PLoS Genet, (2005)	human	60	10-40	3	corrected $p$ value = 0.05	0
Chesler et al., Nat Genet, (2005)	mouse	35	83	5	FDR = 0.05	7
Bystrykh et al., Nat Genet, (2005)	mouse	30	478	136	genome-wide $p < 0.005$	"multiple"
Hubner et al., Nat Genet, (2005)	rat	259	622	1211	$p < 0.05$	2
Mehrabian et al., Nat Genet, (2005)	mouse	111	20107 total	20107 total	LOD > 2	1
DeCook et al., Genetics, (2006)	<i>Arabidopsis</i>	30	3525 total	3525 total	LOD > 3.4	5
Lan et al., PLoS Genet, (2006)	mouse	60	723	5293	LOD > 3.4	15
Wang et al., PLoS Genet, (2006)	mouse	312	2118	4556	$p < 5 \times 10^{-5}$	7
Li et al., PLoS Genet, (2006b)	<i>C. elegans</i>	80	414	308	$p < 0.001$ , FDR = 0.04	1
Keurentjes et al., PNAS, (2007)	<i>Arabidopsis</i>	160	1875	1958	FDR = 0.05	29
McClurg et al., Genetics, (2007)	mouse	32	N.A.	N.A.	N.A.	25
Ernlsson et al., Nature, (2008)	human	470	1970	52	FDR = 0.05	0
Schadt et al., PLoS Biol, (2008)	human	427	3210	242	$p < 1.6 \times 10^{-12}$ , FDR = 0.04	23
Chazalpour et al., PLoS Genet, (2008)	mouse	110	471	701	FDR = 0.1	4
Wu et al., PLoS Genet, (2008)	mouse	28	600	885,840 (Wu & Su, unpublished)	$p < 0.003$	1659

<sup>a</sup>The numbers are based on the statistical procedure and threshold used in the original publication, which can vary widely between papers. Where results based on multiple thresholds were reported, we included the most conservative one in the table. (N.A. not reported in the original paper. FDR: false discovery rate)

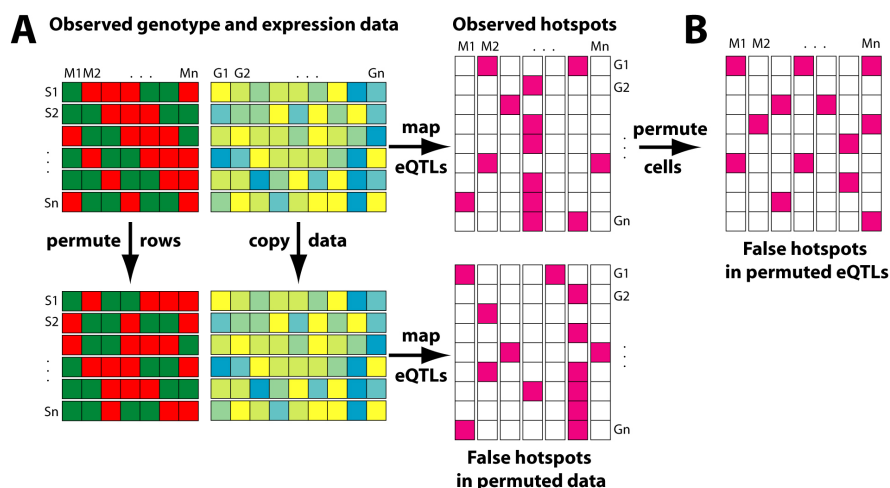
coincide with the position of the gene and are presumably acting in *cis*—e.g., by polymorphisms in the promoter region—distant eQTLs have been found to be more elusive. They seem to show smaller effect sizes and are less consistent, perhaps due to the indirect regulation mechanism, resulting in lower statistical power to detect them and, consequently, an inability to reliably delimit hotspots (de Koning and Haley 2005). While there are typically hundreds to thousands of strong local eQTLs per study, the number of associated hotspots is much lower. For example, a recent very large association study in about 1,000 humans did not find a single significant hotspot (Emilsson *et al.* 2008). Other studies have reported up to about 30 hotspots, far less than the number of significant local eQTLs (Table 4.1). The molecular basis is known for less than a handful of cases. An example is the *Arabidopsis* ERECTA locus, which leads to a drastic phenotypic change in the plant and has broad pleiotropic effects on many molecular (and morphological) traits (Keurentjes *et al.* 2007).

Recently, Wu *et al.* (Wu *et al.* 2008) reported the large-scale identification of hotspots. They studied gene expression in adipose tissue of 28 inbred mouse strains and performed eQTL analysis by genome-wide association analysis. The paper reports the identification of over 1,600 candidate hotspots, each with a minimum hotspot size of 50 target genes. Furthermore, they demonstrated that these hotspots are biologically coherent by showing that in about 25% of cases, the hotspot targets are enriched for functional gene sets derived from Gene Ontology, the KEGG pathways database, and the Ingenuity Pathways Knowledge Base. These findings suggested that genetic polymorphisms can indeed lead to large and consistent biological effects that are visible as eQTL hotspots. However, the authors chose a relatively permissive threshold of  $p = 0.003$  for QTL detection, uncorrected for multiple testing. In total, 886,440 eQTLs were identified at this threshold, i.e., 134 per gene. A permutation test (C. Wu and A. I. Su, unpublished data) shows that this results in a false discovery rate of 64%, largely resulting from multiple testing across 157,000 SNPs and 6,601 probe sets. This relatively permissive threshold was chosen because the focus of the analysis was on patterns of eQTL hotspots and not on individual eQTL associations. Analysis of eQTL patterns is relatively robust to individual false positives, and a permissive threshold allows for relatively greater sensitivity in detecting signal (Wessel *et al.* 2007). The authors observed an enrichment of specific biological functions among the genes in the reported hotspots. The study also reported that enriched categories tended to match the annotation of candidate regulators. Moreover, one predicted regulator was experimentally validated. In sum, these data seem to support the hypothesis that hotspots are downstream of a common master regulator linked to the eQTL. However, we suggest here that these

observations may also be explained by clusters of genes with highly correlated expression. If one gene shows a spurious eQTL, many correlated genes will show the same spurious eQTL, in particular if the false discovery rate for individual eQTLs is very high (de Koning and Haley 2005, Peng *et al.* 2007, Perez-Enciso 2004, Wang *et al.* 2007). There are many nongenetic mechanisms that can create strongly correlated clusters of functionally related genes. On the one hand, such clusters may be a result of a concerted response to some uncontrolled environmental factor. On the other hand, dissected tissue samples can contain slightly varying fractions of individual cell types, leading to cell-type-specific gene clusters, which vary in a correlated manner. The resulting correlation patterns represent potentially confounding effects, both for the correct determination of a significance threshold and for the biological interpretation of the resulting hotspots. Consequently, a key consideration in eQTL analysis is in the effective design of a permutation strategy to assess statistical significance. The approach used in (Wu *et al.* 2008) permuted the observed eQTLs among genes (Figure 4.1b). However, this approach has the disadvantage of ignoring the expression correlation between genes so that their spurious eQTLs no longer cluster along the genome. This permutation strategy leads to a potentially severe underestimate of the null distribution of the size of hotspots, when there are correlated clusters as described above.

An alternative strategy would have been to permute the strain labels as shown in Figure 4.1a, maintaining the correlation of the expression traits while destroying any genetic association (Churchill and Doerge 2008, de Koning and Haley 2005). As discussed above, it is expected that this would result in a more realistic significance threshold and a much smaller number of significant hotspots. Reanalysis of the data from (Wu *et al.* 2008) confirmed this idea: when permuting the strain labels (i.e., randomly swapping the genotypes between animals), the average maximum size of hotspots in the permuted data increases from less than 50 to 986. Consequently, even the largest hotspot in the real data only has a multiple testing corrected  $p$ -value of 0.23. This reanalysis demonstrates that expression correlation can indeed explain a large part of the co-mapping between genes. Such effects may also underlie some of the higher numbers of hotspots reported by some earlier studies (Table 4.1), especially where no appropriate permutation tests were applied to determine the statistical significance of hotspots (de Koning and Haley 2005). Of course, this does not imply that all hotspots are necessarily false positives. As described above, about 5% of the co-mapping clusters in (Wu *et al.* 2008) are not only functionally coherent but also map to a locus that contains a gene of the same functional class. This number is not statistically significant, but it is still suggestive of an en-





**Figure 4.1:** Alternative Permutation Strategies for Determining the Significance of eQTL Hotspots in Linkage and Association Studies. (A) The top panel shows the original data. The genotype matrix contains information about the genotype of each strain ( $S_1 \dots S_n$ ) at each marker position along the genome ( $M_1 \dots M_n$ ). For each strain, the expression of genes  $G_1 \dots G_n$  is measured. Linkage or association mapping combines these two sources of information to yield the eQTL matrix, where each purple entry indicates a significant linkage or association for a gene at a particular locus. The bottom panel illustrates the permutation strategy advocated here, where the strain labels are permuted, so that each strain is assigned the genotype vector of another random strain, while the expression matrix is unchanged. When the mapping is repeated on these permuted data, the correlation structure of gene expression is maintained, leading to an accurate estimate of the clustered distribution of false eQTLs along the genome. (B) shows the permutation strategy used in (Wu *et al.* 2008), where the original eQTL matrix is permuted by assigning the same number of eQTLs to genes randomly. The correlation of gene expression is lost, leading to an underestimate of the clustered pattern of spurious eQTLs.

richment of functional associations ( $p < 0.16$ , false discovery rate = 67%; C. Wu and A. I. Su, unpublished data). Some of these prioritized hotspots could correspond to true hotspots, and indeed one of them has been verified experimentally: cyclin H was validated as a new upstream regulator of cellular oxidative phosphorylation, as well as a transcriptional regulator of genes composing a hotspot (Wu *et al.* 2008).

Other studies, which used much stricter thresholds for defining their hotspots, also demonstrated the potential of interpreting putative hotspots by a closer study of the associated genetic locus (Zhu *et al.* 2008, Stylianou *et al.* 2008). An example is the recent work (Zhu *et al.* 2008): by combining eQTL information, transcription

factor binding sites, and protein–protein interaction data in a Bayesian network approach, they were able to predict causal regulators for nine out of the 13 hotspots (69%) originally reported in (Yvert *et al.* 2003). With integrated methods like these, it should be possible to identify those hotspots that are more than just clusters of co-expressed genes. As a result, the number of identified, functionally relevant hotspots could ultimately increase beyond the small numbers reported in Table 4.1. This would create new opportunities for gene regulatory network reconstruction. In any case, for the time being it seems that distant eQTLs and their hotspots are still scarce and hard to find, and that those that are reported should be interpreted with caution. This rarity of convincing hotspots in genetical genomics studies is intriguing. It could be due to the limited power of the initial studies, but it could also have a more profound reason. For example, it might well be that biological systems are so robust against subtle genetic perturbations that the majority of heritable gene expression variation is effectively buffered and does not lead to downstream effects on other genes, protein, metabolites, or phenotypes (Gibson and Wagner 2000, Carlborg and Haley 2004, Gibson and Dworkin 2004, Le Rouzic and Carlborg 2008). Experimental evidence for phenotypic buffering of protein coding polymorphisms is well established (Queitsch *et al.* 2002, Rutherford and Lindquist 1998). In fact, it has been shown that phenotypic buffering is a general property of complex gene-regulatory networks (Bergman and Siegal 2003). Also, if small heritable changes in transcript levels were transmitted unbuffered throughout the system, there would be a grave danger that genetic recombination would lead to unhealthy combinations of alleles and, consequently, to systems failure. Hotspots with large pleiotropic effects are thus more likely to be removed by purifying selection. If, as thus expected, common alleles are predominantly buffered by the robust properties of the system and hence largely inconsequential for the rest of the molecules in the system, this will have profound consequences for the design and interpretation of genetical genomics studies of complex diseases. Most importantly, it could turn out that even so-called common diseases—like diabetes, asthma, or rheumatoid arthritis—are not necessarily the result of common, small-effect variants in a large number of genes, but are rather caused by changes at a few crucial fragile points of the system (hotspots), which cause large, system-wide disturbances (Iyengar and Elston 2007, Bodmer and Bonilla 2008). Future studies in genetical genomics should aim at further elucidating the striking rarity of eQTL hotspots.

## Chapter 5

# Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*

### ABSTRACT

Recent genetical genomics studies have provided intimate views on gene regulatory networks. Gene expression variations between genetically different individuals have been mapped to the causal regulatory regions, termed expression quantitative trait loci. Whether the environment-induced plastic response of gene expression also shows heritable difference has not yet been studied. Here we show that differential expression induced by temperatures of 16°C and 24°C has a strong genetic component in *Caenorhabditis elegans* recombinant inbred strains derived from a cross between strains CB4856 (Hawaii) and N2 (Bristol). No less than 59% of 308 trans-acting genes showed a significant eQTL-by-environment interaction, here termed plasticity quantitative trait loci. In contrast, only 8% of an estimated 188 cis-acting genes showed such interaction. This indicates that heritable differences in plastic responses of gene expression are largely regulated in trans. This regulation is spread over many different regulators. However, for one group of trans-genes we found prominent evidence for a common master regulator: a transband of 66 coregulated genes appeared at 24°C. Our results suggest widespread genetic variation of differential expression responses to environmental impacts and demonstrate the potential of genetical genomics for mapping the molecular determinants of phenotypic plasticity.

## 5.1 Introduction

Expression quantitative trait loci (eQTLs) are polymorphic genetic loci that cause heritable differences in mRNA concentration. eQTLs have been used in recent genetical genomics studies (Jansen and Nap 2001) to infer the structure of genome-wide gene regulatory networks (Brem *et al.* 2002, Schadt *et al.* 2003, Stranger *et al.* 2005). The definition of eQTLs in these studies is essentially static and does not consider the highly dynamic nature of gene expression. However, mRNA levels respond rapidly to variable ambient conditions such as temperature change. This

has been shown for yeast (Xue *et al.* 2004), bacteria (Kraus *et al.* 2004), and *C. elegans* (GuhaThakurta *et al.* 2002) after exposure to heat shock.

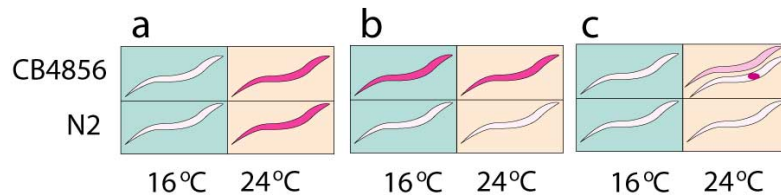
In contrast to these short-term exposures to extreme temperatures, populations under natural conditions are often exposed to longer periods of less extreme temperature changes. The ability to respond to these temperature changes (so-called phenotypic plasticity) differs among genotypes. Phenotypic plasticity to temperature plays an important role in the evolution of life histories in a variable climate (Roff 2002) and is widespread among species. Typical examples are temperature-induced sex determination in reptiles (Crews *et al.* 1994) and seasonal polyphenism in butterflies (Roskam and Brakefield 1996). The detection of temperature-specific proteins was reported by Madi *et al.* (Madi *et al.* 2003), who analyzed proteome temperature plasticity in wild-type *C. elegans*.

Insight into the genetic control of plasticity is a key issue for understanding evolutionary trajectories. Recently, we detected specific QTLs underlying plasticity to temperature in *C. elegans* life-history traits such as growth and fertility (Gutteling *et al.* 2006).

In this chapter we focus on the plasticity of gene expression in *C. elegans* juveniles that have been exposed for their entire life to (different) constant temperatures. We used a genetical genomics approach for detecting loci controlling such gene expression plasticity (plasticity quantitative trait loci [pQTL]). It has been shown that intraspecific evolution of variations in gene expression is to a large extent dominated by intense stabilizing selection (Denver *et al.* 2005). This implies that any beneficial mutation affecting gene expression levels should show its favorable effects selectively in certain environments without disrupting the existing adaptation to other conditions. This is much more likely the case for pQTLs than for nonplastic eQTLs. The “genotype-by-environment” interaction characterizing a pQTL is the prerequisite for adaptive evolution in a fluctuating environment (Levins 2004). In fact, it has been shown that more than half of the regulatory connections in a gene expression network are unique for specific conditions such as cell cycle, sporulation, DNA damage, and stress response (Luscombe *et al.* 2004). Recently, genotype-by-environment interaction was found for genome-wide gene expression among yeast strains (Landry *et al.* 2006).

## 5.2 Results/Discussion

We used a set of 80 recombinant inbred (RI) strains generated from a cross of N2 (Bristol) and CB4856 (Hawaii), representing two genetic and ecological extremes of



**Figure 5.1:** Illustration of Temperature, eQTL, and pQTL (eQTL-by-Temperature Interaction) Effects. Genotype (N2 and CB4856) and temperature (16°C and 24°C) are two factors that might induce differential expression for transcripts. The colors of the animals correspond to the different gene expression levels. (a) Transcript with differential expression induced by temperature. The transcript is overexpressed at 24°C independent of the genotype. (b) Transcript with strong eQTL effect. At both temperatures, worms with N2 genotype at a locus of interest show higher expression. (c) Transcript with pQTL effect. At 16°C, transcripts show low expression in both genotypes. At 24°C, only one allele (e.g., CB4856, as shown here) shows a strong induction of gene expression. If this upregulation is restricted to a specific tissue (the lower worm), it will be diluted in the total body when average of expression is measured (the upper worm). Other possible pQTL patterns can easily be conceived based on this example.

*C. elegans* (Hodgkin and Doniach 1997, de Bono and Bargmann 1998). Their genetic distance amounts to about one polymorphism per 873 base pairs (Wicks *et al.* 2001). Both strains have contrasting behavioral phenotypes (solitary versus gregarious) (de Bono and Bargmann 1998) and differ strikingly in their response to a temperature change (Gutteling *et al.* 2006). We have exposed the RI strains to 16°C and 24°C, temperatures that are known to strongly affect phenotypic characteristics such as body size, lifespan, and reproduction (Gutteling *et al.* 2006). Gene expression patterns were assessed by oligonucleotide microarray hybridization (Genisphere) using a distant pair design, which pairs the RI strains with the largest genetic difference on the same array, to maximize the amount of useful signal for the QTL mapping (Fu and Jansen 2006). The genetic architecture of the 80 RI strains and the description of a dense single nucleotide polymorphism (SNP) map can be found in Protocol S1 and Tables S1–S3.

### 5.2.1 Genome-wide detection of expression and plasticity QTLs

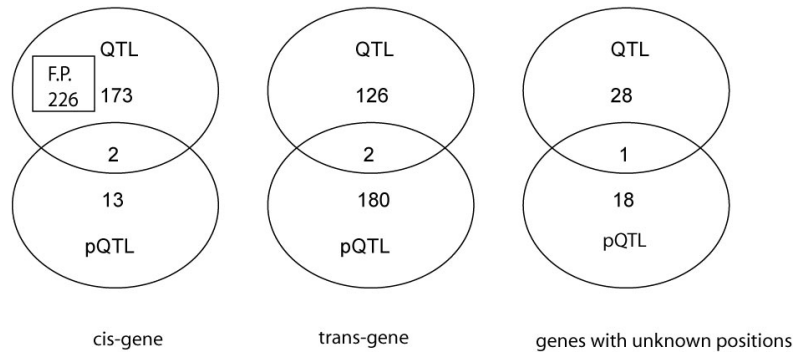
Schematic examples of eQTL, temperature, and eQTL-by-temperature interaction (pQTL) effects are shown in Figure 5.1a–c, respectively. We used a two-step procedure to detect pQTLs. First, we applied a separate eQTL analysis for the expression

data at either temperature (see Materials and Methods). With a genome-wide significance threshold of 4.25 (corresponding to an effective  $p$ -value of 0.001; throughout the chapter, thresholds are in units of  $-\log_{10}p$ ) there are 186 transcripts with significant eQTL effects at 16°C and 279 at 24°C, respectively (42 of these are common for both temperatures), suggesting eQTLs vary significantly between environmental conditions. To detect how much of this difference is due to pQTLs (plasticity regulators), we used the eQTL positions from the separate analyses. We postulated that interaction must happen at positions with eQTL effects and focused on these positions in a joint statistical analysis of data from both temperatures, thereby increasing the power of the method (see Materials and Methods for details). Differential expression for a given gene can result from *cis*-regulation due to variation in the region of the gene itself or from *trans*-regulation by other genes. The criterion used in our analysis is that the putative *cis*-acting QTL peak is within 2 Mb of the transcript. It is worthwhile to notice that the *cis*-QTLs could actually be *trans*-QTLs, due to the limited resolution of the mapping. We found 308 transcripts showing significant *trans*-acting eQTL effects (effective  $p < 0.001$ ) and 182 of these (59%) showed a significant pQTL effect (eQTL-by-temperature interaction) (Figure 5.2). This indicates that a large part of the observed gene expression dynamics differs consistently between the two parental alleles at plasticity-controlling loci.

That the temperature shift indeed leads to a drastic change in the gene regulation network is confirmed by the major differential gene expression observed between the two temperatures (Figure 5.3a). The amount of genes with a significant eQTL is relatively small (Figure 5.3b), while significant pQTLs are even less common, despite their relatively large effect size (Figure 5.3c). This justifies our use of the powerful two-stage statistical analysis outlined above.

### 5.2.2 Test for genetic assimilation

The parental lines of our RI strains originated from two very different thermal environments, and even though they have been maintained for many generations in controlled laboratory conditions, their highly divergent genomes are still expected to reflect the original allelic differences to a large extent. This gives us a unique opportunity to test our data for evidence of the controversial concept of genetic assimilation, whereby originally plastic traits become genetically fixed in a novel environment, e.g., because the original selective pressure favoring plasticity is no longer experienced (Pigliucci *et al.* 2006). In our case, we predict that genetic assimilation would be observed for temperature-related traits in the Hawaiian strain: genes that show strong differential expression in the highly seasonal conditions in Bristol lost

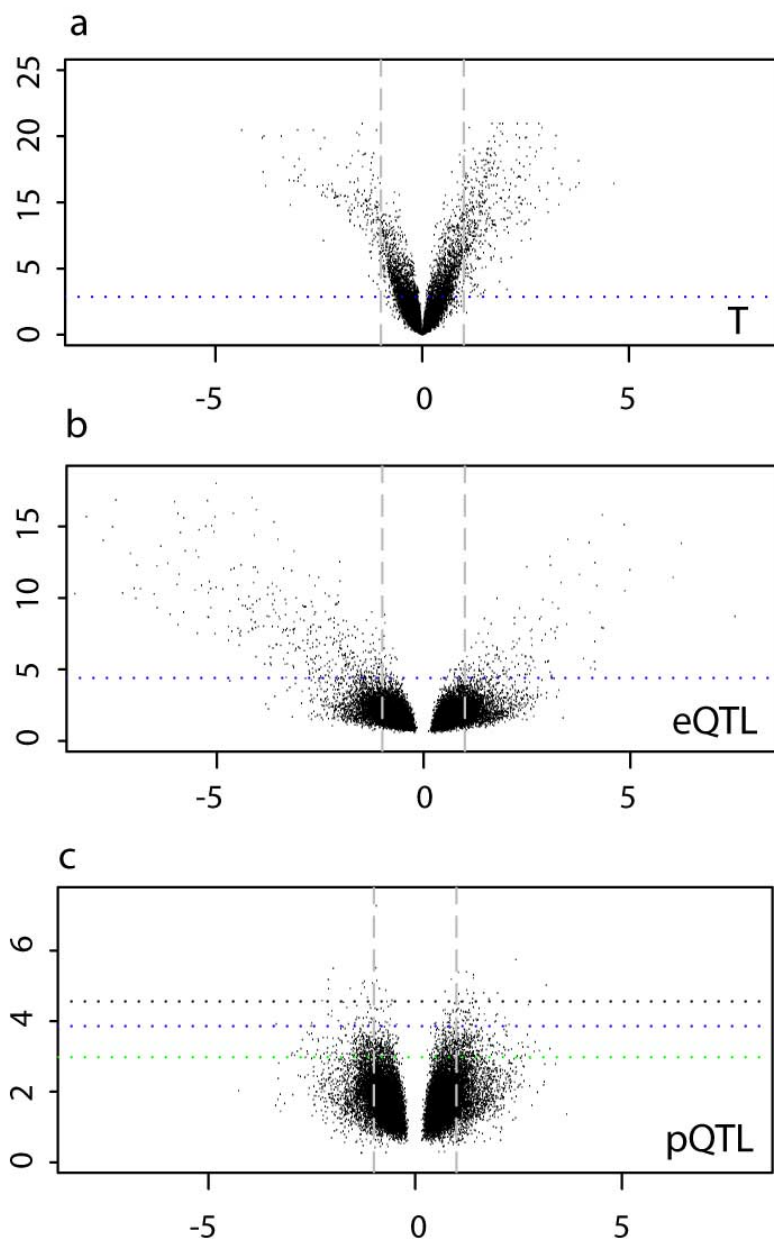


**Figure 5.2:** The figures indicate the number of transcripts detected with significant *cis*- and *trans*-eQTL or pQTL effect ( $p < 0.001$  with FDR of 0.04 after multiple testing correction) in a full ANOVA model (see Materials and Methods for details). In the first Venn diagram, F.P. refers to the number of estimated potential false positive eQTLs.

this behavior in the more constant tropical oceanic climate of Hawaii. This behavior would be reflected in the alleles in our RI strains. However, we find no evidence that genetic assimilation plays a role in the observed expression patterns. Out of 182 genes with pQTL, equal numbers of genes show strong differential expression when the plasticity-controlling *trans* locus carries the Hawaiian allele as when it carries the Bristol allele, and the most extreme differential expression is seen for control by the Hawaiian allele ( $p = 0.002$ , one-sided t-test), exactly the opposite of the predicted pattern. This result may be due to a lack of adaptation of Hawaiian worm strains to their specific environment, possibly due to recent population dispersal.

### 5.2.3 Functional assessment of transband genes

The most prominent case of pQTL in our dataset is found for a group of 66 genes that map to the same genomic region (Figure 5.4a) and in 63 out of 66 cases have a strong eQTL only at 24°C (Figure 5.4b). Of these genes, 41 have a stronger differential expression for the Hawaii allele ( $p = 0.05$ , one-sided Wilcoxon test) (Figure 5.4c). Such a temperature-specific transband (TB, or hotspot) seems extremely unlikely, both statistically ( $p \ll 0.001$ , hypergeometric test) and biologically, because it has been demonstrated recently that natural selection leads to the elimination of mutations in loci that affect many downstream gene expression levels (Denver *et al.* 2005). To test that the TB is not an artifact, we applied a permutation test (Materials and Methods). The results show that the TB does have a strong and significant genetic component (



**Figure 5.3:** Volcano Plots for Temperature, eQTL, and pQTL Effect. The temperature (T) (a), eQTL (b), and pQTL (c) effects for all genes are plotted on the x-axes. (a) Temperature effect  $-\log_{10}$  p-values from intensity-based analysis are plotted on the y-axis. (b and c) eQTL and pQTL  $-\log_{10}$  p-values from full model are plotted on the y-axes. Vertical dashed lines correspond to 2-fold change in expression. The dotted lines indicate the significance thresholds: (a) FDR 0.01; (b)  $p = 0.001$  for single and two-locus search; (c)  $p = 0.001$  for genomewide (black), single-locus (green), and two-locus (blue) search.



$p \ll 0.0001$ ). In addition to three miRNA genes in this region (*cel-mir-48*, *cel-mir-241*, and *cel-mir-257*), potential plasticity regulators for the transband genes are listed in Table S4. Additional analysis of the partial correlation coefficients between TB genes (Materials and Methods) shows that they are only partly controlled by the plasticity regulator at the *cis* position. This suggests that these genes are involved in the same pathway and controlled by a number of shared upstream factors. In fact, the TB genes form a conspicuous biological unit according to a gene ontology analysis (Maere *et al.* 2005), with enrichments in signal transduction ( $p = 0.03$  after multiple testing correction) and cell communication ( $p = 0.04$  after multiple testing correction).

The expression patterns of TB genes are also significantly correlated in an independent dataset (Kim dataset) (Kim *et al.* 2001) as compared with randomly selected genes (one-way Kolmogorov-Smirnov test,  $p \ll 0.001$ ) and they are enriched in the neuronal functional group (coexpression count 6,  $p < 7.9 \times 10^{-14}$ ) (Kim *et al.* 2001). It is particularly interesting to see that the group of 66 TB members contains one gene for an FMRFamide-related neuropeptide (*flp-9*) and four for G-protein coupled receptors (C17H11.1, C48C5.1, C24B5.1, and K10C8.2), all of them uncharacterized (Fisher's exact test,  $p = 0.02$ ). Expression variations of neuropeptides of the FMRFamide-related group (*flp-1* (Nelson *et al.* 1998), *flp-18*, and *flp-21* (Rogers *et al.* 2003)) as well as single amino acid mutations of their G-protein coupled receptor (*npr-1*) (Denver *et al.* 2005) underlie important ecological and behavioral differences among *C. elegans* strains (Denver *et al.* 2005, Nelson *et al.* 1998, Rogers *et al.* 2003). It is therefore tempting to speculate that the TB regulator occurred in two different alleles in the pedigree of the two parental populations (N2 and CB4856) because it controls an adaptive phenotypic difference in response to particular thermal conditions.

Interestingly, we found, in our related study of genotype-by-temperature interaction in classical phenotypic traits, that a fertility QTL maps to the immediate vicinity of our transband and shows the same interaction pattern. This suggests that our TB is possibly involved in fertility regulation or regulated by the same upstream factor(s).

#### 5.2.4 Estimating the rate of false-positives in *cis*-QTL effects

In addition to the *trans*-acting (p)eQTLs, which are the primary focus of the present chapter, previous studies (Manly *et al.* 2005, Doss *et al.* 2005) have also reported numerous *cis*-acting eQTLs, i.e., QTLs that explain expression variation of genes that are physically located at the same position as the QTL. However, as shown in Figure

5.4b, in our data there is a surprisingly high proportion of *cis*-acting QTLs that show a negative eQTL effect ( $p = 1 \times 10^{-9}$ , Fisher's exact test). One likely explanation is the confounding effect of SNPs on array hybridization. Under the assumption that true *cis*-acting QTLs with positive and negative eQTL effects should occur in equal proportions, we estimate that there are about 226 false positives among the *cis*-acting QTLs (402 total *cis*-QTLs minus twice the number of 88 *cis*-QTLs with positive eQTL effects). Following Hughes et al. (Hughes *et al.* 2001), we estimate that, on average, a single mismatch or indel in the ten nucleotides most 5' in our 60-mer probes would result in a significant detectable hybridization difference (about 40% decreased signal). The parental strains, N2 (for which the arrays were designed) and CB4856, differ in their genome sequence by up to one per 873 bp of aligned sequence. Ignoring, for simplicity, the unequal distribution of SNPs in coding and noncoding sequences, we thus estimate the number of genes with one influential SNP to be 238, which corresponds closely to the 226 false positives estimated above. This indicates that *cis* effects are not only less relevant for regulatory gene expression plasticity (topology of the gene regulatory network), but also very prone to hybridization artifacts.

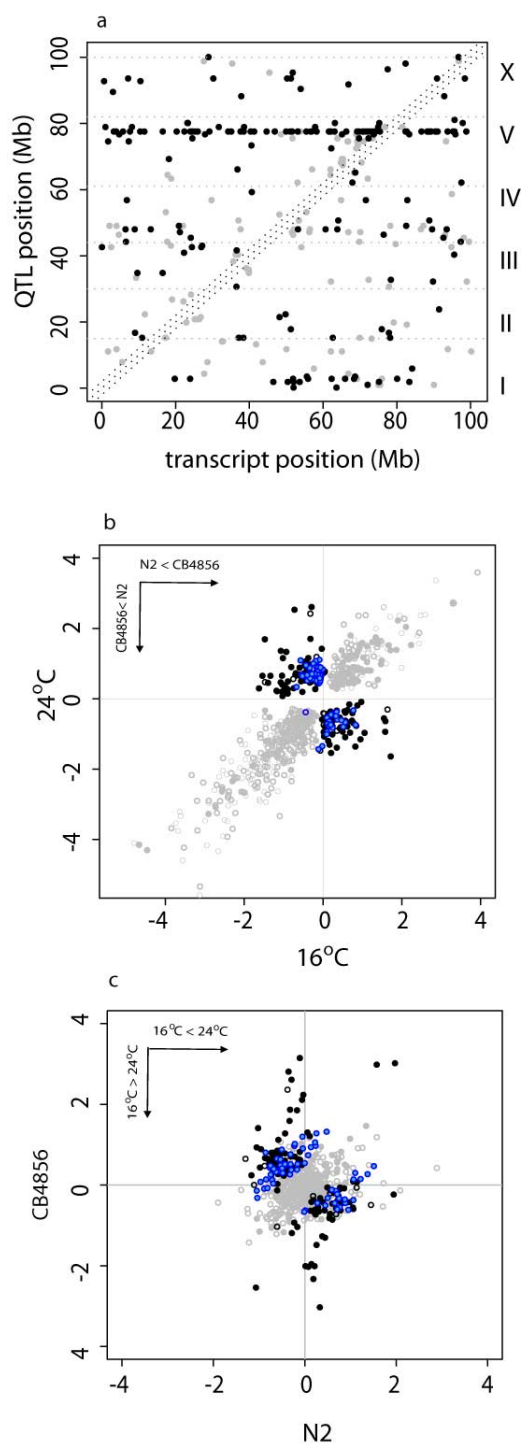
### 5.2.5 Power analysis for plasticity QTLs

Our ability to detect numerous pQTLs is even more striking when we consider that our approach is likely to underestimate the extent of environment-specific genotype effects (pQTLs). This underestimation is due to the fact that such effects have been diluted by measuring the average abundance of transcripts from all cells of *C. elegans* (Figure 5.1c); it is hard to detect a large pQTL effect if such an effect is actually cell-type specific.

To check that the number of pQTLs is not seriously underestimated due to our stringent statistical threshold, which might lead to false negatives, we estimated the detection power of interaction for various QTL effect sizes using simulation (Materials and Methods). We detected 98% of interactions if the difference in QTL effect is larger than two at the two temperatures (a pQTL effect of two, Materials and Methods). This suggests that our detection power is more than sufficient.

## 5.3 Conclusion

Recently the genetic architecture of gene expression has revealed many epistatic interactions in a constant environment (Brem *et al.* 2005). The present results imply that these interactions will change with environmental conditions. In addition,

**Figure 5.4:** Temperature-Dependent eQTL Effects

**Figure 5.4:** (a) Comparisons of eQTL and transcript positions for *trans*-regulated genes with significant eQTL and pQTL effects in the full model. The grey dotted horizontal lines separate the genome into different chromosomes. Grey and black circles indicate *trans*-regulated transcripts with significant eQTL effect and with significant pQTL effect, respectively, in the joint analysis. Among the transcripts with significant eQTL effect at both temperatures, a majority (72%) is *cis*-regulated (not included in the plot), while most of the transcripts (85%) with pQTL effect are *trans*-regulated. A horizontal transband was observed at 77.56 Mb (Chromosome V) by joint analysis. The transcripts falling into the region specified by dotted diagonals have *cis*-regulated eQTL (2 Mb).

(b) Comparison of eQTL effect for transcripts at two temperatures.

(c) Comparison of temperature-induced differential expression (*T* effect) for transcripts at two genotypes.

In (b) and (c), open and closed circles indicate *cis*-regulated and *trans*-regulated transcripts, respectively; grey and black circles are used for genes with significant eQTL and with significant pQTL effect, respectively. *Trans*-regulated transcripts in the 77.56Mb transband are colored blue.

we show that the plasticity of gene expression in *C. elegans* is mainly controlled by *trans*-acting pQTLs (genotype-by-environment interactions). Our results demonstrate widespread heritable variation in gene expression responses to environmental changes, which are used to generate the first comprehensive map of the genetic polymorphisms underlying differences in expression plasticity.

Future studies of ecological adaptation and evolutionary genetics of gene expression will benefit from this molecular genetics perspective, when exploring the plastic patterns of mRNA levels in different cell types, a wider range of environmental conditions, and a larger number of ecotypes.

## 5.4 Materials and methods

### 5.4.1 Genetical genomics experiment

#### Strain culturing

Both N2 and CB parental strains were homozygous. Strains were grown in 9-cm petri dishes at 15 °C or 20 °C on standard nematode growth medium with *Escherichia coli* strain OP50 as a food source and transferred to new dishes by a chunk of agar once a week. Recombinant inbred lines (RILs) were constructed by putting, on each of ten 6-cm dishes, one J4 hermaphrodite of strain N2 with five males of strain CB4856, and vice versa on each of ten other 6-cm dishes to avoid any mater-

nal or paternal effects. Mating was considered to be successful if the ratio of males to hermaphrodites was approximately 1:1 in the F1 hybrids. Approximately 1,500 F1 hermaphrodites were transferred to individual dishes in 24-well multiplates and allowed to self-fertilize at 20°C. This was repeated until F20.

### DNA isolation

For all lines, liquid cultures in S-basal (100 mM NaCl, 50 mM KH<sub>2</sub>PO<sub>4</sub> [pH 6.0], 5 mg/l cholesterol) were started and allowed to develop for one week in 50-ml tissue-culture flasks at 20°C. Cultures were transferred to 10-ml blue caps and centrifuged for 5 min at 4,000 rpm. Pelleted nematodes were transferred to a 1.5-ml Eppendorf tube, washed once with 1 ml M9 buffer, and centrifuged for 3 min at 8,000 rpm. After removal of the supernatant, 300  $\mu$ l lysis buffer (20 mM Tris-HCl [pH 8.0], 2 mM EDTA, 2% Triton X-100) and 5  $\mu$ l proteinase K (10 mg/ml) were added, and samples were left for 3 h at 65°C in a rotary shaker. Samples were washed with 400  $\mu$ l phenol:chloroform:isoamylalcohol (25:24:1) and centrifuged for 3 min at 14,000 rpm, after which the upper layer was transferred to a new tube. This step was repeated once. Next, 30  $\mu$ l 3 M sodium acetate (pH 5.0) and 750  $\mu$ l ice-cold isopropanol was added and samples were centrifuged for 3 min at 14,000 rpm. The DNA was washed once with 1 ml 70% ethanol and subsequently dissolved in 100  $\mu$ l Milli-Q water. 1  $\mu$ l RNase A was added and samples were incubated for 2-3 h at 37°C, after which they were stored at 4°C.

### Genotyping RILs

All markers were selected on the *C. elegans* SNP data website ([http://www.genome.wustl.edu/genome/celegans/celegans\\_snp.cgi](http://www.genome.wustl.edu/genome/celegans/celegans_snp.cgi)). For Chromosomes I, II, III, IV, and X, we selected 20 evenly spaced markers, for Chromosome V we selected 21 markers because this chromosome is larger than the other chromosomes. We selected easily detectable (i.e., with a common restriction enzyme) SNP markers with high  $P_{\text{snp}}$  values ( $P_{\text{snp}} \geq 0.7$ ), of which 75 were already confirmed.

PCR was performed on a Biozym MJ Research PTC-200 Peltier thermal cycler in thin-walled 200- $\mu$ l reaction tubes under the following conditions: 4 min at 94°C; 35 cycles of 45 s at 94°C, 45 s at 56°C, 45 s at 72°C; 5 min at 72°C. Total reaction volume was 10  $\mu$ l, with 5  $\mu$ l 20-fold diluted DNA sample, 1  $\mu$ l 10 PCR buffer (100 mM Tris-HCl [pH 9.0], 15 mM MgCl<sub>2</sub>, 500 mM KCl, 0.1% gelatin, 1% Triton X-100), 0.5  $\mu$ l 50 mM MgCl<sub>2</sub>, a final primer concentration (Gibco-BRL, [www.invitrogen.com](http://www.invitrogen.com); Isogen, [www.isogen-lifescience.com](http://www.isogen-lifescience.com); or Proligo, <http://www.proligo.com>) for each of a 0.4

pmol/ $\mu$ l, a final dNTP (Gibco-BRL) concentration of 0.2 mM, and a final Supertaq polymerase (HT Biotechnology, <http://www.sphaero-q.com/HTbiotechnology.html>) concentration of 0.02 U/ $\mu$ l.

Subsequently, samples were digested by adding 1  $\mu$ l of restriction enzyme buffer and 3 U of the appropriate restriction enzyme (Boehringer; Invitrogen, <http://www.invitrogen.com>; New England Biolabs, <http://www.neb.com>) directly to the sample. BSA was added if necessary. Digestions were performed for 3 h at the appropriate temperature, after which samples were loaded on 1.5%–3% agarose gels (depending on the expected fragment sizes) and run for 1.5 h at 100 V. Suspected mistypings were checked for a second time.

### Marker analysis

The order of markers was not based on a constructed linkage map but on their physical position in the sequenced genome. Physical and F2-derived genetic positions were obtained from Wormbase WS106 (<http://www.wormbase.org>). Marker segregation deviation (segregation distortion) from a 1:1 ratio was analyzed using a  $\chi^2$  test. To correct for Type I errors, we Bonferroni-corrected the significance level of these tests downwards with a factor of 12, which equals the estimated number of independent tests within our dataset: six for the chromosome number multiplied by two for the theoretical number of independent markers on each chromosome (the two outermost ones, which show approximately 50% recombination). Genetic distances between any two neighbouring markers were inferred from recombination fractions using the Kosambi mapping function. Recombination within one chromosome between neighbouring and nonneighbouring markers was analyzed by comparing the observed recombination using a  $\chi^2$  test in which the expected recombination was calculated with the inverse Kosambi function from twice the F2-derived distances between markers to correct for the multiple rounds of meiosis (Dixon 1993).

Association between any two markers on different chromosomes was analyzed for significant deviation from neutrality by comparing the overall number of associations and nonassociations (analogous to (non) recombinants if the markers were close to one another on the same chromosome) for any two markers with a calculated expected number using a  $\chi^2$  test. To obtain a model describing the expected fraction of association based on allele frequency, we performed nonlinear regression on data obtained from a simulation in which we determined the random association between two unlinked loci, each with two alleles, given a specific allele frequency for both alleles at both loci. The random association value finally used as input for

the model was an average based on 1,020 replicates in which for each replicate, 80 marker-to-marker comparisons were randomly selected out of a total of 1,000.

### **Culturing**

All recombinant inbred lines were reared on NGM agar plates seeded with the OP50 strain of *E. coli* as a food source. Stock cultures of OP50 were stored at  $-80^{\circ}\text{C}$ , and the bacterial cultures were grown in autoclaved LB medium (10 g peptone, 10 g yeast extract, 5 g NaCl/l water) for 16 h at  $37^{\circ}\text{C}$  and shaken at 150 rpm. Populations were started with only nonmated hermaphrodites and screened regularly to remove any occurring males.

### **Synchronization**

Experiments were carried out with nematodes belonging to the L3 life stage. To determine the entry into this stage at  $16^{\circ}\text{C}$  and  $24^{\circ}\text{C}$ , the size of the gonads and vulva were monitored. At 72 h of age, nematodes kept at  $16^{\circ}\text{C}$  were at the L3 stage, whereas 40 h of age determined this life stage at  $24^{\circ}\text{C}$ . Populations of each of the RILs were bleached (0.5 M NaOH, 1% hypochlorite) to collect synchronized eggs, which were then inoculated onto fresh dishes. Four replicate dishes of synchronized eggs for each RIL were kept in each of the two temperatures until L3 was reached. The nematodes were then collected and frozen in liquid nitrogen.

### **Probe construction and hybridization**

The parental N2 and CB4856 strains differ in their genome sequence by up to one per 873 bp of aligned sequence (Wicks *et al.* 2001). Koch *et al.* (Koch *et al.* 2000) reported that 85% of the SNPs were found in noncoding DNA. In an attempt to minimize hybridization differences based on SNPs, 60-mer oligonucleotide microarrays were used in this study. The frozen nematode samples were ground and RNA was extracted using the Trizol method, and cleaned up with the RNeasy Micro kit (Qiagen, <http://www1.qiagen.com/>). RNA concentration and quality was measured with a NanoDrop spectrophotometer (<http://www.nanodrop.com>). cDNA was obtained using Array 900 HS kit (Genisphere, <http://www.genisphere.com>) and Superscript II (Invitrogen). The cDNA samples were hybridized to 60-mer oligo arrays using the Genisphere Array 900 HS protocol. The probes on the arrays cover genes all over the genome. These 60-mers (provided by Washington University) were designed to uniquely represent each gene with proximity to the gene 3' end and with a minimum of secondary structure potential. All microarray data have been deposited in

NCBI's Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo>) and are accessible through the GEO Series accession number listed under the Accession Numbers heading in Supporting Information.

### Pairwise design

We adopted a novel distant-pair design for the microarray experiments, which was proposed especially for genetic studies on gene expression (Fu and Jansen 2006). In this design, the 80 RILs are hybridized directly on 40 arrays, in pairs that are maximally genetically different.

### 5.4.2 Data analysis

Full ANOVA model for pQTLs and eQTLs. The expression data of two temperatures were first analyzed separately by the following ANOVA model (Fu and Jansen 2006)

$$y_i = \mu + Q_i + e_i \quad (5.1)$$

where  $y_i$  is the gene's log ratio at the  $i$ th microarray;  $\mu$  is the mean;  $Q_i$  is  $-\beta$ ,  $0$ ,  $\beta$ , for arrays comparing A/B, A/A or B/B, and B/A, respectively;  $\beta$  is the effect of differential allele expression between A and B at a regulatory locus (or nearby marker) under study; letters A and B correspond to N2 and CB4856, respectively; and  $e_i$  is the residual error (see (Fu and Jansen 2006) for details).

Then expression data of two temperatures are combined together and analyzed by a full ANOVA model including T and eQTL\*T effects:

$$y_{ij} = \mu + Q_i + T_j + (QT)_{ij} + e_{ij} \quad (5.2)$$

where  $y_{ij}$  is the gene's log ratio at the  $i$ th microarray ( $i = 1, \dots, n$ ) and  $j$ th temperature,  $T_j$  is the temperature effect for  $j$ th temperature,  $(QT)_{ij}$  is the interaction effect (pQTL) between the  $i$ th eQTL genotype and  $j$ th temperature, and  $e_{ij}$  is the residual error. To increase the power of detecting interaction, we not only did a genome-wide linkage analysis, but also reduced the multiple testing issue by focusing on those three marker positions that show a maximum eQTL in either the full model or one of the two single temperature models. Interaction was assessed at these three positions using significance thresholds determined by simulation. The same strategy was applied for detecting significant eQTL effects.



### Two-stage search for pQTLs

Firstly we did a genome-wide linkage analysis, and then reduced the multiple testing issue by focusing on those three marker positions that show a maximum eQTL in either the full model or one of the two single temperature models. At the strongest eQTL genome position SL (single locus), the corresponding pQTL effect for each transcript was judged to be significant or not. As we expect a pQTL for a gene to occur at the positions with eQTL at one of the two temperatures, we focus on the strongest eQTL genome positions (obtained by separate analysis) for each transcript at 16°C and 24°C. These positions we call TL (two loci, one locus per temperature). At the TL, we checked if the pQTL effect obtained by joint analysis is significant or not. The thresholds were obtained by simulation. A gene is claimed to have a significant interaction effect if it passes the corresponding threshold at one of three positions (SL and TL). The same strategy was applied for detecting significant eQTL effects.

### Determination of genome-wide significance thresholds

To calculate the genome-wide threshold for separate analysis, we performed the following five steps. (1) We simulated trait data by randomly sampling from a standard normal distribution (with zero mean and unit variance) 1,000 times under the null hypothesis of no eQTL. We did this for 16°C and 24°C. (2) We carried out a single marker analysis for all 1,000 runs mimicking 16°C and then for the 1,000 runs mimicking 24°C. (3) At each marker, we obtained the corresponding  $-\log_{10} p$ . (4) We took the maximum overall markers and stored this value. (5) These values were ordered from low to high over all 1,000 runs, and their  $100(1 - \alpha)$  percentile was the estimated critical value (genome-wide threshold).

For the joint analysis the threshold can be obtained in a similar way. After simulating the trait data under the null hypothesis of no eQTL for two temperatures, the joint analysis was applied to the combined data of 16°C and 24°C. Then the genome-wide threshold for eQTL and interaction was obtained at a significant  $p$ -value of 0.001. With the same simulated data, we calculated the  $(-\log)_{10} p$  of interaction effect at SL position or TL positions and stored these values, respectively. At the significance level of 0.001, the thresholds for single locus and two-locus analysis can be obtained. The same strategy was applied for the eQTL effect.

In our analysis, we set the genome-wide  $\alpha$  to be 0.001 at 16°C and 24°C, as well as in the joint analysis. This implies that-with 20,490 transcripts-we expect only  $0.001 \times 20,490 \approx 20.5$  false positives. The threshold of 4.25 was obtained for the

separate analyses at both temperatures. For the joint analysis, the genome-wide threshold for eQTL is 4.50 and the single-locus threshold is 4.41. For the interaction effect, the genomewide threshold is 4.56 while the single-locus threshold and two-locus threshold are 2.98 and 3.88, respectively.

### Estimation of temperature-induced differential expression (T effect) based on intensity data

The intensity-based analysis considers the model

$$y_{ijk} = \mu + Q_i + T_j + (QT)_{ij} + S_k + D_g + e_{ijk} \quad (5.3)$$

where  $y_{ij}$  is the gene's log intensity at the  $i$ th microarray ( $i = 1, \dots, n$ ) and  $j$ th temperature;  $T_j$  is the temperature effect for the  $j$ th temperature;  $(QT)_{ij}$  is the interaction effect (pQTL) between the  $i$ th eQTL genotype and  $j$ th temperature,  $S_k$  is the random spot effect,  $D_g$  is the effect of the  $g$ th dye, and  $e_{ijk}$  is the residual error. Firstly, the QTL effects of two temperatures estimated from ratio data using the full model as described in the main text were used to replace the  $Q_i$  and  $(QT)_{ij}$  terms by constant values in the intensity-based model. Then temperature-induced differential expression effects were estimated from the remaining model.

### Coexpression of transband genes in Kim dataset

The experiments in the Kim dataset (Kim *et al.* 2001) compare RNA between mutant and wild-type strains or between worms grown under different conditions. The dataset consists of expression of 19,738 genes in 553 experiments. 56 out of 66 of our TB genes are found in the Kim dataset. We calculated all pairwise Pearson correlation coefficients among these 56 genes. Then we randomly chose the same number of genes from the Kim dataset 10,000 times and calculated the correlation coefficients of each pair of them. The resulting distribution is compared with that among the original TB genes by a one-way Kolmogorov-Smirnov test ( $p$ -value  $\ll 10^{-10}$ ).

### Permutation test for the transband

We used the real gene expressions of transband genes (i.e., the structure of correlation is kept unchanged), but reassigned different genomes to the different TB randomly to disturb the association between trait and genotype. From 10,000 permutations, the maximum genome-wide number of QTL for each permutation is

stored and the 99.9 percentile corresponding to a  $-\log_{10}p$  of 6 was obtained. The results show that the TB does have a strong and significant genetic component ( $p \ll 0.0001$ ).

#### **cis-factor test for transband**

Pearson correlation coefficients (zero order) were first calculated for the trait data of transband genes at 24°C. Then first order partial correlation coefficients conditioning on the genotype of the transband position (marker 97th) were calculated according to the following formula:

$$r_{x,y,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2 - r_{yz}^2}} \quad (5.4)$$

where  $r_{xy}$ ,  $r_{xz}$ , and  $r_{yz}$  are the Pearson correlation coefficient of gene expression between x and y, x and z, and y and z, respectively. We simulated random trait data for the same number of genes as in the TB and calculated corresponding zero and first order correlation coefficients. The results show that the first order partial correlation coefficient on genotype for TB genes decreases significantly from zero order coefficients. However, they are still larger than those for random traits. This indicates that the TB genes are only partly controlled by the master regulator at the QTL position and that these genes are involved in the same pathway and controlled by a number of shared upstream factors.

#### **Power of detection for pQTL by full model**

Compared with the total number of transcripts, only about 0.8% of 20,000 genes had a detectable pQTL effect, i.e., a surprisingly low proportion of regulatory connections seem to respond differentially to the major environmental change in the two genotypes. To check that this is not due to our stringent threshold, which might lead to false negatives, we estimated the detection power of pQTL for various eQTL effect sizes using simulation. We simulated the expression data for 1,000 genes with an eQTL effect size of B but opposite sign at two temperatures. Then the strategy of searching for pQTL used in real data was applied for the simulated data. The detected proportion of genes with significant pQTL indicates the power of our two-stage search method. With varying B between 0 and 5 with interval 0.25, the power of detection for pQTLs can be estimated. We detect 98% of interactions if the eQTL effect is larger than 1 and has opposite signs at the two temperatures, which corresponds to a pQTL effect of 2. This suggests that our detection power is more than

sufficient.

#### **Master regulator for transband searching**

There are 66 genes with significant pQTL at 77.56 Mb (Chromosome V). It is likely that there is a *cis*-acting master regulator at the QTL position. We first averaged the pQTL profiles for the transband genes and then took a 1.5 dropoff ( $-\log_{10}p$ ) to obtain genome region 75.91 – 79.33 Mb as the searching region. There are 1,180 potential candidates in total with a physical location in this region (819 potential candidates had a measured expression level in our dataset). We divided them into different groups according to their eQTL and pQTL effect and their annotation (see Table S4). The top candidates might be the genes that themselves have a significant pQTL effect (e.g., Y75B12B.3), and eQTL effect, ( e.g., *nhr-54* and *nhr-116*) involved in transcription factor activity, and map in *cis*; i.e., have a possible regulatory polymorphism in their promoter region.

#### **URLs**

This supporting information of this chapters is in <http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.0020222#s4>

## **5.5 Acknowledgments**

We thank G. de Haan, D. Denver, B. Walsh, L. van de Zande, and J. Bakker for stimulating discussions and comments. Nematode strains were provided by the *Caenorhabditis* Genetics Center, Minnesota, United States. We also thank S. R. Wicks for providing primers and Jos Molthoff of Plant Research International for allowing us to use the hybridization equipment.

## Chapter 6

# Expression quantitative trait loci are highly sensitive to cellular differentiation state

### ABSTRACT

*Genetical genomics is a strategy for mapping gene expression variation to expression quantitative trait loci (eQTLs). We performed a genetical genomics experiment in four functionally distinct but developmentally closely related hematopoietic cell populations isolated from the BXD panel of recombinant inbred mouse strains. This analysis allowed us to analyze eQTL robustness/sensitivity across different cellular differentiation states. Although we have identified a large number (365) of “static” eQTLs that were consistently active in all four cell types, we found a much larger number (1283) of “dynamic” eQTLs showing cell-type-dependence, and out of which 140, 45, 531, and 295 eQTLs were preferentially active in stem, progenitor, erythroid and myeloid cells, respectively. A detailed investigation of those dynamic eQTLs showed that in many cases the eQTL specificity was associated with expression changes in the target gene. We found no evidence for target genes that were regulated by distinct eQTLs in different cell types, suggesting that large-scale changes within functional regulatory networks are uncommon. Our results demonstrate that heritable differences in gene expression are highly sensitive to the developmental stage of the cell population under study. Therefore, future genetical genomics studies should aim at studying multiple well-defined and highly-purified cell types in order to construct as comprehensive a picture of the changing functional regulatory relationships as possible.*

## 6.1 Introduction

Genetical genomics uses quantitative genetics on a panel of densely genotyped individuals to map genomic loci that modulate gene expression (Jansen and Nap 2001). The quantitative trait loci identified in this manner are referred to as expression quantitative trait loci, or eQTLs (Schadt *et al.* 2003). Most genetical genomics studies that have thus far been reported have analyzed single cell types or compared developmentally unrelated and distant cell types (Bystrykh *et al.* 2005,

Chesler *et al.* 2005, Hubner *et al.* 2005, Petretto *et al.* 2006, Monks *et al.* 2004, Morley *et al.* 2004). Here, we report the first application of genetical genomics to study eQTL dynamics across closely related cell types during cellular development. We show results that discriminate between eQTLs that are consistently active or “static” and those that are cell-type-dependent or “dynamic”.

We used the hematopoietic system as a model to analyze how the genome of a single stem cell is able to generate a large variety of morphologically and functionally distinct differentiated cells. Differentiation of hematopoietic stem cells towards mature, lineage-committed blood cells is associated with profound changes in gene expression patterns. The search for differentially expressed genes, most notably for those transcripts exclusively present in stem cells and not in their more differentiated offspring, has been successful and has provided valuable insight into the molecular nature of stem cell self-renewal (Ivanova *et al.* 2002, Chambers *et al.* 2007, Kiel *et al.* 2005, Forsberg *et al.* 2005). Yet, complementary approaches were needed to elucidate the dynamic regulatory pathways that are underlying the robust differentiation program leading to blood cell production.

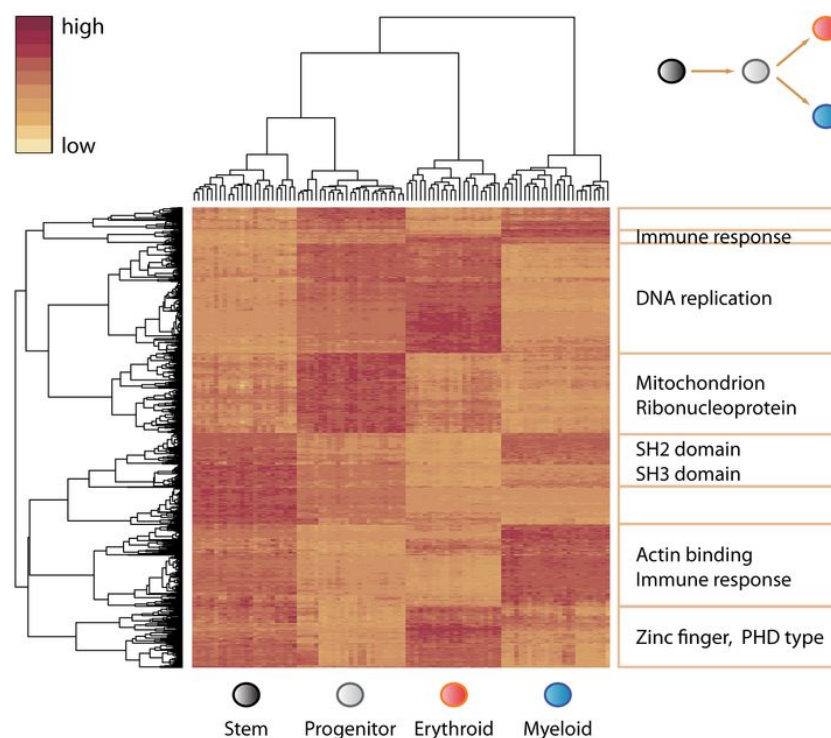
We describe a genetic analysis of variation in gene expression across four functionally distinct, but developmentally related hematopoietic cell populations. Our data reveal complex cell-stage specific patterns of heritable variation in transcript abundance, demonstrating the plasticity of gene regulation during hematopoietic cell differentiation.

## 6.2 Results

### 6.2.1 Genetic regulation of gene expression

We evaluated genome-wide RNA transcript expression levels in purified Lin<sup>-</sup>Sca-1<sup>+</sup>c-Kit<sup>+</sup> multi-lineage cells, committed Lin<sup>-</sup>Sca-1<sup>-</sup>c-Kit<sup>+</sup> progenitor cells, erythroid TER-119<sup>+</sup> cells, and myeloid Gr-1<sup>+</sup> cells, isolated from the bone marrow of ~25 genetically related and fully genotyped BXD – C57BL/6 (B6) X DBA/2 (D2) – recombinant inbred mouse strains (Peirce *et al.* 2004). In this study, we exploit the fact that the purified cell populations are closely related, sometimes just a few cell divisions apart on the hematopoietic trajectory. The Lin<sup>-</sup>Sca-1<sup>+</sup>c-Kit<sup>+</sup> cell population contains all stem cells with long-term repopulating ability, but also includes multipotent progenitors that still have lymphoid potential. Although long-term repopulating stem cells are known to only make up a fraction of the Lin<sup>-</sup>Sca-1<sup>+</sup>c-Kit<sup>+</sup> population, for simplicity we will refer to this population as stem cells. The

$\text{Lin}^- \text{Sca-1}^- \text{c-Kit}^+$  cell population does not contain stem cells and lymphoid precursors, but does include common progenitors of the myeloid and erythroid lineages (Bryder *et al.* 2006). Finally, TER-119<sup>+</sup> cells and Gr-1<sup>+</sup> cells are fully committed to the erythroid and myeloid lineages, respectively. Unsupervised clustering of the most varying transcripts demonstrated that each of the four cell populations could easily be recognized based on expression patterns across all four cell types (Figure 6.1 and Table S1).



**Figure 6.1:** Mean expression levels for all probes in the four cell types. Unsupervised clustering including all probes for the 96 RNA samples follows cell-type (top hierarchical tree), while clustering of the 876 most varying probes reveals distinct categories of genes that show cell-type-specific expression (left hierarchical tree). The heat map shows the expression patterns of those probes and selected enriched gene categories in each major cluster. Discriminatory genes are enriched in various functional classes, including SH2/SH3 domain containing transcription factors for stem cells, mitochondrial genes for progenitor cells, genes involved in DNA replication and zinc fingers for erythroid cells, and immunoglobulin type genes for myeloid cells (all  $p$ -values  $< 0.05$ ). For genes that belong to each of these clusters, see Table S1.

We observed strong and biologically significant variation in gene expression during hematopoietic differentiation, independent of mouse strain. However, the genetical genomics strategy, in which we focus on inter-strain gene expression differences, allows for a far more comprehensive understanding of the genetic regulatory links underlying this variation. QTL mapping of gene expression traits allows us to identify eQTLs; genomic regions that have a regulatory effect on those expression traits. Two types of eQTLs can be distinguished, i.e., those that map near (less than 10 Mb from) the gene which encodes the transcript (*local*) and those that map elsewhere in the genome (*distant*) (Rockman and Kruglyak 2006). Together, *local* and *distant* eQTLs constitute a genome-wide overview of the gene regulatory networks that are active in the cell type under study. The strongest eQTLs were found for genes that were expressed only in mouse strains carrying one specific parental allele, suggesting that *local* regulatory elements are distinct between the two alleles. Cases of such allele-specific expression included *H2-Ob* and *Apobec3*. These transcripts were only detectable in strains that carried the B6 allele of the gene (see Figures S1a-b). A global view of heritable variation in gene expression indicated that the strongest eQTLs are not associated with the most highly expressed genes, and that for most probes the expression difference between the B6 and D2 alleles is small (see Figures S1c-d).

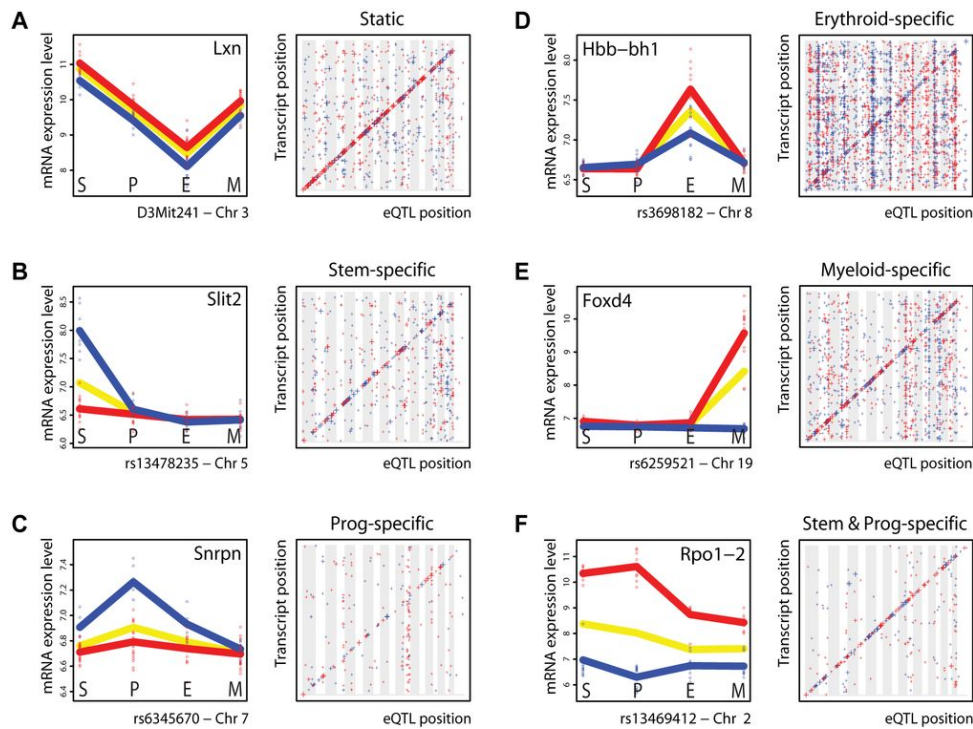
Since the focus of this project is to study the influence of cellular differentiation state on regulatory links, we used ANOVA to distinguish between “static” eQTLs that show consistent genetic effects across the four cell types and “dynamic” eQTLs that are sensitive to cellular state (i.e., eQTLs that have a statistically significant genotype-by-cell-type interaction). We further partitioned dynamic eQTLs into different categories on the basis of their dynamics along the differentiation trajectory.

### 6.2.2 Cell-type-independent static eQTLs

The first eQTL category comprises genes that have static eQTLs across all four cell types under study. Variation in *Lxn* expression is shown as a representative example (Figure 6.2a, left panel). *Lxn* expression has previously been shown to be higher in B6 stem cells compared to D2 stem cells, and to be negatively correlated with stem cell numbers (Liang *et al.* 2007). In our dataset *Lxn* showed clear expression dynamics (it was most highly expressed in stem cells), and was indeed more strongly expressed in cells carrying the B6 allele, but the expression difference between mice carrying the B6 or D2 allele remained constant across all cell types.

In total, we identified 365 probes that displayed a static eQTL at threshold  $p <$





**Figure 6.2:** Identification of static and dynamic eQTLs.

(A) Genome-wide identification of cell-type-independent static eQTLs. (Left panel) *Lxn* mRNA levels were analyzed in all 4 cell types. Each circle represents an individual sample (strain). The yellow line shows mean expression levels across all strains. The red and blue lines indicate mean *Lxn* expression levels in strains that carry the B6 or D2 *Lxn* allele, respectively. The genetic effect of parental alleles on *Lxn* expression levels was consistent in all cell types. (Right panel) Individual probes that detected a transcript that was consistently controlled by the same eQTL in all 4 cell types. The y-axis indicates the physical position of the encoding gene, the x-axis provides the genomic position of the marker with strongest linkage statistics. Vertical gray and white bandings indicate different chromosomes, ranging from chromosome 1 to X. The size of each symbol reflects the strength of the genetic association: eQTLs with  $p$ -values  $< 10^{-8}$  are represented by the largest crosses,  $p$ -values between  $10^{-6}$  and  $10^{-8}$  are shown with medium crosses, while small crosses refer to eQTLs with  $p$ -values between  $10^{-4}$  and  $10^{-6}$ . The color coding (red and blue) indicates the parental allele of the eQTL that caused a higher gene expression (B6 is red and D2 is blue).

(B-E) Genome-wide identification of transcripts that are controlled by cell-type-specific eQTLs. (Left panels) Expression data for some transcripts that were affected by cell-type-specific eQTLs (B: *Slit2* in stem cells, C: *Snrpn* in progenitor cells, D: *Hbb-bh1* in erythroid cells and E: *Foxd4* in myeloid cells). (Right panels) Genome-wide distribution of eQTLs that were preferentially/uniquely detected in each of the four cell populations.

(F) Transcripts that were controlled by eQTLs in both stem and progenitor cells. An example is *Rpo1-2*. Full lists of all genes belonging to the eQTL (sub)categories shown here are provided in Table S2.

$10^{-6}$  (FDR = 0.02). Among the 268 *locally*-regulated probes in this category was *H2-D1*. The histocompatibility gene *H2-D1* is known to be polymorphic between B6 and D2 mice, and would therefore be expected to be in the static eQTL category. The remaining 97 probes mapped to *distant* eQTLs, i.e., their heritable expression variation was affected by the same *distant* locus in all four cell types (Table 6.1).

**Table 6.1:** Overview of static and dynamic eQTLs ( $p < 10^{-6}$ ): number of probes and the number of associated markers.

eQTL category	eQTL subcategory		#probes	#markers	#probes/#markers
<i>Static</i>	All	<i>Local</i>	268	161	1.66
		<i>Distant</i>	97	76	1.28
		Total	365	213	1.71
<i>Dynamic</i>	All	<i>Local</i>	642	282	2.88
		<i>Distant</i>	641	276	2.32
		Total	1283	445	2.88
	Stem-specific	<i>Local</i>	87	66	1.32
		<i>Distant</i>	53	42	1.26
		Total	140	105	1.33
	Progenitor-specific	<i>Local</i>	32	27	1.19
		<i>Distant</i>	13	12	1.08
		Total	45	39	1.15
	Erythroid-specific	<i>Local</i>	131	90	1.46
		<i>Distant</i>	400	164	2.44
		Total	531	223	2.38
	Myeloid-specific	<i>Local</i>	163	121	1.35
		<i>Distant</i>	132	72	1.83
		Total	295	179	1.65

All probes that belonged to the static eQTL category are graphically depicted in an eQTL dot plot displaying the genomic positions of the eQTLs compared to the genomic positions of the genes by which the variably expressed transcripts were encoded (Figure 6.2a, right panel). Whereas in this plot *local* eQTLs appear on the diagonal, *distant* eQTLs appear elsewhere. In general, as has been reported before in eQTL studies, transcripts that were *locally* regulated showed strong linkage statistics. Not surprisingly, the statistical association between genotype and variation in transcript abundance for those transcripts that were controlled by *distant* loci was weaker. These genes are likely to be controlled by multiple loci, each contributing

only partially to the phenotype, thereby limiting their detection and validation in the current experimental sample size. A list of all transcripts with significant static eQTLs is provided in Table S2.

### 6.2.3 Cell-type-dependent dynamic eQTLs

The second eQTL category comprises genes that have dynamic eQTLs across all four cell types under study. In total, we identified 1283 eQTLs ( $p < 10^{-6}$ , FDR = 0.021) that showed different genetic effects in different cell types, indicating that eQTLs are highly sensitive to cellular differentiation state (Table 6.1). Within this dynamic eQTL category, the first four subcategories are composed of eQTLs that were preferentially active in only one of the four cell types we analyzed (Figures 6.2b-e).

For example, *Slit2* mapped to a strong eQTL that was active only in stem cells. *Slit2* mRNA was only detected in the most primitive hematopoietic cell compartment in those BXD strains that carried the D2 allele at rs13478235, a SNP that mapped 629 kb away from the *Slit2* gene (Figure 6.2b, left panel). *Slit2* encodes an excreted chemorepellent molecule that is known to be expressed in embryonic stem cells (Katoh and Katoh 2005), to be involved in neurogenesis (Wang *et al.* 1999) and angiogenesis (Wang *et al.* 2003a), and to inhibit leukocyte chemotaxis (Wu *et al.* 2001). We found a total of 140 genes that have eQTLs that are preferentially/selectively active in stem cells (Figure 6.2b, right panel, largest symbols, Table 6.1). These 140 genes included well-known candidate stem cell genes such as *Angpt1*, *Ephb2*, *Ephb4*, *Foxa3*, *Fzd6*, and *Hoxb5*. Interestingly, many transcripts with as yet unknown (stem cell) function were transcriptionally affected by stem-cell-specific eQTLs. Candidate novel stem cell genes include *Msh5*, and *Trim47*, in addition to a large collection of completely unannotated transcripts.

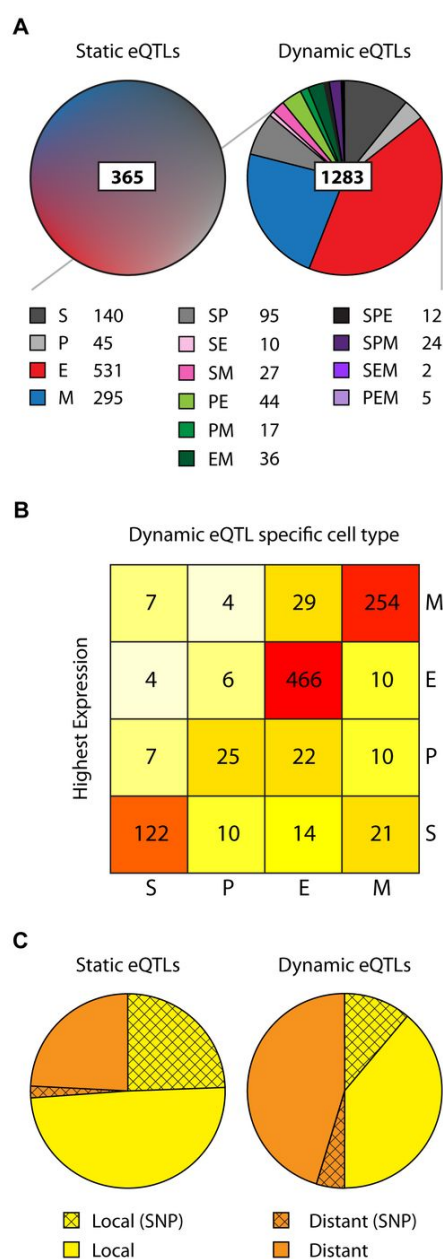
A total of 45, 531, and 295 eQTLs were found to be preferentially/selectively active in progenitors, erythroid cells, and myeloid cells, respectively (Table 6.1). Very distinct patterns of cell-type-specific gene regulation emerged when these eQTLs were visualized in genome-wide dot plots (Figures 6.2c-e). Using genome-wide  $p$ -value thresholds of  $p < 10^{-6}$ , we identified 53 *distantly*-regulated transcripts in stem cells, 13 in progenitor cells, 400 in erythroid cells, and 132 in myeloid cells. In erythroid and myeloid cells most of these transcripts mapped to relatively few genomic loci; these trans-bands are statistically significant, as assessed by a permutation approach taking expression correlation into account (see Methods) (Breitling *et al.* 2008a). Typically, transcripts mapping to a common marker showed a directional bias towards either B6 or D2 expression patterns.

In addition to the relatively simple eQTL dynamics that we have thus far illustrated, more complex eQTL dynamics were also detected using this approach. For example, *Rpo1-2* is a transcript that shows a strong *local* eQTL in the two non-committed lineages included in our study, but shows a much weaker genetic effect in erythroid and myeloid cells (Figure 6.2f). Whereas in mice carrying the B6 allele of *Rpo1-2* the overall expression of the gene decreased substantially during differentiation of progenitor to erythroid cells, in mice carrying the D2 allele expression slightly increased. This observation hints at complex regulatory mechanisms underlying the expression of this gene. Full lists of genes in each dynamic eQTL subcategory described thus far are supplied in Table S2. Additional subcategories and their exact definitions are explained more extensively in the Methods section, and complete results of all dynamic eQTLs are available in Table S3.

#### 6.2.4 Detailed analysis of static and dynamic eQTLs

eQTL dynamics can be caused by transcription factors being switched on/off upon cellular differentiation, or by a transcription factor showing changed specificity due to variations in regulatory input. We found that most (>75%) of the dynamic eQTLs are active in only one of the four cell types under study (Figure 6.3a). A more detailed analysis revealed that in the majority of cases the genes with a cell-type-specific eQTL were also most highly expressed in that particular cell type (Figure 6.3b). Next, we explored whether we could find transcripts that were regulated by distinct eQTLs in different cell types (see Methods). Such eQTL “swapping” would indicate major changes in transcriptional regulation networks. We could find no evidence for such cases. However, given our limited population size we have a low power to detect multiple eQTLs, so swapping eQTLs may still exist but remain undetected in our experimental setting.

It has been described that not all *local* eQTLs in genetical genomics experiments reflect actual expression differences between mouse strains, but rather indicate differential hybridization caused by polymorphisms in the sequences recognized by the probes (Alberts *et al.* 2007). For this reason, we divided both the static and dynamic eQTL categories in *local* and *distant* eQTLs, and indicated the number of probes that hybridized to sequences that are known to contain polymorphisms (Figure 6.3c). As expected, the static eQTL category contained a higher number of such potential false *local* eQTLs. If these false positive eQTLs could be removed, the relative abundance of dynamic eQTLs would be higher, indicating that our study may even conservatively underestimate the level of eQTL dynamics.



**Figure 6.3:** Quantitative overview of static and dynamic eQTLs. (A) Pie charts presenting all 365 static and 1283 dynamic eQTLs that were detected with  $p < 10^{-6}$ . Dynamic eQTLs are subdivided in all 14 categories of interaction eQTLs. (B) Matrix showing the four cell-type-dependent dynamic eQTL categories and the cell type in which the gene was expressed most highly. (C) All static and dynamic eQTLs are subdivided in local and distant eQTLs. Shown is which number of eQTLs was detected by Illumina probes that hybridize to sequences that are known to contain polymorphisms (SNPs) between the two parental strains.

### 6.3 Discussion

We found that many eQTLs are highly sensitive to the developmental state of the cell population under study. Even when the purified cells were only separated by a few cell divisions, eQTLs demonstrated a remarkable plasticity. Furthermore, we provide evidence that the cell-stage-sensitivity of eQTLs is often intertwined with gene expression variation during development. We did not identify target genes that were regulated by distinct eQTLs in different cell types, suggesting that large-scale changes within transcriptional regulation networks are not common.

The fact that eQTLs appear to be highly cell-type-dependent highlights the importance of using well-characterized purified cell types in eQTL studies. In particular, eQTL studies of physiological or disease processes (Schadt *et al.* 2005, Goring *et al.* 2007, Emilsson *et al.* 2008, Chen *et al.* 2008) should target the relevant cell type as precisely as possible, i.e. they should use cells or tissues directly involved in the patho-physiological process. This could even mean that several different cell types need to be separately studied, in particular if developmental trajectories are affected (Li *et al.* 2008). Using unfractionated bone marrow cells, we would have missed many of the diverse and dynamic patterns that we uncovered here, both at the expression level and at the genetic regulatory level. Even so, the four cell populations that we studied are still heterogeneous and further subfractionation of these populations based on different sets of markers would have resulted in even more precise regulatory maps.

Many genetical genomics experiments have used highly heterogeneous samples, in which mRNA from a variety of different cell types was pooled (Chesler *et al.* 2005, Hubner *et al.* 2005, Li *et al.* 2006b, West *et al.* 2006, Keurentjes *et al.* 2007, Whiteley *et al.* 2008). In such mixed samples it is usually impossible to ensure that the contribution of individual cell types to the mixture is the same across samples. As a result, important parts of the variation in gene expression could arise from different sample compositions. For example, if in whole brain samples a heritable morphological or developmental trait leads to an increased size of some brain regions, this can cause apparent hotspots for transcripts that are specific for those particular regions. Our data provide a valuable tool for studying the exact consequences of sample heterogeneity on eQTL mapping: a further study could simulate a collection of samples made of computed mixtures of different hematopoietic cells in defined proportions. Clearly, cell purification strategies are essential to identify those cell-type-specific eQTLs that would otherwise be “masked” in heterogeneous cell populations. Therefore, future genetical genomics studies should be realized on as many cell types or cellular differentiation states as possible, and ideally even on the scale of individual

cells.

All data presented in this chapter were deposited in the online database GeneNetwork ([www.genenetwork.org](http://www.genenetwork.org)), an open web resource that contains genotypic, gene expression, and phenotypic data from several genetic reference populations of multiple species (e.g. mouse, rat and human) and various cell types and tissues (Chesler *et al.* 2004, Wang *et al.* 2003b). It provides a valuable tool to integrate gene networks and phenotypic traits, and also allows cross-cell type and cross-species comparative gene expression and eQTL analyses. Our data can aid in the identification of candidate modulators of gene expression and/or phenotypic traits (Gerrits *et al.* 2008), and as such can serve as a starting point for hypothesis-driven research in the fields of stem cell biology and hematology.

## 6.4 Materials and methods

### 6.4.1 Ethics statement

All animal experiments were approved by the Groningen University Animal Care Committee.

### 6.4.2 Recombinant inbred mice

Female BXD recombinant inbred mice were originally purchased from The Jackson Laboratory and housed under clean conventional conditions. Mice were used between 3 and 4 months of age.

### 6.4.3 Cell purification

Bone marrow cells were flushed from the femurs and tibias of three mice and pooled. After standard erythrocyte lysis, nucleated cells were stained with either a panel of biotin-conjugated lineage-specific antibodies (containing antibodies to CD3e, CD11b (Mac1), CD45R/ B220, Gr-1 (Ly-6G and Ly-6C) and TER-119 (Ly-76)), fluorescein isothiocyanate (FITC)-conjugated antibody to Sca-1 and allophycocyanin (APC)-conjugated antibody to c-Kit, or with biotin-conjugated TER-119 antibody and FITC-conjugated antibody to Gr-1. After being washed, cells were incubated with streptavidin-phycoerythrin (PE) (all antibodies were purchased from Pharmingen). Cells were purified using a MoFlo flowcytometer (BeckmanCoulter) and were immediately collected in RNA lysis buffer. Lineage-depleted (Lin-) bone marrow cells were defined as the 5% of cells showing the least PE intensity.

#### 6.4.4 RNA isolation and Illumina microarrays

Total RNA was isolated using the RNeasy Mini kit (Qiagen) in accordance with the manufacturer's protocol. RNA concentration was measured using a Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies). The RNA quality and integrity was determined using Lab-on-Chip analysis on an Agilent 2100 Bioanalyzer (Agilent Technologies). Biotinylated cRNA was prepared using the Illumina Total-Prep RNA Amplification Kit (Ambion) according to the manufacturer's specifications starting with 100 ng total RNA. Per sample, 1.5  $\mu$ g of cRNA was used to hybridize to Sentrix Mouse-6 BeadChips (Illumina). Hybridization and washing were performed by ServiceXS according to the Illumina standard assay procedures. Scanning was carried out on the Illumina BeadStation 500. Image analysis and extraction of raw expression data were performed with Illumina Beadstudio v2.3 Gene Expression software with default settings and no normalization. The raw expression data from all four cell types were first log<sub>2</sub> transformed and then quantile normalized as a single group.

#### 6.4.5 Clustering of genes

For cluster analysis we retained only genes having a minimal fold change of 2 (difference of 1 in log<sub>2</sub> scale) in either direction in mean expression on the transition from Lin<sup>-</sup>Sca-1<sup>+</sup>c-Kit<sup>+</sup> to Lin<sup>-</sup>Sca-1<sup>-</sup>c-Kit<sup>+</sup> and on the transition from Lin<sup>-</sup>Sca-1<sup>-</sup>c-Kit<sup>+</sup> to TER-119<sup>+</sup> or to Gr-1<sup>+</sup>. This filter reduced the dataset to 876 probes. We then computed the distance matrix for this group of probes, using the absolute Pearson correlation. Using this distance matrix, we applied the hierarchical clustering algorithm. From the resulting tree, 8 different clusters emerged from a manually chosen threshold. We then submitted each of these clusters to DAVID to identify enriched functional annotations (Dennis *et al.* 2003).

#### 6.4.6 Full ANOVA model for eQTL mapping

The expression data of the four cell types were firstly corrected for batch effect and then analyzed separately by the following ANOVA model:  $y_i = \mu + Q_i + e_i$  where  $y_i$  is the gene's log intensity on the  $i$ th microarray;  $\mu$  is the mean;  $Q_i$  is the genotype effect under study; and  $e_i$  is the residual error.

Next, expression data of the four cell types were combined and analyzed by a full ANOVA model including the cell type effect (CT) and the eQTL CT interaction effect:  $y_{ij} = \mu + CT_j + Q_i + (Q \times CT)_{ij} + e_{ij}$  where  $y_{ij}$  is the gene's log intensity at the



$i$ th microarray ( $i = 1, \dots, n$ ) and  $j$ th cell type;  $CT_j$  is the  $j$ th cell type effect;  $(Q \times CT)_{ij}$  is the interaction effect between the  $i$ th eQTL genotype and  $j$ th cell type, and  $e_{ij}$  is the residual error. The batch effect was included as one of the factors. For each probe, we performed a genome-wide linkage analysis to identify the two markers that showed the most significant main QTL effect and interaction effect, respectively.

#### 6.4.7 Local and distant eQTLs

We defined an eQTL as *local* if it was located within less than 10 Mb from the gene. All other eQTLs were considered *distant*.

#### 6.4.8 Classification of eQTLs

The ANOVA yields significance  $p$ -values for the main QTL effect  $Q_i$  and the interaction effect  $(Q \times CT)_{ij}$  for each probe at each marker. A small  $p$ -value for the interaction effect indicates that the eQTL effect is different between the cell types. This significant difference can be due to very diverse patterns, with different biological interpretations. It is therefore necessary to classify interaction eQTLs based on these patterns. To achieve this classification, for every interaction eQTL we evaluated the strength of the effect in each cell type by calculating the difference between the mean expression of both genotypes. The cell type for which the effect was the strongest was labeled "High". The cell type whose effect was most different from the strongest effect was labeled "Low". The remaining two cell types were assigned to the group they resembled most closely. This classification allowed us to define 14 categories of interaction eQTLs. Additionally, we identified eQTLs that have a consistent effect across all four cell types. This category of consistent eQTLs consists of all probes satisfying the following three conditions: the gene has a significant main effect  $Q_i$  at marker  $m$ ; for the same marker  $m$ , the interaction  $(Q \times CT)_{ij}$  is not significant; the mean eQTL effect across cell types has a coefficient of variation smaller than 0.3.

#### 6.4.9 Estimating the FDR for the main QTL effect

We permuted the strain labels in the genotype data 100 times, maintaining the correlation of expression traits while destroying any genetic association. Then we applied the full ANOVA model and stored the genome-wide minimum  $p$ -value for each transcript. Based on the resulting empirical distribution of  $p$ -values, we estimated that a threshold of  $-\log_{10}p = 6$  corresponds to a false discovery rate (Storey

and Tibshirani 2003) of 0.02 for the main QTL effect. The 99.9th percentile of the number of significant eQTLs per marker (i.e., the minimum size of statistically significant “eQTL hotspots”) is 28.

#### 6.4.10 Estimating the FDR for interaction QTL effect

We estimated the residuals of the full ANOVA model after fitting all factors up to the main QTL effect at each marker for each transcript (Anderson and ter Braak 2003). Then we permuted the strain labels and applied the ANOVA model  $y_{ij} = Q_i + CT_j + Q \times CT_{ij} + e_{ij}$  to the permuted residuals at each marker for each transcript and stored the genome-wide minimum  $p$ -value. Based on 100 permutations and the resulting empirical distribution of  $p$ -values, we estimated that a threshold of  $-\log_{10}p = 6$  corresponds to a false discovery rate of 0.021 for interacting QTL effect. The 99.9th percentile of the number of significant eQTLs per marker (i.e., the minimum size of statistically significant “interaction hotspots”) is 8.

#### 6.4.11 Detection of swapping eQTLs

Swapping eQTLs are those transcripts that show one eQTL in one cell type, but another eQTL in another cell type. From the full model mapping described above, we obtained 1283 transcripts with a significant interaction effect between genotype (first marker) and cell type. After taking into account the genetic and interaction effects of the first marker, we scanned the genome excluding the region of the first marker (window size = 30cM) and tested if there was a significant interaction effect between genotype and cell type and whether this new interaction effect was classified in a different cell type category (see above Classification of eQTLs), which would indicate a swapping eQTL.

This means, for each transcript, a two-marker full model mapping was applied using the following model:

$$y_{ij} = \mu + CT_j + Q_i^* + (Q^* \times CT)_{ij} + Q_i + (Q \times CT)_{ij} + Q_i^* Q_i + e_{ij}$$

where  $y_{ij}$  is the gene’s log intensity at the  $i$ th microarray ( $i = 1, \dots, n$ ) and  $j$ th cell type;  $CT_j$  is the  $j$ th cell type effect;  $Q_i^*$  and  $(Q^*CT)_{ij}$  are the main genotype effect at first marker and interaction effect between cell type and the genotype effect at this marker, where the first marker is defined as the marker with maximal interaction effect from previous one-marker full model mapping;  $Q_i$  is the genotype effect of the second marker;  $(Q \times CT)_{ij}$  is the interaction effect between the  $i$ th genotype and  $j$ th cell type,  $Q_i^* Q_i$  is the epistasis effect and  $e_{ij}$  is the residual error.

### 6.4.12 URLs

All raw data were deposited at GEO (<http://www.ncbi.nlm.nih.gov/geo/>).  
All processed data presented in this chapter were deposited at GeneNetwork ([www.genenetwork.org](http://www.genenetwork.org)) (Chesler *et al.* 2004, Wang *et al.* 2003b).

## 6.5 Acknowledgments

We thank Guus Smit and Sabine Spijker for providing BXD mice, Geert Mesander and Henk Moes for assistance in cell sorting, and Arthur Centeno and Rob W. Williams for depositing our data in [www.genenetwork.org](http://www.genenetwork.org).



## Chapter 7

# Global genetic robustness of the alternative splicing machinery in *C. elegans*

### ABSTRACT

*Alternative splicing is considered a major mechanism for creating multicellular diversity from a limited repertoire of genes. Different isoforms can be produced at the same time in the same cell type and their ratios can be the same or different between divergent genotypes. Here, we performed the first study of genetic variation controlling alternative splicing patterns by comprehensively identifying quantitative trait loci affecting the differential expression of transcript isoforms in a large recombinant inbred population of C. elegans, using a new generation of whole-genome very-high-density oligonucleotide microarrays. These arrays provide 3 million measured intensity values for each sample, which allowed us to detect heritable differences in gene expression with exquisite sensitivity and resolution. Using 60 experimental lines, we were able to detect 435 genes with substantial heritable variation for at least one exon, of which 36% were regulated at a distance (in trans). Nonetheless, we find only a very small number of examples of heritable variation in alternative splicing (22 transcripts), and most of these genes co-localize with the associated genomic loci. Our findings suggest that the regulatory mechanism of alternative splicing in C. elegans is robust towards genetic variation at the genome-wide scale. This is in striking contrast to earlier observations in humans, which showed much less genetic robustness.*

## 7.1 Introduction

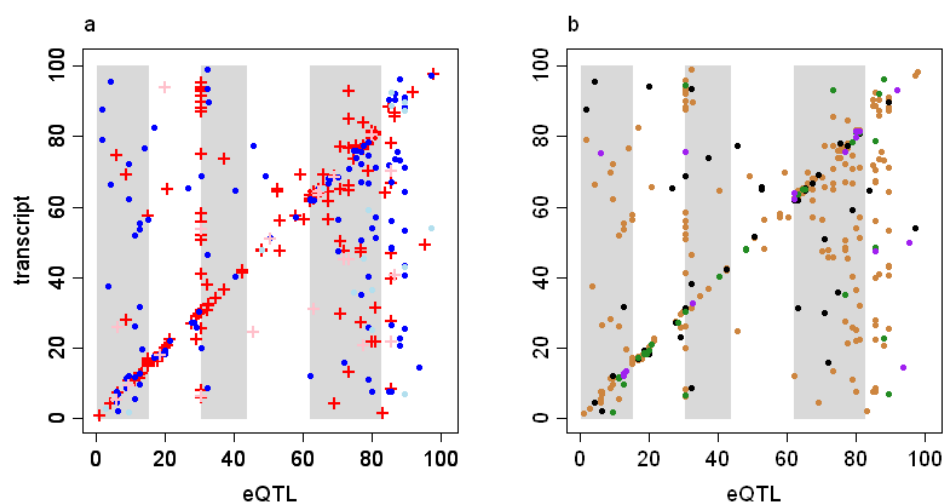
Alternative splicing of pre-mRNAs is part of gene regulation and a major mechanism for increasing the protein repertoire and the resulting phenotypic diversity. Recently, in individual cases variations in number and ratio of splice variants have also been found in *C. elegans* in different developmental stages (Barberan-Soler and Zahler 2008b), tissues (Kuroyanagi *et al.* 2007) and genotypes (Fischer *et al.* 2008). However, the much smaller number of alternative splicing (Kim *et al.* 2007), and the strong evolutionary conservation of splicing events in *C. elegans* (Barberan-

Soler and Zahler 2008a) have been interpreted as signifying a fundamental difference in the way that worms and vertebrates generate diversity from their genetic information. The relative rarity of splicing and high degree of stabilizing selection are seen as having parallels in the limited cellular complexity and highly conserved, rigid developmental programs (Zhao *et al.* 2008) in worms compared to humans. If this is a general trend, and not restricted to just individual cases of splicing, the conservation of splicing patterns should be reflected at the whole-genome level.

In this chapter we explore this question by extending the genetical genomics strategy (Jansen and Nap 2001) to the characterization of the genetic factors contributing to variations in alternative splicing in 60 recombinant inbred (RI) *C. elegans* strains. This powerful new strategy, also known as expression genetics (Schadt *et al.* 2003), has emerged in recent years as a versatile tool to study the genetic basis of gene expression by integrating transcriptomics and classical quantitative genetics (Mackay *et al.* 2009). In this approach, molecular profiling on a large population of densely genotyped individuals is used to map genomic loci that modulate gene expression. This leads to the identification of expression quantitative trait loci (eQTLs), i.e. polymorphic genetic loci that cause heritable differences in mRNA concentration. eQTLs that are found in close vicinity of the transcript-encoding gene are called local or cis eQTL, while influential loci elsewhere in the genome are known as distant or trans eQTL. Using high-resolution tiling microarray we were able to extend this concept to the detection of genetic determinants of alternative splicing, so-called asQTLs, and to the detailed quantification of the genetic robustness of the alternative splicing machinery in *C. elegans* on a genome-wide scale.

## 7.2 Results

Here, we performed the first genome-wide analysis of genetic variation of alternative splicing in *C. elegans* using a comprehensive tiling microarray. We used 60 recombinant inbred lines of a cross between two very diverse strains, Bristol (N2) and Hawaii (CB4856), which have been genotyped using 121 markers (Li *et al.* 2006b). By using tiling array data, with multiple probes targeting every exon of each gene, we obtained a more comprehensive and sensitive picture of heritable variation of gene expression than possible with previous technologies. It also allows us to dissect the genetic component for differences in isoform-specific gene expression. Thus we can detect alternative splicing quantitative trait loci (asQTL), the genome region controlling variation in isoform-specific expression. Two categories of asQTLs can be



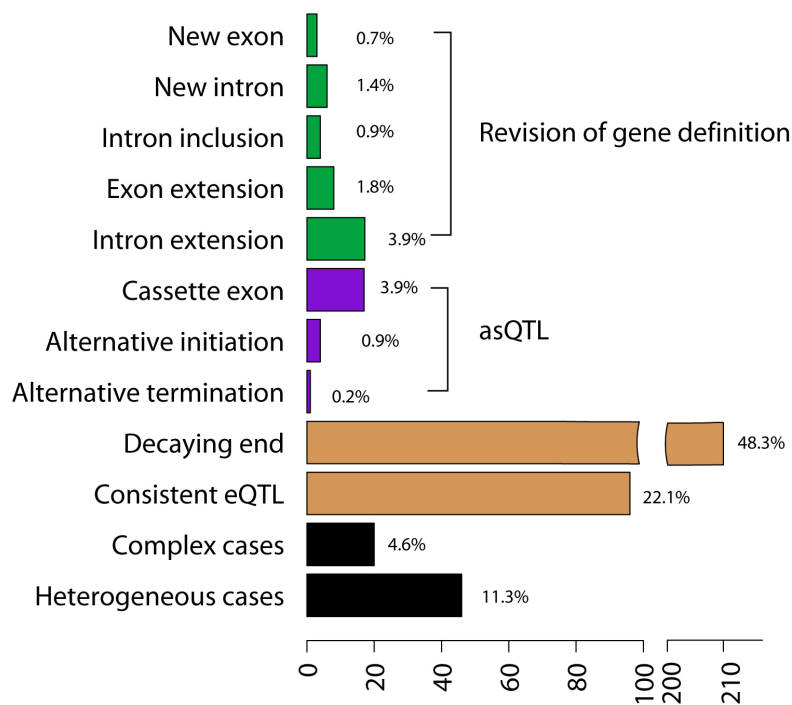
**Figure 7.1:** Mapping location (a) and type (b) of heritable variation in gene expression.

(a) Each dot in the figure represents a single transcript. The physical position of each transcript is indicated on the y-axis, and the position of the locus that is most strongly associated with variation of the corresponding transcript level is shown on the x-axis. Transcripts on and off the diagonal are locally- and distantly-regulated, respectively. The different symbols/colors discriminate the parental allele of the eQTL that caused a higher expression (N2 is indicated by a red cross and CB4856 by a blue dot). Transcripts that overlap with another gene on the genome according to the WormBase genome annotation are shown in pink ((N2>CB4856)) and light blue (N2<CB4856).

(b) The same eQTL with a different coloring scheme: colors discriminate consistent eQTL (brown) and asQTL (purple). The relative rarity of the latter category is clearly visible. The few cases that are observed are mostly restricted to the cis diagonal, i.e. they are caused by local variation of the gene sequence close to the affected splice site. Variation in trans-acting splicing factors seems to be extremely rare. Transcripts with revised gene definition are indicated in green. Genes with complex and heterogeneous eQTL patterns are shown in black.

distinguished, i.e. those that map in close vicinity to the gene itself (*local*) and those that map elsewhere in the genome (*distant*). Local-activity can be explained, for example, by altered functional motifs in exonic splicing enhancers that will affect the splicing activity. The mechanism of *distant* regulation is often more complicated and can possibly be explained by a polymorphism in an auxiliary splicing factor (e.g. SR protein) that modulate the activity of the spliceosome. In this case we would ex-

pect to see a genetic master regulator at the locus of the splicing factor controlling isoform ratios for large groups of transcripts.



**Figure 7.2:** Classification of genes showing heritable expression variation (eQTL)

The 435 transcripts were classified into different groups according to their eQTL pattern: Consistent eQTLs (brown) lead to consistent expression differences in all exons of a gene, non-consistent ones are indicating the need for revised gene annotations (green; 8.7%) or potential heritable differences in splicing (i.e. transcripts with alternative splicing QTL)(Kwan *et al.* 2008). The latter (purple) are subdivided in three classes according to the position of the alternatively spliced exon; they comprise a total of 5% of all cases, compared to 55% out of total 324 transcripts with significant eQTL that showed heritable isoform changes in humans (Kwan *et al.* 2008). Complex cases (black) contain indications for multiple event types, e.g. various exons with different patterns of heritable difference. Some cases (10.6%) show very heterogeneous eQTL patterns across probes and exons.

Using a nonparametric effect size testing, corrected for genotype imbalance (Materials and Methods) and corresponding to a  $p$ -value of 0.001 (Wilcoxon's test), we detected 435 genes with substantial heritable variation for at least one exon. The comparison of gene position and associated polymorphisms shows that most eQTLs



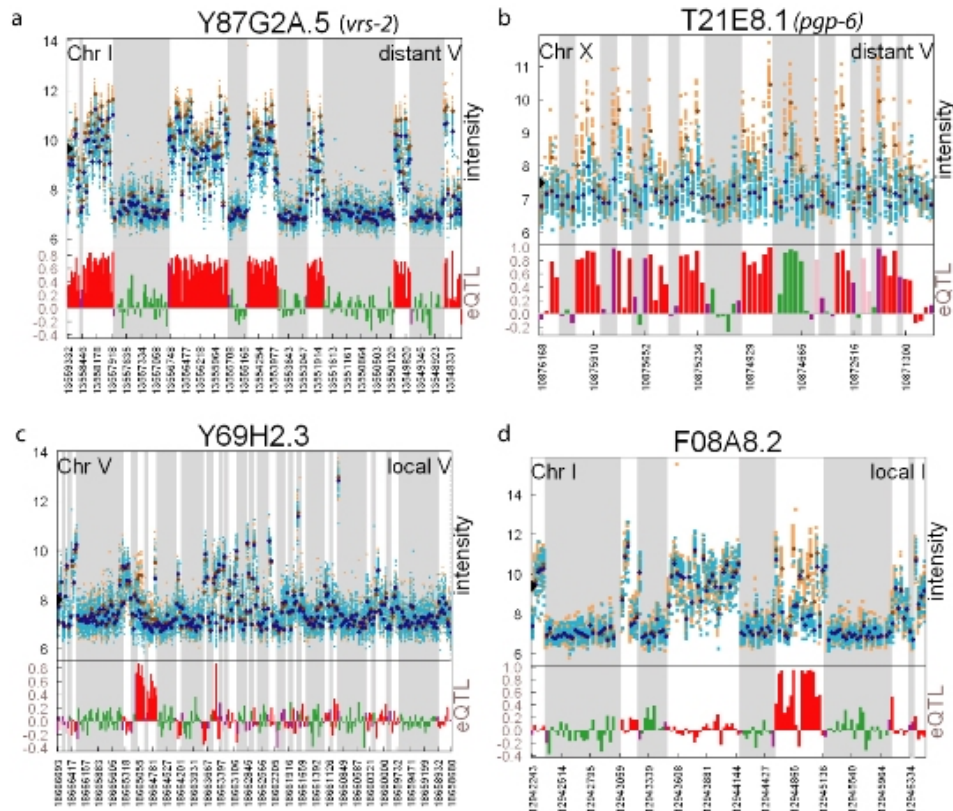
map in close proximity to the affected gene (*local* eQTL; 277 genes or 64%; Figure 7.1). There are 158 eQTLs mapping to another chromosome (*distant* eQTL). 267 genes show higher expression in carriers of the N2 allele than in CB4856 carriers, including 53 cases of known gene deletions in the CB4856 strain (Maydan *et al.* 2007).

A large majority of eQTLs (319, or 70.4%) lead to a consistent differential expression across all exons of the affected gene. Interestingly, the genetic effects (eQTL size) of these consistent eQTLs shows a strong correlation (Spearman's  $\rho = 0.78$ ) with a previous experiment using cDNA micorarrays (Li *et al.* 2006b). As shown in Figure 7.2, 8.7% of cases show evidence for a necessary refinement of existing gene definitions, predominantly by expanding known exons (plotted results for all genes are available at [www.wormplot.org](http://www.wormplot.org) for detailed examination). In contrast, we find only 22 genes that show evidence for genetic variation of alternative splicing, i.e. an exon-specific asQTL (Figure 7.3). This genome-wide evidence for the genetic robustness of the alternative splicing machinery is consistent with the earlier indication that individual alternative splicing events in *C. elegans* are highly conserved and hardly tolerate genetic variation. Note, however, that variation in alternative splicing events restricted to a specific cell or tissue type can be diluted in measurements on whole-worm mRNA. In addition, 77% of asQTL were found to be *locally* regulated. This agrees with recent findings that alternative splicing can be regulated without involvement of an auxiliary splicing factor, by cis-acting RNA sequences that can function as splicing silencer (Yu *et al.* 2008).

Our results show that most of the reported asQTLs have strong genetic effects (qualitative on-off patterns). We found only few cases of subtle quantitative effects on alternative splicing. Despite the large population used in this study, technical noise and biological variation might limit our ability to detect subtle shifts in isoform proportions. In order to detect more quantitative effects (Figure 7.4), more precise technology such as deep-sequencing would be required. Even then, reliable detection of changes in isoform proportions will depend on extremely large read numbers.

### 7.3 Conclusions

Our genome-wide study provides the first genome-wide evidence supporting earlier hypotheses that in *C. elegans* the alternative splicing machinery exhibits a general genetic robustness, and only a minor fraction of genes shows heritable variation in splicing forms and relative abundance. This observation points to a profound difference in the regulation of the alternative splicing machinery compared to humans,



**Figure 7.3:** Expression intensity and eQTL effect per probe along the genome for selected genes

(a) Detecting consistent heritable differences in gene expression with high resolution. Nearly 300 probes cover the area of this gene, Y87G2A.5. Exon probes show consistently high expression (median intensity=9.64), compared to intron probes. However, there is huge variation between probes, which makes the clear delimitation of exon boundaries challenging. In contrast, there is a clear and highly consistent differential expression between carriers of different alleles (N2 and CB4856). This so-called eQTL effect, indicated by red bars, is highly reproducible across all exon probes within the gene. In this example, the average expression difference between the two alleles is approximately 2.4-fold. In total, there are 306 genes with similar consistent expression differences. It should be noted that a majority of genes show consistently lower intensity (and thus lower eQTL effect) in the 3' untranslated region (UTR) indicating the decaying end of transcript (Kolmogorov-Smirnov test,  $p$  value  $< 2.2 \times 10^{-16}$  ).

(b) Refining existing genome annotations. The exon probes within gene T21E8.1 show consistently higher expression for individuals carrying the N2 allele than the CB4856 allele. Additionally, several adjacent probes within the sixth intron show the same differential expression pattern, suggesting that this intron contains a previously unannotated additional exon. This would not have been detectable based on the absolute expression levels, due to the high inter-probe variability. We find a total of 41 genes which require refined annotation according to the eQTL pattern, mostly extensions of known exons and redefinitions of the transcript start and end sites.

**Figure 7.3:** (c) and (d) Detecting heritable variation in alternative splicing. These genes do not show heritable expression differences in general, but individual exons show consistently lower signal for carriers of the CB4856 allele. This suggests that these exons are specifically removed by alternative splicing in one of the two alleles. In both cases, this alternative splicing variation is determined by a local sequence variation (QTL mapping in cis). The first example (Y69H2.3) has been confirmed experimentally (Barberan-Soler and Zahler 2008a). We find 22 comparable instances of heritable differences in splicing patterns.

which parallels the differences in cellular diversity and developmental flexibility in the two species and has important consequences for interpreting future studies using *C. elegans* as a model organism for metazoan splicing.

## 7.4 Materials and methods

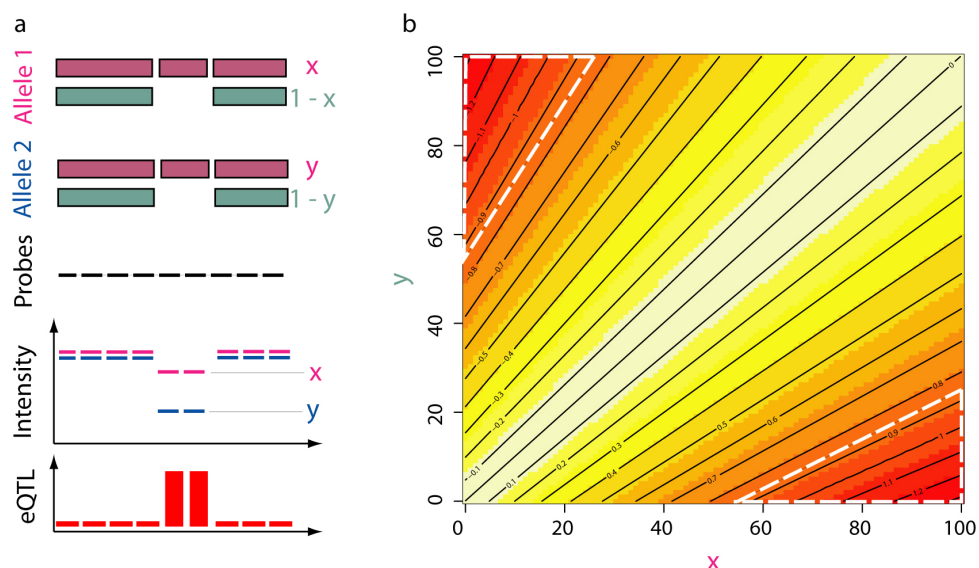
### 7.4.1 Worm samples, genotyping and Affymetrix GeneChips

*C. elegans* recombinant inbred lines were generated and genotyped as described in (Li *et al.* 2006b). mRNA was isolated from 60 RILs reared under standard condition and hybridized to Affymetrix 1.0 *C. elegans* tiling arrays. The hybridization was done by ServiceXS (Leiden, The Netherlands). Since polymorphisms in the probe region can lead to spurious *local* eQTLs (Alberts *et al.* 2007), 80903 probes (out of  $2.9 \times 10^6$  probes on each array) with known SNP (including predicted SNP) were removed for the subsequent analysis. Each probe is annotated as exonic, intronic, or intergenic, when the entire probe of 25bp falls in one of the three regions, respectively. Probes spanning exon-intron boundaries are labeled as boundary probes.

### 7.4.2 Data analysis

#### Preprocessing of raw data

The raw gene expression data from 60 microarrays (one RIL per array) were log transformed and quantile normalized. Subsequently, the normalized intensity data were corrected for batch effects using the following linear model:  $y_i = \mu + B_i + e_i$  where  $y_i$  is the genes intensity on the  $i$ th microarray ( $i = 1, \dots, 60$ );  $\mu$  is the mean;  $B_i$  is the batch effect defined as the date of hybridization and measurement and treated as a categorical variable; and  $e_i$  is the residual error.



**Figure 7.4:** Schematic illustration (a) and power of detection (b) for quantitative changes in alternative splicing.

(a) We consider a transcript with two alternative splicing forms: the second exon is included in isoform 1 but excluded in isoform 2 (cassette exon). Under allele A,  $x\%$  of the entire transcript amount are of isoform 1, while isoform 2 is expressed at  $(1-x)\%$ . Similarly, under allele B, the isoform 1 is expressed at  $y\%$ , and isoform 2 at  $(1-y)\%$ . Without loss of generality, we assume that the total transcript amount is 1, thus the detected signal for the 2nd exon is  $x$  and  $y$  under allele 1 and 2, respectively. The difference between these signals ( $x-y$ ) will be detected as our asQTL effect.

(b) The asQTL effect size changes for different combinations of  $x$  and  $y$ . The white dotted line corresponds to our QTL threshold; changes in transcript isoform ratios outside the dotted line are reliably detectable for the population size used.

### Differential expression between genotypes (eQTL)

We used a robust and powerful statistical approach to associate microarray probe intensity and genotype data in the face of widely different hybridization properties of individual probes. Instead of computing significance of a statistical test, we evaluated a non-parametric effect size (Cliff's delta (Cliff 1996)) for all 3 million probes at each genomic marker. To increase the robustness of the procedure, the median effect size of probes within each exon was taken as expression QTL effect size of this exon.

The raw gene expression data were quantile normalized and corrected for batch effects using a linear model. For each probe on the array we compute the eQTL effect size using Cliff's non-parametric Delta statistic:

$$\Delta = \frac{\#(X_{i1} > X_{i2}) - \#(X_{i1} < X_{i2})}{n_1 \times n_2} \quad (7.1)$$

where  $n_1$  and  $n_2$  are the number of carriers of the N2 and CB4856 allele, and  $\#(X_{i1} > X_{i2})$  is the number of possible pair-wise comparisons where the expression level of gene  $i$  in an N2 carrier is larger than in a CB4856 carrier. The genotype information of the 60 RILs was previously described (Li *et al.* 2006b). For an individual probe, a value of  $\Delta = 0.45$  corresponds to a  $p$ -value = 0.001 in a Wilcoxon rank sum test. As several positions in the genome show a strongly imbalanced genotype ratio (i.e. the number RILs carrying the N2 allele is far larger than the number of RILs carrying the CB4856 allele at a particular locus), the corresponding threshold ( $p$ -value) for each marker at significance level = 0.001 was obtained first, taking the locus-specific imbalance into account. The thresholds of distorted genome regions are expected to be larger than those of balanced marker positions. Subsequently these marker-dependent thresholds were applied to identify significant eQTL.

#### Summarizing the eQTL effect for exons

To increase the robustness of the procedure, the median effect size of probes within each exon was taken as representing the expression QTL effect size of this exon for each genomic marker. Subsequently, the eQTL profile at the marker with maximal summarized eQTL effect was obtained. To achieve a reliable estimate of eQTL effect size, only exons covered by more than 3 probes were considered here. Transcripts with a summarized eQTL effect larger than the threshold for at least one exon were declared as having a significant eQTL and were used for further analysis.

### 7.4.3 Classification of eQTLs pattern

There are 435 transcripts with a significant eQTL for at least one exon were examined in greater details and manually classified as shown in Figure 7.1. By visualizing the intensity level and eQTL size of the entire transcript, we firstly classified transcripts as consistent eQTL if there is an association of majority of probes significant at threshold of  $\Delta = 0.45$ . Consistent eQTLs lead to consistent expression differences in all exons of a gene, non-consistent ones are indicating the need for revised gene annotations or potential heritable differences in splicing (i.e. transcripts with

alternative splicing QTL) (Kwan *et al.* 2008). The former are subdivided into five subcategories: new exons, new introns, intron inclusions, exon extensions and intron extensions. The latter are subdivided in three classes according to the position of the alternatively spliced exon: cassette exon, alternative initiation and alternative termination. Transcripts showing evidence for multiple types of variation, e.g. having various exons with different patterns of heritable difference, were classified as complex cases. Heterogeneous cases contain transcripts showing very diverse eQTL patterns across probes and exons and belonging to none of the above-mentioned categories.

To validate the classification procedure, all classifications were performed independently by two researchers, and inconsistent cases checked in more detail. A complete list of classifications and the corresponding plots is available in the Supplementary Table 1 and the corresponding plots for all genes are available at [www.wormplot.org](http://www.wormplot.org).

#### 7.4.4 Permutation

A permutation approach was used to estimate the empirical false discovery rates for the detection of genetically regulated alternative splicing. We permute sample labels in the genotype matrix and keep the correlation structure between traits and the correlation structure between markers; this makes this empirical procedure perfectly suited to a non-biased estimation of the significance under the multiple dependence properties of the data (Breitling *et al.* 2008a). The permuted data were reanalyzed for all genes at chromosome IV to keep the computational burden within reasonable limits: we repeated the QTL detection and classification as we did for the real data. Based on a total of 67,000 permuted instances of genes, we estimated the false discovery rate for genetically regulated alternative splicing case being <1%.

#### 7.4.5 Deleted genes

We validated our ability to detect heritable expression differences by examining published gene deletions in CB4856 worms (Maydan *et al.* 2007). These genes should show consistently variable expression according to the local genotype. Of 531 CB4856-deleted genes, about 10% (53 genes) are detected as expressed in our experiment. All of these genes show consistent eQTL across all probes with larger expression in N2 allele, well above our threshold. This confirms the sensitivity of our approach.

#### 7.4.6 Comparison with previous experiment

As a further validation step, we compared the detected eQTL to those observed in an earlier study using cDNA microarrays (Li *et al.* 2006b). Nearly half of the top-500 highly expressed genes (231 genes) are shared in the two experiments. The eQTL effect size also shows strong correlation (*locally* regulated QTL:  $r = 0.72$ , *distantly* regulated:  $r = 0.48$ ). Several strong *distant* eQTL were found in both experiments including ZK488.6, F10D2.9 (*fat-7*), 56H6.5 (*gmd-2*), C38D9.2, T21E8.1 (*pgp-6*), C05A9.1 (*pgp-5*), F15D4.5.

#### 7.4.7 Power to detect quantitative changes in alternative splicing

Generally, the genetic effect on the abundance of transcript isoforms can be quantitative rather qualitative (shifts in isoform ratios, rather than on-off effects). We calculated the expected effect size for all possible shifts of isoform ratio, assuming that two isoforms differ only by the presence or absence of one exon, and that there is no overall expression difference (Figure 7.2). It turns out that the difference in abundance of transcript isoforms should be at least about 1.86-fold to be picked up in our study. This means that our method has sufficient power to identify quantitative changes in isoform ratio like 90:10 (allele 1)→20:80 (allele 2) or 60:40 (allele 1)→12:88 (allele 2).

#### 7.4.8 Supporting information

The NCBI Gene Expression omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo>) accession number for the tiling data discussed in the chapter is GSE15778.

### 7.5 Acknowledgments

This work was supported by EU FP7 PANACEA 222936 and the Netherlands Organization for Scientific Research, NWO-86504001.





## Chapter 8

---

# Summarizing discussion

### ABSTRACT

*Genetical genomics integrates data from multiple molecular levels such as the transcriptome, proteome and metabolome by mapping their variation in a population to polymorphic genetic loci. This systems genetics approach is increasingly used to identify molecular traits involved in the pathology of diseases and to elucidate the networks underlying complex phenotypes. Recent studies have pushed the genetical genomics concept further towards data integration and interpretation within and across molecular levels, and have also revealed remaining challenges. The focus of this chapter is to discuss these challenges and their possible solutions in the following three following areas: (1) experimental design, (2) setting significance thresholds, and (3) defining gene and QTL networks. Finally, we explore how future genetical genomics studies might benefit from the advent of new methods that aim at removing large pervasive variation components that are caused by uncontrolled factors in omics datasets.*

## 8.1 Introduction

Genetical genomics (Jansen 2003, Jansen and Nap 2001) uses classical genetics approaches of Quantitative Trait Locus (QTL) mapping to link or associate the variation in traits from multiple molecular levels (such as transcriptomics, proteomics and metabolomics) to genetic loci harboring genotypic polymorphisms. Genetical genomics has become a popular systems genetics strategy (Sieberts and Schadt 2007) for unraveling molecular regulatory networks: a PubMed search on relevant keywords currently yields 191 scientific publications (webCite 2010), 39% of which were published in 2009/10. Pioneering experiments have demonstrated the high heritability of an extensive range of molecular traits (mainly mRNAs but also protein and metabolite abundance as measured with mass spectrometry or nuclear magnetic resonance, see Box 8.1 for a summary of special features of molecular traits) in numerous model species (including yeasts, plants, worms, flies, mice, rats and humans) (Brem *et al.* 2002, Bystrykh *et al.* 2005, Chesler *et al.* 2005, DeCook

*et al.* 2006, Petretto *et al.* 2006, Ruden *et al.* 2009, Schadt *et al.* 2003), and they have exposed the plasticity of the QTLs that control those traits with respect to environmental condition, tissue type or cellular context (Dimas *et al.* 2009, Ge *et al.* 2009, Gerrits *et al.* 2009, Li *et al.* 2006b, Smith and Kruglyak 2008). Genetical genomics studies that integrate 'classical' phenotypes (such as height or disease susceptibility) with multiple traits from molecular levels have improved our understanding of how genetic variation propagates through biological systems (Fu *et al.* 2009) and have suggested molecular pathways through which some genetic variants can cause diseases (Emilsson *et al.* 2008, Moffatt *et al.* 2007, Schadt *et al.* 2008). While scaling up classical quantitative genetics approaches to the study of thousands of omics traits opens new avenues for the dissection of molecular mechanisms that regulate biological systems, it is also accompanied by a whole new range of specific challenges. These challenges are intrinsic to the high-throughput nature of the measurements, to technical aspects of the profiling technologies used, to the statistical issues introduced by the untargeted multifactorial perturbation that underlies the approach, and to the complexity of the molecular networks under study. In this review, we inspect important issues that arise at each step (Figure 8.1) of a genetical genomics study from experimental design to result interpretation. We provide accordingly recommendations allowing a reliable and efficient use of the power of genetical genomics. This review will focus primarily on gene expression profiling, but many of the issues raised are also applicable to other "omics" technologies.

#### **BOX 8.1: Special features of molecular traits**

Many types of molecular traits studied in genetical genomics experiments have specific properties that can be helpful in the analysis and interpretation of the data. For example, the location of the genes coding for the transcripts or the proteins studied is usually known. This extra information compared to classical non-molecular phenotypes can be used to gain a deeper insight into the mechanistic details of the underlying biological processes.

##### **Isoforms and modifications**

Molecular traits can often be observed in different forms. For example, transcripts are spliced into different variants. It was shown that this alternative splicing could itself be mapped to genetic variation (Kwan *et al.* 2008, Li *et al.* 2010a). On a 2D-gel, the same protein often migrates to more than one spot. This can be explained by post-translational modifications such as phosphorylation, and mapping the genetic basis for variations in the phosphoproteome and kinome is an exciting prospect. Similarly with mass spectrometry techniques, different forms of the same protein will yield different sets of peaks. Moreover, in order to remove the multiple testing induced by the observation of the same protein through multiple mass peaks, statistical methods that automatically connect those peaks can advantageously be used (Dijkstra and Jansen 2009).

**BOX 8.1(continued): Special features of molecular traits*****Local and distant QTLs***

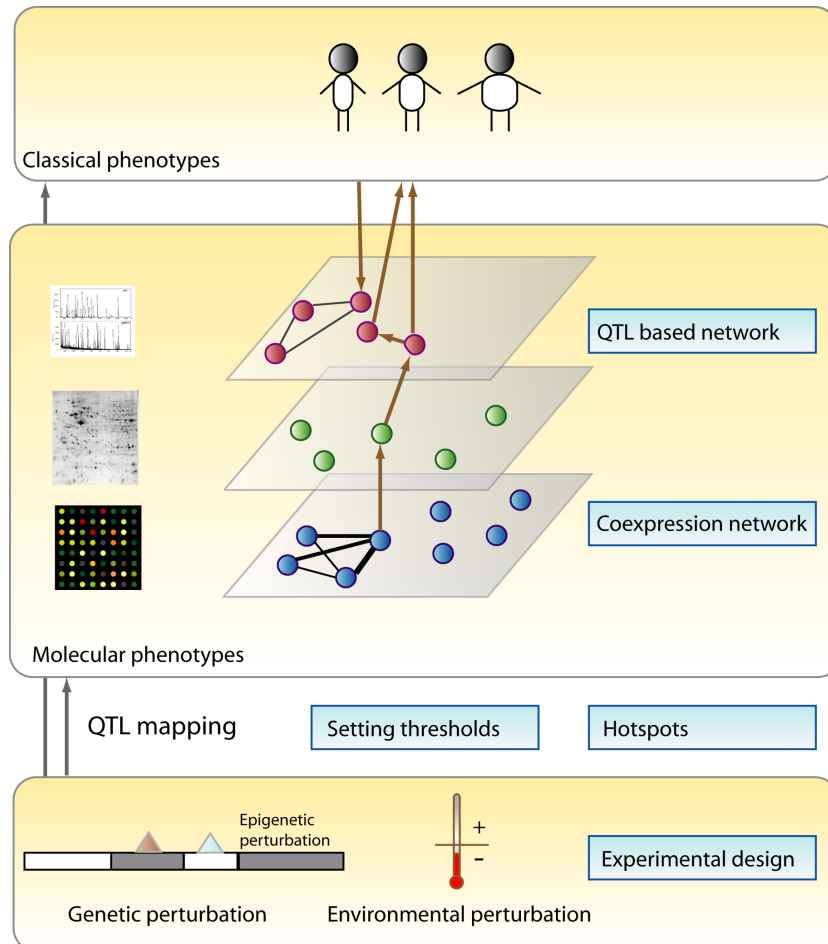
Protein and mRNA eQTLs can be classified into *Local* and *distant* QTLs depending on the relative location of the QTL and the gene coding for the measured mRNA or protein. The identification of *distant* QTLs tends to be less reliable than that of *local* QTLs: this can be attributed to two main reasons. First, the effects of *distant* QTLs are biologically more indirect and therefore harder to pick up. Second, when one tests for a *distant* QTL, the loci tested are genome wide, as opposed to just the gene's locus in the case of a local effect, therefore the power to detect *distant* QTLs is much lower than the power to detect *local* QTLs. This is consistent with the fact that distant QTLs tend to be more difficult to replicate than *local* QTLs (Peirce et al. 2006).

***False local QTLs***

Frequently, strong local linkage affecting transcripts reported intensity is actually the result of a technical artifact rather than a biological differential expression. Microarray probes are most often designed following reference assemblies that are based on "mainstream" laboratory strains (in the case of model organisms) or a few individuals (in the case of humans). As a result of this, the hybridization efficiency of these probes can vary from one individual to another when the probe's target sequence is polymorphic (Alberts et al. 2007). In this case, the differential intensity between the two alleles will reflect a difference in hybridization efficiency rather than a difference in true gene expression signal. It is therefore necessary to mask all probes containing known sequence polymorphisms between the parental strains studied. Alternatively, for short-oligonucleotide arrays, which typically contain multiple probes per gene, Alberts et al. suggest to include probe effects in the statistical model, which can potentially separate "true" differential mRNA expression from "ghost" effects caused by polymorphisms (Alberts et al. 2005).

## 8.2 Designing a genetic experiment for thousands of phenotypes

Many of the considerations that apply to the experimental design of a classical genetic study also apply to genetical genomics. However, because the number of traits studied in a genetical genomics experiments is of a much higher magnitude (tens of thousands typically), a few specific issues need to be taken into account when deciding the population type, the sample size, and the assignment of samples to different treatments or conditions. In this section, we address the consequences of these decisions, and we introduce approaches for optimizing the power and resolution in genetical genomics experiments.



**Figure 8.1:** Genetical genomics as a systems genetics strategy. Genetical genomics integrates data from multiple molecular levels (middle box) and classical phenotypes (upper box) by relating their variation to common multifactorial perturbations (a combination of (epi)genetic and environmental factors as shown in the lower box). Methods to form undirected association networks within a molecular level (e.g. coexpression network) or to draw directed edges between phenotypes use QTL information and allow relating classical phenotypes such as obesity to relevant genes, metabolites or proteins.

### 8.2.1 Population

Just as in any genetic study, the first critical step in designing a genetical genomics experiment is the choice of a population to be studied, which will determine the

ensuing mapping strategy: linkage or association mapping. In genetical genomics studies, multiple testing caused by the mapping of large numbers of phenotypes reduces the available statistical power. Linkage mapping on an inbred population such as recombinant inbred lines (RILs), F2 intercrosses or backcrosses provides enough power to perform eQTL studies with relatively small sample sizes. Fully inbred populations (immortal lines) allow collecting different types of phenotypes on distinct but genetically identical individuals, which is a valuable advantage in systems biology experiment where invasive procedures are needed to collect various phenotypes. However, linkage genetical genomics studies in general provide a relatively poor resolution, i.e. the confidence intervals surrounding a QTL span large genome regions of typically several million base pairs. This results in additional efforts needed to identify the actual polymorphism causing the QTL, for instance using independent information such as gene annotation (Franke *et al.* 2006), or generating congenic strains. RILs can be generated from more than two parental strains; for example, the mouse collaborative cross (Churchill *et al.* 2004) and the Multiparent Advanced Generation Inter-Cross (MAGIC) for *Arabidopsis thaliana* (Kover *et al.* 2009) use 8 and 19 founder lines, respectively. Other types of crosses, such as Advanced Intercross Lines (AIL) (Darvasi and Soller 1995, Rockman and Kruglyak 2008) or Heterogeneous Stock (HS) for rat (Hansen and Spuhler 1984), introduce more recombinations and therefore have improved resolution. Association studies performed on natural or outbred populations on the other hand, have less power because a much larger number of smaller genomic regions are tested for QTL, leading to a drop in statistical significance caused by increased multiple testing and because of the large imbalance of the allele frequencies of genotypes. Since association studies allow for a much finer mapping of the QTL than that obtained with linkage analysis, there is a trade-off to consider between power and resolution when choosing the mapping strategy. Genome-wide association studies (GWAS) have naturally been used to perform genetical genomics studies in humans (Dubois *et al.* 2010, Emilsson *et al.* 2008, Goring *et al.* 2007, Heap *et al.* 2009, Stranger *et al.* 2005) and are emerging in model organisms studies using outbred populations (Ghazalpour *et al.* 2008).

### 8.2.2 Combining studies

Combining information from different studies can further increase the power and resolution in eQTL mapping. Meta-analysis of multiple datasets is a strategy widely used in GWAS of classical traits but is only starting to be explored in the context of genetical genomics (Dubois *et al.* 2010, Heap *et al.* 2009). Meta-analyses use sta-

tistical methods for combining p-values (Whitlock 2005), because combining directly data from different experiments is hampered by heterogeneity issues (e.g. different microarray platforms). As a result the power increases: combining their own peripheral blood dataset with the HapMap B-cell dataset, Heap *et al.* report close to 40% additional eQTLs that were not detected in the individual eQTL scans. Also, the combination of association and linkage mapping, a procedure commonly used in classical genetics studies, has recently been applied to eQTL studies (Gatti *et al.* 2009). A linkage study is first performed to identify eQTL regions with satisfactory power; an association study is then performed to refine the eQTL found by the linkage study. This association step can be performed using a relaxed statistical significance threshold since only the regions identified in the linkage step are tested.

### 8.2.3 Sample assignment for molecular profiling

After the choice of a population to be studied, molecular profiling can be conducted using high-throughput technologies. In this step, random assignment of experimental units is a fundamental principle of experimental design which ensures that a treatment of interest is not confounded with other factors (Fisher 1951, Wit and McClure 2004). While the genotypes are naturally randomized in the process of meiotic recombination and segregation, randomization must be enforced for other relevant factors during the design of genetical genomics experiments. In order to optimize the design for statistical power, the best way is to increase the sample size; but a smart assignment of samples to experimental units can further maximize the information that can be extracted from the data without any additional costs. For example, it was suggested to pair the most genetically distant individuals on two-color microarrays so as to maximize the number of informative genetic contrasts (Fu and Jansen 2006, Lam *et al.* 2008). Two-color arrays are no longer widely used, but the basic idea can be elegantly generalized: in genetical genomics experiments studying environmental perturbation, one aims at achieving the most accurate estimate of the QTL effects and QTL-by-environment interaction effects of interest, either in one or more regions of special interest, such as a previously detected phenotypic QTL, or across the entire genome. In this case, genotyped individuals can be 'intelligently' selected and distributed across multiple environments using an optimization algorithm to minimize the sum of variance of the parameter estimates of interest (Lam *et al.* 2008, Li *et al.* 2008, Li *et al.* 2009).

### 8.3 Significance thresholds for eQTL detection

The large number of molecular traits (tens of thousands) and markers (from 100s to millions) that are tested in a genetical genomics study requires the significance level for linkage or association to be rigorously adjusted to control the number of false positive results. Bonferroni correction in this context tends to be too conservative, and in genetical genomics studies, it is more appropriate to control false discovery rate (FDR) (Benjamini and Hochberg 1995). In practice, the approaches for calculating the significance thresholds accounting for multiple testing used in genetical genomics are mostly relying on permutations (Breitling *et al.* 2008a, Churchill and Doerge 1994), since standard approaches (Storey and Tibshirani 2003) work under the assumption that there is relatively mild dependence of the tests, which is not the case in genetical genomics where important correlations exist between traits and between neighboring markers. Permuting aims at breaking the biological relationship between genotypes and traits so that any QTL detected in the permuted dataset is a false positive, which allows estimating the FDR by providing an estimate of the number of false positives to be expected in the original data. By permuting only the sample labels in the genotype data, both the correlation structure between traits and the correlation structure between markers is conserved, which makes this empirical procedure perfectly suited to a non-biased estimation of the significance under the multiple dependences present in the data. If a major correlation structure is causing large groups of genes to be associated with the genotypes at random genomic loci, forming spurious hotspots of eQTLs, such permutations would also be likely to lead to hotspots being mapped by chance and therefore identify the hotspots as not significant (Breitling *et al.* 2008a). Thousands of permutations are usually required to ensure accuracy of the FDR estimates, but methods approximating the tail of the distribution may allow for extrapolation from a smaller number of permutations and reduce the computational burden (Knijnenburg *et al.* 2009). When the statistical models used for mapping contain genetic, environmental and interacting factors, the appropriate permutation strategy may be difficult to determine as certain situations require different permutation procedures to be used for individual terms in the ANOVA model, including restricted permutation, permutation of whole groups of units, permutation of some forms of residuals or some combination of these (Anderson and ter Braak 2003).

Special situations require some additional adjustments to the significance threshold used. Firstly, testing for a *local* eQTL effect (a QTL affecting a gene lying in a nearby locus on the same chromosome) involves testing the genotypes at only one restricted genome region as opposed to the whole genome when scanning for dis-

tant genetic effects. Therefore detection of *local* eQTLs is affected to a much lesser extent to multiple testing and it is advisable to use a relaxed threshold for the detection of *local* QTLs. Secondly, in the presence of imbalanced allele frequencies (occurring randomly or caused by segregation distortion) in an experimental population, one of the genotype group may have a very limited size yielding unreliable estimate of mean within that group, which in turn may influence the accuracy of the p-value estimates. The same issue is usually avoided in association studies where SNPs with very low minor allele frequency (e.g. below 5%) are simply excluded, at the risk of missing important biological phenomena (Dickson *et al.* 2010).

## 8.4 Defining gene and QTL networks

In addition to the genetic dissection of phenotypic variation using QTL mapping techniques, systems geneticists are interested in reconstructing the biological networks that connect genes, proteins and other traits based on their observed genetic (co-)variation. In this context, biological networks are often defined by graphical models that are composed of nodes representing traits such as gene expression levels and edges representing (causal, correlational or mechanistic) relationships between these nodes. In current genetical genomics studies, there are two main types of approaches for the inference of such networks (i) methods for identifying coexpression networks on the basis of (partial) correlations between traits; (ii) methods for identifying QTL networks on the basis of QTL underlying variation and coexpression.

### 8.4.1 Correlation-based networks

Coexpression networks are undirected networks in which edges connect genes that have correlated expression behaviors over a set of samples. In the genetical genomics context, these samples come from genetically diverse individuals, possibly observed over multiple conditions. The coexpression similarity between genes can be measured using different metrics, the most commonly used being Pearson's correlation. Coexpression of two genes does not prove a causal relationship; however under the principle of "guilt by association", it can be used to predict similar gene functions and is indicative of possible co-regulation.

Coexpression methods can be divided between unweighted and weighted approaches. Unweighted approaches use cutoffs to define a minimum level of correlation (or of another coexpression metric) required to draw an edge between two



genes. Those cutoffs are typically determined using permutations (Butte *et al.* 2000, Carter *et al.* 2004). Weighted approaches on the other hand consider all genes to be interconnected albeit with different strengths (weights) (Zhang and Horvath 2005). Weighted approaches typically transform the correlation (e.g. with a power function) to emphasize the weight of higher correlations, a practice known as “soft thresholding”. A natural step following the generation of the global network in a co-expression analysis is to identify modules of coexpressed genes, i.e. find communities of highly interconnected nodes within the graph. Such coexpressed modules can then be studied as putative functional units, thereby considerably reducing the dimensionality of the data. Different approaches have been proposed, many of which are inspired by social network research. Chesler *et al.* choose to focus on sets of genes in which all nodes are interconnected; such sets are termed “cliques” (Chesler *et al.* 2005). Searching for cliques in a network containing thousands of nodes poses a serious computational burden and several algorithms have been designed to alleviate it (Baldwin *et al.* 2005). An alternative is the use of the topological overlap measure (TOM): this metric allows grouping together genes that share the same neighbors in the correlation graph (Ravasz *et al.* 2002, Zhang and Horvath 2005), but without the strong constraint imposed by cliqueness.

One strategy to identify important genes within coexpressed modules has been to focus on highly connected genes. Connectivity (also known as degree) represents the amount of edges reaching a gene in the coexpression network in the unweighted case, or the sum of the correlation strength with all other genes (correlations) in the weighted case. Genes with high connectivity, termed “hubs”, have been claimed to be enriched for essential genes (Carter *et al.* 2004). Connectivity is therefore used to prioritize between genes belonging to modules of interest.

Similar correlation-based approaches can be used to study metabolites (Kose *et al.* 2001). Steuer discussed the important differences existing in the correlation structure of metabolites compared to that of genes because of the specific biochemical characteristics of metabolic networks, in which molecules rather than information is flowing along pathways (Steuer 2006). A promising perspective is the profiling of multiple classes of macromolecules in the same samples in order to form correlation networks integrating genes, metabolites, and possibly proteins (Fu *et al.* 2009).

By using partial correlations, i.e. conditioning on selected other nodes in the network, it is possible to remove indirect edges from the network (Bing and Hoeschele 2005, de la Fuente *et al.* 2004, Keurentjes *et al.* 2006). Since large scale changes in coexpression may indicate rewiring of the transcriptional network, recent work has

focused on the identification of such changes between different conditions in what is known as differential coexpression analysis (Choi and Kendziorski 2009, Tesson *et al.* 2010). One limitation of correlation-based networks is that they are undirected and do not use explicitly the genotypic variation, therefore lacking the causal information that is needed to identify the drivers of biological processes.

#### 8.4.2 QTL-based networks

The interest of using multiple QTL co-localization information for the reconstruction of trait networks has been noted early on (Jansen and Nap 2001). The basic idea is that QTLs from upstream regulators should also be QTLs of the associated downstream traits, providing a simple means to order traits from causal to reactive. Moreover, when two genes map to the same eQTL, one locally and one distantly, the gene with the *local* eQTL is likely to regulate the gene with a *distante*QTL (Jansen 2003). In practice, the application of these ideas has been hampered by two limitations of most available datasets. Firstly, the lack of power of current genetical genomics experiments does not allow for deconvolution of traits into multiple QTL (one or two QTL per trait are detected at best, and discrimination between a weak but existing QTL and absence of any QTL effect is difficult). Secondly, in experiments with low mapping resolution, it is often impossible to discriminate between two distinct neighbouring QTLs, and one shared QTL (statistical methods provide 'parsimonious' models, but this does not exclude that reality is more complex).

Building on the aforementioned fundamental principles, Bayesian modeling concepts for causal inference have been adapted to assist in the extraction of regulatory evidence from genetical genomics data. If a trait T1 regulates a trait T2, then variation in T1 will be propagated to T2. When some of T1's variation can be accounted for by a QTL, this QTL will also explain some of the variation in T2. The regression of T2 on T1 corrects T2 for the variation propagated from T1, including the QTL variation: this independence of T2 and the QTL conditional on T1 is used as evidence for the fact that T1 is causal for (regulates) T2. Different statistical testing frameworks have been proposed to use this conditional independence property. For example, model selection approaches have been used to identify the causal relationship among traits that is best supported by the data (Li *et al.* 2006a, Schadt *et al.* 2005). Chen *et al.* provided a method to quantify the likelihood of each causal link (Chen *et al.* 2007). Recently, Millstein *et al.* further formularize a similar idea into a hypothesis test which results in a quantitative estimation of significance in terms of p-value (Millstein *et al.* 2009). Chaibub Neto *et al.* propose a likelihood-based method to compare graph configurations in which the non-propagated vari-

ation present in the downstream trait is explicitly modeled by non-shared QTL(s) (Chaibub Neto *et al.* 2008). The performances of those methods in terms of power, false positive and false negative rates are strongly dependent on sample size, QTL effect sizes, genotype frequencies and measurement errors (Li *et al.* 2010b).

Some attempts have been made to combine co-expression networks with QTL-based causal inference: either by orienting undirected edges of coexpression networks (Aten *et al.* 2008, Chaibub Neto *et al.* 2008) or by inferring causal relationships between entire modules and clinical traits by studying the eigengenes representing those modules or selected genes from those modules (Chen *et al.* 2008, Plaisier *et al.* 2009).

### 8.4.3 Hotspots

A particular case of QTL-based networks is that of QTL hotspots: specific loci that control a large number of genes distantly. Hotspots may be the consequence of one single polymorphism with major direct effects: for example, a polymorphic transcription factor affecting multiple targets. Hotspots could also be the result of the indirect downstream effects of a single polymorphism. A handful of such eQTL hotspots have been biologically validated. For example, using a small interfering RNA (siRNA) knockdown of selected candidate genes, Wu *et al.* were able to observe a phenotypic change consistent with the predicted function of the hotspot genes in mouse (Wu *et al.* 2008). Also, a variant in the *ERECTA* gene was found to cause variation in a number of molecular traits (transcripts, proteins and metabolites) as well as classical phenotypes (Fu *et al.* 2009). Another possible hotspot cause is the co-localization of multiple polymorphisms with unrelated effects: the hotspot QTL would then typically lie within a SNP dense region or a gene rich region and this can be tested.

If the hotspot is the result of a single polymorphism, one might expect that genes whose expression is affected by this polymorphism should belong to a common biological pathway or process, at least if the effect is reasonably direct. For that reason, one of the first tests performed on the genes affected by a hotspot is often a gene annotation enrichment analysis such as GSEA (Backes *et al.* 2007, Dennis *et al.* 2003) or iGA (Breitling *et al.* 2004). It is, however, important to remember that this annotation enrichment alone does not prove the validity of the hotspots: as detailed in Box 8.2 with the tissue purity scenario, spurious hotspots would be expected to contain biologically related genes as well.

The search for a “master regulator” within the hotspot QTL interval is challenging since typically many candidate genes lie in the QTL confidence interval due

to the lack of resolution in most genetical genomics linkage studies (see also the earlier section). Interestingly, loci harboring eQTL hotspots were not found to be enriched for transcription factors in a yeast study (Yvert *et al.* 2003), and the majority of hotspots turns out to be due to very indirect effects on gene expression. In order to prioritize genes within the list of candidate regulators, multiple independent sources of information can be utilized (Franke *et al.* 2006). Statistical evidence such as correlation of the hotspot genes with the candidate regulator or the presence of a *local* eQTL for the regulator can be integrated with biological evidence such as the relevance of the functional annotations associated with the candidate gene. Sequence information can also be used. Is the candidate gene polymorphic between the two parental strains? Is there evidence of enrichment of certain transcription factor binding sites within the hotspot target genes that would provide clues on the involvement of a certain regulator? Finally, it is important to remember that the regulators underlying the QTL may not be protein-coding genes but could also be miRNAs, or structural or epigenetic mechanisms. For integrating these different pieces of information, the rank product method can be applied to prioritize the candidate regulators by multiplication across the ranks positions of candidate genes in each prioritization step (Breitling *et al.* 2004, Keurentjes *et al.* 2007).

Regulatory links derived from the above-mentioned approaches should only be considered as putative and should be experimentally validated. Experimental validation techniques include gene knockouts, transgenic animals, RNA interference-based knockdowns and chemical perturbations.

#### 8.4.4 Non-genetic variation

The aim of genetical genomics studies is to dissect genetic variation and trace its propagation from DNA to molecular and classical traits shedding some light onto biological pathways and processes. However, previous studies have shown that many non-genetic factors contribute to variation in the data (Churchill 2002). For example, hybridization batches, population stratification, the preparation of samples by different technicians, variations in sample composition and purity can introduce strong expression changes in large groups of genes, creating a pervasive correlation structure. When those factors are known, their effects can be corrected (by including them in a regression), as is routinely done with controlled factors such as sex or age. Unfortunately, in many cases, the relevant factors are “hidden” (e.g. varying physiological state or tissue purity in the samples), and failure to account for them will have a crucial influence at many levels of a genetical genomics study.

**BOX 8.2: Confounding factors for eQTL hotspots.****Known confounding factors**

Many uncontrolled factors, such as hybridization batches, can result in highly correlated traits and clusters of QTLs. These non-biological sources of variation can be introduced during the preparation of samples, the hybridization and the measurements. Known batch effects can be filtered out from the data using a batch correction method, such as a linear regression. In some cases, however, the confounding factors are unknown and this strategy cannot be applied. When applying a genetical genomics strategy to multiple tissues/conditions, the scale of the experiment is usually doubled or tripled. It is almost impossible to conduct the whole experiment at once. The resulting batch differences become influential and it should be noted that this effect can be different in multiple tissues/conditions. In such cases, a batch-by-condition interaction effect should be considered in order to properly remove the unwanted variation caused by batches (Mead 1988).

**Unknown confounding factors**

Even if all known batch effects induced by sample processing have been explicitly corrected for, some variation arising from other untraceable factors will always remain. One example of a parameter difficult to trace is cellular state when studying the dynamics of gene expression during differentiation. When samples from a given cell differentiation stage are collected using certain markers (e.g. cell surface markers), the variation affecting those markers may have important consequences on QTL detection. Dissected tissue samples can contain slightly varying fractions of individual cell types, leading to cell-type specific gene clusters that vary in a correlated manner. If the fraction of a certain cell-type in the tissues happens to correlate with the genotype at a particular marker, all those cell-type specific genes will erroneously map to this marker. Therefore, it is suggested to perform genetical genomics experiments on samples from highly purified tissues or even at a single cell level (Gerrits *et al.* 2009); however, this would come with an inherent cost in terms of additional experimental and biological noise (Ozbudak *et al.* 2002).

Firstly, the presence of large non-genetic variation components in traits reduces the proportion of variance explained by genetic factors, thereby limiting the power for QTL detection.

Secondly, the large scale correlation induced by non-genetic variation can be dominant and cause coexpression methods to fail to capture the correlation stemming from genetic variation. Indeed, the impact of genetic variation on the correlation structure has been observed to be weak. For example, Ayroles *et al.* noted that the heritability of gene expression traits and connectivity are negatively correlated among a panel of *D. melanogaster* inbred lines (Ayroles *et al.* 2009), and suggested that hub genes are protected against genetic variation by purifying selection so that genetic variation does not have widespread effects within the coexpression network. Instead, it is likely that the global correlation structure observed in most experiments stems from variation in non-genetic factors such as sex, tissue compo-

sition, physiological state or experimental conditions. Eliminating such sources of correlation is at least as crucial to QTL based causal inference as it is to coexpression analyses, since correlated error terms are known to confuse causal inference methods (Li *et al.* 2010b).

Thirdly, several studies have suggested that most hotspots are likely to be spurious (Breitling *et al.* 2008a, de Koning and Haley 2005) and may be a mere statistical consequence of the correlation structure induced by hidden factors inherent to the data. If, by chance, one of those hidden factors correlates (even moderately) with the genotype distribution at one of the markers, a QTL hotspot containing most of these genes will be spuriously detected at this marker. In Box 8.2, several scenarios leading to such wrong interpretations of the data are detailed.

Since this confounding variation is usually not completely resolved by randomization of the experimental design (Churchill 2002) or through low-level normalization techniques (Yang *et al.* 2002), unsupervised methods to identify and control the variation associated with hidden factors have recently been proposed. For example, surrogate variable analysis (SVA) identifies the variables representing consistent expression signatures due to sources other than the factors of interest and these can be included in subsequent analyses as covariates to reduce dependency across genes, stabilize false discovery rate estimates, and improve the reproducibility of the analysis (Leek and Storey 2007). Kang *et al.* proposed inter-sample correlation emended (ICE) eQTL mapping which estimates the total correlation between samples and incorporates it into a linear mixed model as a variance component accounting for a random effect (Kang *et al.* 2008). These methods result in the removal of most hotspots and in a more sensitive detection of *local* QTL signals. From the simulations results by Kang *et al.*, ICE is able to correct for a mixture of strong and moderate confounding effects while SVA is able to correct only for a number of strong confounding factors. However, as Kang *et al.* acknowledge, true regulatory hotspots may affect a large number of genes and cause inter-sample correlation. Therefore, when correcting for this inter-sample correlation, one risks removing true hotspots as well. Finally, Dubois *et al.* identified large scale effects via Principal Component Analysis (PCA) on a comprehensive collection of unrelated datasets spanning multiple tissues and conditions, and used these principal components (termed transcriptional components) to correct for variation in their own independent data, resulting in improved power to detect QTLs (Dubois *et al.* 2010).

Importantly, even some of the correlation induced by genetic variation may falsely suggest a pathway relationship. For example, in an inbred genetic panel such as an F2 or RILs, large fractions of the genome are under strong linkage disequilibrium as

few recombination events are present within the studied population. Consequently, all genes under local genetic control within those areas are linked to a common genotype and therefore are (somewhat) correlated but this correlation between neighboring genes does not reflect membership to a common functional group. It is possible to remove such linkage disequilibrium-induced correlation by working with expression data corrected for local genotypes (Heap *et al.* 2009).

## 8.5 Conclusions

The adaptation of old concepts from classical genetics and epidemiology to the new postgenomic fields is establishing itself as a major research area with the potential to elucidate the biological processes leading to complex phenotypes. As standard good practices are adopted by the community for the design, statistical analysis and biological interpretation of genetical genomics experiments, the trend of these genetic studies will be to go deeper (integrating more molecular levels (Ferrara *et al.* 2008, Fu *et al.* 2009, Johannes *et al.* 2008) and broader (larger sample sizes, combining genetic perturbation with other factors such as environmental factors) (Jansen *et al.* 2009, Li *et al.* 2008). The pervasive correlation structure stemming from (mainly poorly understood) physiological and technical factors within genomics datasets is appearing as the main challenge slowing down the path towards new discovery. Promising new approaches that tackle this confounding variation (Dubois *et al.* 2010, Kang *et al.* 2008, Leek and Storey 2007) are emerging and already proving to be beneficial as they improve the power to detect QTL while eliminating spurious findings. The application of these new approaches to network reconstruction (Chaibub Neto *et al.* 2008, Chen *et al.* 2007, Schadt *et al.* 2005, Zhang and Horvath 2005) promises to be accompanied by new breakthroughs by removing one of the major obstacles on the way towards reliable network inference (Li *et al.* 2010b).





---

## Bibliography

- Akey J. M., Biswas S., Leek J. T. and Storey J. D.: 2007, On the design and analysis of gene expression studies in human populations, *Nat Genet* **39**(7), 807–8; author reply 808–9.
- Alberts R., Terpstra P., Bystriykh L. V., de Haan G. and Jansen R. C.: 2005, A statistical multiprobe model for analyzing cis and trans genes in genetical genomics experiments with short-oligonucleotide arrays, *Genetics* **171**(3), 1437–9.
- Alberts R., Terpstra P., Li Y., Breitling R., Nap J. P. and Jansen R. C.: 2007, Sequence polymorphisms cause many false cis eqtls, *PLoS ONE* **2**(7), e622.
- Anderson M. and ter Braak C.: 2003, Permutation tests for multi-factorial analysis of variance, *Journal of Statistical Computation and Simulation* **73**, 85–113.
- Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T., Harris M. A., Hill D. P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J. C., Richardson J. E., Ringwald M., Rubin G. M. and Sherlock G.: 2000, Gene ontology: tool for the unification of biology. the gene ontology consortium, *Nat Genet* **25**(1), 25–9.
- Aten J. E., Fuller T. F., Lusi A. J. and Horvath S.: 2008, Using genetic markers to orient the edges in quantitative trait networks: the neo software, *BMC Syst Biol* **2**, 34.
- Ayroles J. F., Carbone M. A., Stone E. A., Jordan K. W., Lyman R. F., Magwire M. M., Rollmann S. M., Duncan L. H., Lawrence F., Anholt R. R. and Mackay T. F.: 2009, Systems genetics of complex traits in drosophila melanogaster, *Nat Genet* **41**(3), 299–307.

- Backes C., Keller A., Kuentzer J., Kneissl B., Comtesse N., Elnakady Y. A., Muller R., Meese E. and Lenhof H. P.: 2007, Genetrail—advanced gene set enrichment analysis, *Nucleic Acids Res* **35**(Web Server issue), W186–92.
- Baldwin N. E., Chesler E. J., Kirov S., Langston M. A., Snoddy J. R., Williams R. W. and Zhang B.: 2005, Computational, integrative, and comparative methods for the elucidation of genetic coexpression networks, *J Biomed Biotechnol* **2005**(2), 172–80.
- Barberan-Soler S. and Zahler A. M.: 2008a, Alternative splicing and the steady-state ratios of mrna isoforms generated by it are under strong stabilizing selection in *caenorhabditis elegans*, *Mol Biol Evol* **25**(11), 2431–7.
- Barberan-Soler S. and Zahler A. M.: 2008b, Alternative splicing regulation during *c. elegans* development: splicing factors as regulated targets, *PLoS Genet* **4**(2), e1000001.
- Benjamini Y. and Hochberg Y.: 1995, Controlling the false discovery rate: a practical and powerful approach to multiple testing., *J Roy Statist Soc* **57**, 289–300.
- Bergman A. and Siegal M. L.: 2003, Evolutionary capacitance as a general feature of complex gene networks, *Nature* **424**(6948), 549–52.
- Bing N. and Hoeschele I.: 2005, Genetical genomics analysis of a yeast segregant population for transcription network inference, *Genetics* **170**(2), 533–42.
- Blow N.: 2007, Genomics: the personal side of genomics, *Nature* **449**(7162), 627–30.
- Bodmer W. and Bonilla C.: 2008, Common and rare variants in multifactorial susceptibility to common diseases, *Nat Genet* **40**(6), 695–701.
- Boer M. P., Wright D., Feng L., Podlich D. W., Luo L., Cooper M. and van Eeuwijk F. A.: 2007, A mixed-model quantitative trait loci (qtl) analysis for multiple-environment trial data using environmental covariables for qtl-by-environment interactions, with an example in maize, *Genetics* **177**(3), 1801–13.
- Breitling R., Amtmann A. and Herzyk P.: 2004, Iterative group analysis (iga): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments, *BMC Bioinformatics* **5**, 34.
- Breitling R., Li Y., Tesson B. M., Fu J., Wu C., Wiltshire T., Gerrits A., Bystrykh L. V., de Haan G., Su A. I. and Jansen R. C.: 2008a, Genetical genomics: spotlight on qtl hotspots, *PLoS Genet* **4**(10), e1000232.

- Breitling R., Vitkup D. and Barrett M. P.: 2008b, New surveyor tools for charting microbial metabolic maps, *Nat Rev Microbiol* **6**(2), 156–61.
- Brem R. B. and Kruglyak L.: 2005, The landscape of genetic complexity across 5,700 gene expression traits in yeast, *Proc Natl Acad Sci U S A* **102**(5), 1572–7.
- Brem R. B., Storey J. D., Whittle J. and Kruglyak L.: 2005, Genetic interactions between polymorphisms that affect gene expression in yeast, *Nature* **436**(7051), 701–3.
- Brem R. B., Yvert G., Clinton R. and Kruglyak L.: 2002, Genetic dissection of transcriptional regulation in budding yeast, *Science* **296**(5568), 752–5.
- Bryder D., Rossi D. J. and Weissman I. L.: 2006, Hematopoietic stem cells: the paradigmatic tissue-specific stem cell, *Am J Pathol* **169**(2), 338–46.
- Buck M. J. and Lieb J. D.: 2004, Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments, *Genomics* **83**(3), 349–60.
- Butte A. J., Tamayo P., Slonim D., Golub T. R. and Kohane I. S.: 2000, Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks, *Proc Natl Acad Sci U S A* **97**(22), 12182–6.
- Bystrykh L., Weersing E., Dontje B., Sutton S., Pletcher M. T., Wiltshire T., Su A. I., Vellenga E., Wang J., Manly K. F., Lu L., Chesler E. J., Alberts R., Jansen R. C., Williams R. W., Cooke M. P. and de Haan G.: 2005, Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics', *Nat Genet* **37**(3), 225–32.
- Carlborg O. and Haley C. S.: 2004, Epistasis: too often neglected in complex trait studies?, *Nat Rev Genet* **5**(8), 618–25.
- Carter S. L., Brechbuhler C. M., Griffin M. and Bond A. T.: 2004, Gene co-expression network topology provides a framework for molecular characterization of cellular state, *Bioinformatics* **20**(14), 2242–50.
- Chaibub Neto E., Ferrara C. T., Attie A. D. and Yandell B. S.: 2008, Inferring causal phenotype networks from segregating populations, *Genetics* **179**(2), 1089–100.
- Chambers S. M., Boles N. C., Lin K. Y., Tierney M. P., Bowman T. V., Bradfute S. B., Chen A. J., Merchant A. A., Sirin O., Weksberg D. C., Merchant M. G., Fisk C. J., Shaw C. A. and Goodell M. A.: 2007, Hematopoietic fingerprints: an expression database of stem cells and their progeny, *Cell Stem Cell* **1**(5), 578–91.

- Chen L. S., Emmert-Streib F. and Storey J. D.: 2007, Harnessing naturally randomized transcription to infer regulatory relationships among genes, *Genome Biol* **8**(10), R219.
- Chen Y., Zhu J., Lum P. Y., Yang X., Pinto S., MacNeil D. J., Zhang C., Lamb J., Edwards S., Sieberts S. K., Leonardson A., Castellini L. W., Wang S., Champy M. F., Zhang B., Emilsson V., Doss S., Ghazalpour A., Horvath S., Drake T. A., Lusis A. J. and Schadt E. E.: 2008, Variations in dna elucidate molecular networks that cause disease, *Nature* **452**(7186), 429–35.
- Chesler E. J., Lu L., Shou S., Qu Y., Gu J., Wang J., Hsu H. C., Mountz J. D., Baldwin N. E., Langston M. A., Threadgill D. W., Manly K. F. and Williams R. W.: 2005, Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function, *Nat Genet* **37**(3), 233–42.
- Chesler E. J., Lu L., Wang J., Williams R. W. and Manly K. F.: 2004, Webqtl: rapid exploratory analysis of gene expression and genetic networks for brain and behavior, *Nat Neurosci* **7**(5), 485–6.
- Cheung V. G., Spielman R. S., Ewens K. G., Weber T. M., Morley M. and Burdick J. T.: 2005, Mapping determinants of human gene expression by regional and genome-wide association, *Nature* **437**(7063), 1365–9.
- Choi Y. and Kendzierski C.: 2009, Statistical methods for gene set co-expression analysis, *Bioinformatics* **25**(21), 2780–6.
- Churchill G. A.: 2002, Fundamentals of experimental design for cDNA microarrays, *Nat Genet* **32 Suppl**, 490–5.
- Churchill G. A., Airey D. C., Allayee H., Angel J. M., Attie A. D., Beatty J., Beavis W. D., Belknap J. K., Bennett B., Berrettini W., Bleich A., Bogue M., Broman K. W., Buck K. J., Buckler E., Burmeister M., Chesler E. J., Cheverud J. M., Clapcote S., Cook M. N., Cox R. D., Crabbe J. C., Crusio W. E., Darvasi A., Deschepper C. F., Doerge R. W., Farber C. R., Forejt J., Gaile D., Garlow S. J., Geiger H., Gershenfeld H., Gordon T., Gu J., Gu W., de Haan G., Hayes N. L., Heller C., Himmelbauer H., Hitzemann R., Hunter K., Hsu H. C., Iraqi F. A., Ivandic B., Jacob H. J., Jansen R. C., Jepsen K. J., Johnson D. K., Johnson T. E., Kempermann G., Kendzierski C., Kotb M., Kooy R. F., Llamas B., Lammert F., Lassalle J. M., Lowenstein P. R., Lu L., Lusis A., Manly K. F., Marcucio R., Matthews D., Medrano J. F., Miller D. R., Mittelman G., Mock B. A., Mogil J. S., Montagutelli X., Morahan G., Morris D. G., Mott R., Nadeau J. H., Nagase H., Nowakowski

- R. S., O'Hara B. F., Osadchuk A. V., Page G. P., Paigen B., Paigen K., Palmer A. A., Pan H. J., Peltonen-Palotie L., Peirce J., Pomp D., Pravenec M., Prows D. R., Qi Z., Reeves R. H., Roder J., Rosen G. D., Schadt E. E., Schalkwyk L. C., Seltzer Z., Shimomura K., Shou S., Sillanpaa M. J., Siracusa L. D., Snoeck H. W., Spearow J. L., Svenson K. *et al.*: 2004, The collaborative cross, a community resource for the genetic analysis of complex traits, *Nat Genet* **36**(11), 1133–7.
- Churchill G. A. and Doerge R. W.: 1994, Empirical threshold values for quantitative trait mapping, *Genetics* **138**(3), 963–71.
- Churchill G. A. and Doerge R. W.: 2008, Naive application of permutation testing leads to inflated type I error rates, *Genetics* **178**(1), 609–10.
- Cliff D.: 1996, Answering ordinal questions with ordinal data using ordinal statistics, *Multivariate Behavioral Research* **31**, 331–350.
- Cox J. and Mann M.: 2007, Is proteomics the new genomics?, *Cell* **130**, 395–398.
- Crews D., Bergeron J. M., Bull J. J., Flores D., Tousignant A., Skipper J. K. and Wibbels T.: 1994, Temperature-dependent sex determination in reptiles: proximate mechanisms, ultimate outcomes, and practical applications, *Dev Genet* **15**(3), 297–312.
- Darvasi A. and Soller M.: 1995, Advanced intercross lines, an experimental population for fine genetic mapping, *Genetics* **141**(3), 1199–207.
- Darvasi A. and Soller M.: 1997, A simple method to calculate resolving power and confidence interval of qtl map location, *Behav Genet* **27**(2), 125–32.
- de Bono M. and Bargmann C. I.: 1998, Natural variation in a neuropeptide y receptor homolog modifies social behavior and food response in *c. elegans*, *Cell* **94**(5), 679–89.
- de Koning D. J. and Haley C. S.: 2005, Genetical genomics in humans and model organisms, *Trends Genet* **21**(7), 377–81.
- de la Fuente A., Bing N., Hoeschele I. and Mendes P.: 2004, Discovery of meaningful associations in genomic data using partial correlation coefficients, *Bioinformatics* **20**(18), 3565–74.
- DeCook R., Lall S., Nettleton D. and Howell S. H.: 2006, Genetic regulation of gene expression during shoot development in arabidopsis, *Genetics* **172**(2), 1155–64.

- Dennis, G. J., Sherman B. T., Hosack D. A., Yang J., Gao W., Lane H. C. and Lempicki R. A.: 2003, David: Database for annotation, visualization, and integrated discovery, *Genome Biol* **4**(5), P3.
- Denver D. R., Morris K., Streelman J. T., Kim S. K., Lynch M. and Thomas W. K.: 2005, The transcriptional consequences of mutation and natural selection in *caenorhabditis elegans*, *Nat Genet* **37**(5), 544–8.
- Dickson S. P., Wang K., Krantz I., Hakonarson H. and Goldstein D. B.: 2010, Rare variants create synthetic genome-wide associations, *PLoS Biol* **8**(1), e1000294.
- Dijkstra M. and Jansen R. C.: 2009, Optimal analysis of complex protein mass spectra, *Proteomics* **9**(15), 3869–76.
- Dimas A. S., Deutsch S., Stranger B. E., Montgomery S. B., Borel C., Attar-Cohen H., Ingle C., Beazley C., Gutierrez Arcelus M., Sekowska M., Gagnebin M., Nisbett J., Deloukas P., Dermitzakis E. T. and Antonarakis S. E.: 2009, Common regulatory variation impacts gene expression in a cell type-dependent manner, *Science* **325**(5945), 1246–50.
- Dixon L. K.: 1993, Use of recombinant inbred strains to map genes of aging, *Genetica* **91**(1-3), 151–65.
- Doss S., Schadt E. E., Drake T. A. and Lusis A. J.: 2005, Cis-acting expression quantitative trait loci in mice, *Genome Res* **15**(5), 681–91.
- Dubois P. C., Trynka G., Franke L., Hunt K. A., Romanos J., Curtotti A., Zhernakova A., Heap G. A., Adany R., Aromaa A., Bardella M. T., van den Berg L. H., Bockett N. A., de la Concha E. G., Dema B., Fehrmann R. S., Fernandez-Arquero M., Fiatal S., Grandone E., Green P. M., Groen H. J., Gwilliam R., Houwen R. H., Hunt S. E., Kaukinen K., Kelleher D., Korponay-Szabo I., Kurppa K., Macmathuna P., Maki M., Mazzilli M. C., McCann O. T., Mearin M. L., Mein C. A., Mirza M. M., Mistry V., Mora B., Morley K. I., Mulder C. J., Murray J. A., Nunez C., Oosterom E., Ophoff R. A., Polanco I., Peltonen L., Platteel M., Rybak A., Salomaa V., Schweizer J. J., Sperandeo M. P., Tack G. J., Turner G., Veldink J. H., Verbeek W. H., Weersma R. K., Wolters V. M., Urcelay E., Cukrowska B., Greco L., Neuhausen S. L., McManus R., Barisani D., Deloukas P., Barrett J. C., Saavalainen P., Wijmenga C. and van Heel D. A.: 2010, Multiple common variants for celiac disease influencing immune gene expression, *Nat Genet* **42**, 295–302.

- Duffy D. L. and Martin N. G.: 1994, Inferring the direction of causation in cross-sectional twin data: theoretical and empirical considerations, *Genet Epidemiol* **11**(6), 483–502.
- Emilsson V., Thorleifsson G., Zhang B., Leonardson A. S., Zink F., Zhu J., Carlson S., Helgason A., Walters G. B., Gunnarsdottir S., Mouy M., Steinthorsdottir V., Eiriksdottir G. H., Bjornsdottir G., Reynisdottir I., Gudbjartsson D., Helgadóttir A., Jonasdóttir A., Jonasdóttir A., Styrkarsdóttir U., Gretarsdóttir S., Magnusson K. P., Stefansson H., Fossdal R., Kristjansson K., Gislason H. G., Stefansson T., Leifsson B. G., Thorsteinsdóttir U., Lamb J. R., Gulcher J. R., Reitman M. L., Kong A., Schadt E. E. and Stefansson K.: 2008, Genetics of gene expression and its effect on disease, *Nature* **452**(7186), 423–8.
- Fehrmann R. S., de Jonge H. J., Ter Elst A., de Vries A., Crijns A. G., Weidenaar A. C., Gerbens F., de Jong S., van der Zee A. G., de Vries E. G., Kamps W. A., Hofstra R. M., Te Meerman G. J. and de Bont E. S.: 2008, A new perspective on transcriptional system regulation (tsr): towards tsr profiling, *PLoS One* **3**(2), e1656.
- Ferrara C. T., Wang P., Neto E. C., Stevens R. D., Bain J. R., Wenner B. R., Ilkayeva O. R., Keller M. P., Blasiolo D. A., Kendziorski C., Yandell B. S., Newgard C. B. and Attie A. D.: 2008, Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling, *PLoS Genet* **4**(3), e1000034.
- Fischer S. E., Butler M. D., Pan Q. and Ruvkun G.: 2008, Trans-splicing in *c. elegans* generates the negative miRNA regulator eri-6/7, *Nature* **455**(7212), 491–6.
- Fisher R. A.: 1947, *The design of experiments*, 4th ed. edn, Oliver and Boyd, Edinburgh.
- Fisher R. A.: 1951, *The design of experiments*, 6th edn, UK: Oliver and Boyd, London.
- Forsberg E. C., Prohaska S. S., Katzman S., Heffner G. C., Stuart J. M. and Weissman I. L.: 2005, Differential expression of novel potential regulators in hematopoietic stem cells, *PLoS Genet* **1**(3), e28.
- Fournier M. V., Carvalho P. C., Magee D. D., Carvalho M. G. C. and Appasani K.: 2007, Experimental design for gene expression analysis, *Bioarrays From Basics to Diagnostics*, Humana Press, p. 29.
- Franke L., van Bakel H., Fokkens L., de Jong E. D., Egmont-Petersen M. and Wijmenga C.: 2006, Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes, *Am J Hum Genet* **78**(6), 1011–25.

- Fu J. and Jansen R. C.: 2006, Optimal design and analysis of genetic studies on gene expression, *Genetics* **172**(3), 1993–9.
- Fu J., Keurentjes J. J., Bouwmeester H., America T., Verstappen F. W., Ward J. L., Beale M. H., de Vos R. C., Dijkstra M., Scheltema R. A., Johannes F., Koornneef M., Vreugdenhil D., Breitling R. and Jansen R. C.: 2009, System-wide molecular evidence for phenotypic buffering in arabidopsis, *Nat Genet* **41**(2), 166–7.
- Gatti D. M., Harrill A. H., Wright F. A., Threadgill D. W. and Rusyn I.: 2009, Replication and narrowing of gene expression quantitative trait loci using inbred mice, *Mamm Genome* **20**(7), 437–46.
- Ge B., Pokholok D. K., Kwan T., Grundberg E., Morcos L., Verlaan D. J., Le J., Koka V., Lam K. C., Gagne V., Dias J., Hoberman R., Montpetit A., Joly M. M., Harvey E. J., Sinnett D., Beaulieu P., Hamon R., Graziani A., Dewar K., Harmsen E., Majewski J., Goring H. H., Naumova A. K., Blanchette M., Gunderson K. L. and Pastinen T.: 2009, Global patterns of cis variation in human cells revealed by high-density allelic expression analysis, *Nat Genet* **41**(11), 1216–22.
- Gerrits A., Dykstra B., Otten M., Bystrykh L. and de Haan G.: 2008, Combining transcriptional profiling and genetic linkage analysis to uncover gene networks operating in hematopoietic stem cells and their progeny, *Immunogenetics* **60**(8), 411–22.
- Gerrits A., Li Y., Tesson B. M., Bystrykh L. V., Weersing E., Ausema A., Dontje B., Wang X., Breitling R., Jansen R. C. and de Haan G.: 2009, Expression quantitative trait loci are highly sensitive to cellular differentiation state, *PLoS Genet* **5**(10), e1000692.
- Ghazalpour A., Doss S., Kang H., Farber C., Wen P. Z., Brozell A., Castellanos R., Eskin E., Smith D. J., Drake T. A. and Lusis A. J.: 2008, High-resolution mapping of gene expression using association in an outbred mouse stock, *PLoS Genet* **4**(8), e1000149.
- Gibson G. and Dworkin I.: 2004, Uncovering cryptic genetic variation, *Nat Rev Genet* **5**(9), 681–90.
- Gibson G. and Wagner G.: 2000, Canalization in evolutionary genetics: a stabilizing theory?, *Bioessays* **22**(4), 372–80.
- Gibson G. and Weir B.: 2005, The quantitative genetics of transcription, *Trends Genet* **21**(11), 616–23.



- Goring H. H., Curran J. E., Johnson M. P., Dyer T. D., Charlesworth J., Cole S. A., Jowett J. B., Abraham L. J., Rainwater D. L., Comuzzie A. G., Mahaney M. C., Almasy L., Maccluer J. W., Kissebah A. H., Collier G. R., Moses E. K. and Blangero J.: 2007, Discovery of expression qtls using large-scale transcriptional profiling in human lymphocytes, *Nat Genet* **39**(10), 1208–16.
- GuhaThakurta D., Palomar L., Stormo G. D., Tedesco P., Johnson T. E., Walker D. W., Lithgow G., Kim S. and Link C. D.: 2002, Identification of a novel cis-regulatory element involved in the heat shock response in *caenorhabditis elegans* using microarray gene expression and computational methods, *Genome Res* **12**(5), 701–12.
- Gutteling E. W., Riksen J. A., Bakker J. and Kammenga J. E.: 2006, Mapping phenotypic plasticity and genotype-environment interactions affecting life-history traits in *caenorhabditis elegans*, *Heredity* pp. advance online publication, doi:10.1038/sj.hdy.6800894.
- Haley C. S. and de Koning D. J.: 2007, Towards in vitro genetics, *Trends Genet* **23**(8), 382–6.
- Hansen C. and Spuhler K.: 1984, Development of the national institutes of health genetically heterogeneous rat stock, *Alcohol Clin Exp Res* **8**(5), 477–9.
- Heap G. A., Trynka G., Jansen R. C., Bruinenberg M., Swertz M. A., Dinesen L. C., Hunt K. A., Wijmenga C., Vanheer D. A. and Franke L.: 2009, Complex nature of snp genotype effects on gene expression in primary human leucocytes, *BMC Med Genomics* **2**, 1.
- Hodgkin J. and Doniach T.: 1997, Natural variation and copulatory plug formation in *caenorhabditis elegans*, *Genetics* **146**(1), 149–64.
- Hoheisel J. D.: 2006, Microarray technology: beyond transcript profiling and genotype analysis, *Nat Rev Genet* **7**(3), 200–10.
- Hubner N., Wallace C. A., Zimdahl H., Petretto E., Schulz H., Maciver F., Mueller M., Hummel O., Monti J., Zidek V., Musilova A., Kren V., Causton H., Game L., Born G., Schmidt S., Muller A., Cook S. A., Kurtz T. W., Whittaker J., Pravenec M. and Aitman T. J.: 2005, Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease, *Nat Genet* **37**(3), 243–53.
- Hughes T. R., Mao M., Jones A. R., Burchard J., Marton M. J., Shannon K. W., Lefkowitz S. M., Ziman M., Schelter J. M., Meyer M. R., Kobayashi S., Davis

- C., Dai H., He Y. D., Stephaniants S. B., Cavet G., Walker W. L., West A., Coffey E., Shoemaker D. D., Stoughton R., Blanchard A. P., Friend S. H. and Linsley P. S.: 2001, Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer, *Nat Biotechnol* **19**(4), 342–7.
- Ivanova N. B., Dimos J. T., Schaniel C., Hackney J. A., Moore K. A. and Lemischka I. R.: 2002, A stem cell molecular signature, *Science* **298**(5593), 601–4.
- Iyengar S. K. and Elston R. C.: 2007, The genetic basis of complex traits: rare variants or "common gene, common disease"?, *Methods Mol Biol* **376**, 71–84.
- Jansen R. C.: 2003, Studying complex biological systems using multifactorial perturbation, *Nat Rev Genet* **4**(2), 145–51.
- Jansen R. C. and Nap J. P.: 2001, Genetical genomics: the added value from segregation, *Trends Genet* **17**(7), 388–91.
- Jansen R. C. and Nap J. P.: 2004, Regulating gene expression: surprises still in store, *Trends Genet* **20**(5), 223–5.
- Jansen R. C., Tesson B. M., Fu J., Yang Y. and McIntyre L. M.: 2009, Defining gene and qtl networks, *Curr Opin Plant Biol* **12**(2), 241–6.
- Johannes F., Colot V. and Jansen R. C.: 2008, Epigenome dynamics: a quantitative genetics perspective, *Nat Rev Genet* **9**(11), 883–90.
- Kanehisa M. and Goto S.: 2000, Kegg: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res* **28**(1), 27–30.
- Kang H. M., Ye C. and Eskin E.: 2008, Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots, *Genetics* **180**(4), 1909–25.
- Katoh Y. and Katoh M.: 2005, Comparative genomics on slit1, slit2, and slit3 orthologs, *Oncol Rep* **14**(5), 1351–5.
- Kerr M. K. and Churchill G. A.: 2001, Experimental design for gene expression microarrays, *Biostatistics* **2**(2), 183–201.
- Keurentjes J. J., Fu J., de Vos C. H., Lommen A., Hall R. D., Bino R. J., van der Plas L. H., Jansen R. C., Vreugdenhil D. and Koornneef M.: 2006, The genetics of plant metabolism, *Nat Genet* **38**(7), 842–9.

- Keurentjes J. J., Fu J., Terpstra I. R., Garcia J. M., van den Ackerveken G., Snoek L. B., Peeters A. J., Vreugdenhil D., Koornneef M. and Jansen R. C.: 2007, Regulatory network construction in arabidopsis by using genome-wide gene expression quantitative trait loci, *Proc Natl Acad Sci U S A* **104**(5), 1708–13.
- Kiel M. J., Yilmaz O. H., Iwashita T., Yilmaz O. H., Terhorst C. and Morrison S. J.: 2005, Slam family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells, *Cell* **121**(7), 1109–21.
- Kim E., Magen A. and Ast G.: 2007, Different levels of alternative splicing among eukaryotes, *Nucleic Acids Res* **35**(1), 125–31.
- Kim S. K., Lund J., Kiraly M., Duke K., Jiang M., Stuart J. M., Eizinger A., Wylie B. N. and Davidson G. S.: 2001, A gene expression map for caenorhabditis elegans, *Science* **293**(5537), 2087–92.
- Kirkpatrick S., Gelatt, C. D. J. and Vecchi M. P.: 1983, Optimization by simulated annealing, *Science* **220**(4598), 671–680.
- Kirst M., Myburg A. A., De Leon J. P., Kirst M. E., Scott J. and Sederoff R.: 2004, Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cdna microarray data in an interspecific backcross of eucalyptus, *Plant Physiol* **135**(4), 2368–78.
- Knijnenburg T. A., Wessels L. F., Reinders M. J. and Shmulevich I.: 2009, Fewer permutations, more accurate p-values, *Bioinformatics* **25**(12), i161–8.
- Koch R., van Luenen H. G., van der Horst M., Thijssen K. L. and Plasterk R. H.: 2000, Single nucleotide polymorphisms in wild isolates of caenorhabditis elegans, *Genome Res* **10**(11), 1690–6.
- Kose F., Weckwerth W., Linke T. and Fiehn O.: 2001, Visualizing plant metabolomic correlation networks using clique-metabolite matrices, *Bioinformatics* **17**(12), 1198–208.
- Kover P. X., Valdar W., Trakalo J., Scarcelli N., Ehrenreich I. M., Purugganan M. D., Durrant C. and Mott R.: 2009, A multiparent advanced generation inter-cross to fine-map quantitative traits in arabidopsis thaliana, *PLoS Genet* **5**(7), e1000551.
- Kraus P. R., Boily M. J., Giles S. S., Stajich J. E., Allen A., Cox G. M., Dietrich F. S., Perfect J. R. and Heitman J.: 2004, Identification of cryptococcus neoformans temperature-regulated genes with a genomic-dna microarray, *Eukaryot Cell* **3**(5), 1249–60.

- Kulp D. C. and Jagalur M.: 2006, Causal inference of regulator-target pairs by gene mapping of expression phenotypes, *BMC Genomics* **7**, 125.
- Kuroyanagi H., Ohno G., Mitani S. and Hagiwara M.: 2007, The fox-1 family and sup-12 coordinately regulate tissue-specific alternative splicing in vivo, *Mol Cell Biol* **27**(24), 8612–21.
- Kwan T., Benovoy D., Dias C., Gurd S., Provencher C., Beaulieu P., Hudson T. J., Sladek R. and Majewski J.: 2008, Genome-wide analysis of transcript isoform variation in humans, *Nat Genet* **40**(2), 225–31.
- Lam A. C., Fu J., Jansen R. C., Haley C. S. and de Koning D. J.: 2008, Optimal design of genetic studies of gene expression with two-color microarrays in outbred crosses, *Genetics* **180**(3), 1691–8.
- Lan H., Chen M., Flowers J. B., Yandell B. S., Stapleton D. S., Mata C. M., Mui E. T., Flowers M. T., Schueler K. L., Manly K. F., Williams R. W., Kendziorski C. and Attie A. D.: 2006, Combined expression trait correlations and expression quantitative trait locus mapping, *PLoS Genet* **2**(1), e6.
- Landry C. R., Oh J., Hartl D. L. and Cavalieri D.: 2006, Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes, *Gene* **366**(2), 343–51.
- Le Rouzic A. and Carlborg O.: 2008, Evolutionary potential of hidden genetic variation, *Trends Ecol Evol* **23**(1), 33–7.
- Leek J. T. and Storey J. D.: 2007, Capturing heterogeneity in gene expression studies by surrogate variable analysis, *PLoS Genet* **3**(9), 1724–35.
- Levins R.: 2004, Toward a population biology, still, in M. K. U. Rama S. Singh (ed.), *The evolution of Population Biology*, Cambridge University Press.
- Li R., Lyons M. A., Wittenburg H., Paigen B. and Churchill G. A.: 2005, Combining data from multiple inbred line crosses improves the power and resolution of quantitative trait loci mapping, *Genetics* **169**(3), 1699–709.
- Li R., Tsaih S. W., Shockley K., Stylianou I. M., Wergedal J., Paigen B. and Churchill G. A.: 2006a, Structural model analysis of multiple quantitative traits, *PLoS Genet* **2**(7), e114.
- Li Y., Alvarez O. A., Gutteling E. W., Tijsterman M., Fu J., Riksen J. A., Hazendonk E., Prins P., Plasterk R. H., Jansen R. C., Breitling R. and Kammenga J. E.: 2006b,

- Mapping determinants of gene expression plasticity by genetical genomics in *c. elegans*, *PLoS Genet* **2**(12), e222.
- Li Y., Breitling R. and Jansen R. C.: 2008, Generalizing genetical genomics: getting added value from environmental perturbation, *Trends Genet* **24**(10), 518–24.
- Li Y., Breitling R., Snoek L. B., van der Velde K. J., Swertz M. A., Riksen J., Jansen R. C. and Kammenga J. E.: 2010a, Global genetic robustness of the alternative splicing machinery in *c. elegans*, *Genetics* **186**, 405.
- Li Y., Swertz M. A., Vera G., Fu J., Breitling R. and Jansen R. C.: 2009, designgg: an r-package and web tool for the optimal design of genetical genomics experiments, *BMC Bioinformatics* **10**, 188.
- Li Y., Tesson B. M., Churchill G. A. and Jansen R. C.: 2010b, Critical reasoning on causal inference in genome-wide linkage and association studies, *Trends in Genetics* (**in press**).
- Liang Y., Jansen M., Aronow B., Geiger H. and Van Zant G.: 2007, The quantitative trait gene latexin influences the size of the hematopoietic stem cell population in mice, *Nat Genet* **39**(2), 178–88.
- Liu B., de la Fuente A. and Hoeschele I.: 2008, Gene network inference via structural equation modeling in genetical genomics experiments, *Genetics* **178**(3), 1763–76.
- Luscombe N. M., Babu M. M., Yu H., Snyder M., Teichmann S. A. and Gerstein M.: 2004, Genomic analysis of regulatory network dynamics reveals large topological changes, *Nature* **431**(7006), 308–12.
- Mackay T. F., Stone E. A. and Ayroles J. F.: 2009, The genetics of quantitative traits: challenges and prospects, *Nat Rev Genet* **10**(8), 565–77.
- Madi A., Mikkat S., Ringel B., Ulbrich M., Thiesen H. J. and Glocker M. O.: 2003, Mass spectrometric proteome analysis for profiling temperature-dependent changes of protein expression in wild-type *caenorhabditis elegans*, *Proteomics* **3**(8), 1526–34.
- Maere S., Heymans K. and Kuiper M.: 2005, Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks, *Bioinformatics* **21**(16), 3448–9.
- Manly K. F., Wang J. and Williams R. W.: 2005, Weighting by heritability for detection of quantitative trait loci with microarray estimates of gene expression, *Genome Biol* **6**(3), R27.

- Maydan J. S., Flibotte S., Edgley M. L., Lau J., Selzer R. R., Richmond T. A., Pofahl N. J., Thomas J. H. and Moerman D. G.: 2007, Efficient high-resolution deletion discovery in *caenorhabditis elegans* by array comparative genomic hybridization, *Genome Res* **17**(3), 337–47.
- McClurg P., Janes J., Wu C., Delano D. L., Walker J. R., Batalov S., Takahashi J. S., Shimomura K., Kohsaka A., Bass J., Wiltshire T. and Su A. I.: 2007, Genomewide association analysis in diverse inbred mice: power and population structure, *Genetics* **176**(1), 675–83.
- Mead R.: 1988, *The design of experiments*, Cambridge University Press, Cambridge.
- Mehrabian M., Allayee H., Stockton J., Lum P. Y., Drake T. A., Castellani L. W., Suh M., Armour C., Edwards S., Lamb J., Lusic A. J. and Schadt E. E.: 2005, Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits, *Nat Genet* **37**(11), 1224–33.
- Millstein J., Zhang B., Zhu J. and Schadt E. E.: 2009, Disentangling molecular relationships with a causal inference test, *BMC Genet* **10**, 23.
- Moffatt M. F., Kabesch M., Liang L., Dixon A. L., Strachan D., Heath S., Depner M., von Berg A., Bufe A., Rietschel E., Heinzmann A., Simma B., Frischer T., Willis-Owen S. A., Wong K. C., Illig T., Vogelberg C., Weiland S. K., von Mutius E., Abecasis G. R., Farrall M., Gut I. G., Lathrop G. M. and Cookson W. O.: 2007, Genetic variants regulating *ormdl3* expression contribute to the risk of childhood asthma, *Nature* **448**(7152), 470–3.
- Monks S. A., Leonardson A., Zhu H., Cundiff P., Pietrusiak P., Edwards S., Phillips J. W., Sachs A. and Schadt E. E.: 2004, Genetic inheritance of gene expression in human cell lines, *Am J Hum Genet* **75**(6), 1094–105.
- Morley M., Molony C. M., Weber T. M., Devlin J. L., Ewens K. G., Spielman R. S. and Cheung V. G.: 2004, Genetic analysis of genome-wide variation in human gene expression, *Nature* **430**(7001), 743–7.
- Nelson L. S., Rosoff M. L. and Li C.: 1998, Disruption of a neuropeptide gene, *flp-1*, causes multiple behavioral defects in *caenorhabditis elegans*, *Science* **281**(5383), 1686–90.
- Ozbudak E. M., Thattai M., Kurtser I., Grossman A. D. and van Oudenaarden A.: 2002, Regulation of noise in the expression of a single gene, *Nat Genet* **31**(1), 69–73.

- Pearl J.: 2000, *Causality: models, reasoning, and inference*, 2nd ed edn, Cambridge University Press.
- Peirce J. L., Lu L., Gu J., Silver L. M. and Williams R. W.: 2004, A new set of bxd recombinant inbred lines from advanced intercross populations in mice, *BMC Genet* **5**, 7.
- Peng J., Wang P. and Tang H.: 2007, Controlling for false positive findings of trans-hubs in expression quantitative trait loci mapping, *BMC Proc* **1 Suppl 1**, S157.
- Perez-Enciso M.: 2004, In silico study of transcriptome genetic variation in outbred populations, *Genetics* **166**(1), 547–54.
- Petretto E., Mangion J., Dickens N. J., Cook S. A., Kumaran M. K., Lu H., Fischer J., Maatz H., Kren V., Pravenec M., Hubner N. and Aitman T. J.: 2006, Heritability and tissue specificity of expression quantitative trait loci, *PLoS Genet* **2**(10), e172.
- Pigliucci M., Murren C. J. and Schlichting C. D.: 2006, Phenotypic plasticity and evolution by genetic assimilation, *J Exp Biol* **209**(Pt 12), 2362–7.
- Plaisier C. L., Horvath S., Huertas-Vazquez A., Cruz-Bautista I., Herrera M. F., Tusie-Luna T., Aguilar-Salinas C. and Pajukanta P.: 2009, A systems genetics approach implicates *usf1*, *fads3*, and other causal candidate genes for familial combined hyperlipidemia, *PLoS Genet* **5**(9), e1000642.
- Ptacek J., Devgan G., Michaud G., Zhu H., Zhu X., Fasolo J., Guo H., Jona G., Breitzkreutz A., Sopko R., McCartney R. R., Schmidt M. C., Rachidi N., Lee S. J., Mah A. S., Meng L., Stark M. J., Stern D. F., De Virgilio C., Tyers M., Andrews B., Gerstein M., Schweitzer B., Predki P. F. and Snyder M.: 2005, Global analysis of protein phosphorylation in yeast, *Nature* **438**(7068), 679–84.
- Queitsch C., Sangster T. A. and Lindquist S.: 2002, Hsp90 as a capacitor of phenotypic variation, *Nature* **417**(6889), 618–24.
- Ravasz E., Somera A. L., Mongru D. A., Oltvai Z. N. and Barabasi A. L.: 2002, Hierarchical organization of modularity in metabolic networks, *Science* **297**(5586), 1551–5.
- Rockman M. V. and Kruglyak L.: 2006, Genetics of global gene expression, *Nat Rev Genet* **7**(11), 862–72.
- Rockman M. V. and Kruglyak L.: 2008, Breeding designs for recombinant inbred advanced intercross lines, *Genetics* **179**(2), 1069–78.

- Roff D. A.: 2002, *Life History Evolution*, Sinauer Associates Inc, Sunderland.
- Rogers C., Reale V., Kim K., Chatwin H., Li C., Evans P. and de Bono M.: 2003, Inhibition of caenorhabditis elegans social feeding by fmrfamide-related peptide activation of npr-1, *Nat Neurosci* **6**(11), 1178–85.
- Rosa G. J., de Leon N. and Rosa A. J.: 2006, Review of microarray experimental design strategies for genetical genomics studies, *Physiol Genomics* **28**(1), 15–23.
- Roskam J. C. and Brakefield P. M.: 1996, A comparison of temperature-induced polyphenism in african bicyclus butterflies from a seasonal savannah-rainforest ecotone., *Evolution* **50**, 2360–2372.
- Ruden D. M., Chen L., Possidente D., Possidente B., Rasouli P., Wang L., Lu X., Garfinkel M. D., Hirsch H. V. and Page G. P.: 2009, Genetical toxicogenomics in drosophila identifies master-modulatory loci that are regulated by developmental exposure to lead, *Neurotoxicology* **30**(6), 898–914.
- Rutherford S. L. and Lindquist S.: 1998, Hsp90 as a capacitor for morphological evolution, *Nature* **396**(6709), 336–42.
- Sargon J.: 1958, The estimation of economic relationships using instrumental variables., *Econometrica* **26**, 393–415.
- Schadt E. E., Lamb J., Yang X., Zhu J., Edwards S., Guhathakurta D., Sieberts S. K., Monks S., Reitman M., Zhang C., Lum P. Y., Leonardson A., Thieringer R., Metzger J. M., Yang L., Castle J., Zhu H., Kash S. F., Drake T. A., Sachs A. and Lusis A. J.: 2005, An integrative genomics approach to infer causal associations between gene expression and disease, *Nat Genet* **37**(7), 710–7.
- Schadt E. E., Molony C., Chudin E., Hao K., Yang X., Lum P. Y., Kasarskis A., Zhang B., Wang S., Suver C., Zhu J., Millstein J., Sieberts S., Lamb J., GuhaThakurta D., Derry J., Storey J. D., Avila-Campillo I., Kruger M. J., Johnson J. M., Rohl C. A., van Nas A., Mehrabian M., Drake T. A., Lusis A. J., Smith R. C., Guengerich F. P., Strom S. C., Schuetz E., Rushmore T. H. and Ulrich R.: 2008, Mapping the genetic architecture of gene expression in human liver, *PLoS Biol* **6**(5), e107.
- Schadt E. E., Monks S. A., Drake T. A., Lusis A. J., Che N., Colinayo V., Ruff T. G., Milligan S. B., Lamb J. R., Cavet G., Linsley P. S., Mao M., Stoughton R. B. and Friend S. H.: 2003, Genetics of gene expression surveyed in maize, mouse and man, *Nature* **422**(6929), 297–302.



- Sieberts S. K. and Schadt E. E.: 2007, Moving toward a system genetics view of disease, *Mamm Genome* **18**(6-7), 389–401.
- Smith E. N. and Kruglyak L.: 2008, Gene-environment interaction in yeast gene expression, *PLoS Biol* **6**(4), e83.
- Solberg L. C., Baum A. E., Ahmadiyeh N., Shimomura K., Li R., Turek F. W., Churchill G. A., Takahashi J. S. and Redei E. E.: 2004, Sex- and lineage-specific inheritance of depression-like behavior in the rat, *Mamm Genome* **15**(8), 648–62.
- Spirtes P., Glymour C. and Scheines R.: 1993, *Causation, Prediction, and Search*, Springer-Verlag, New York.
- Stegle O., Parts L., Durbin R. and Winn J.: 2010, A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies, *PLoS Comput Biol* **6**(5), e1000770.
- Stephens M. and Balding D. J.: 2009, Bayesian statistical methods for genetic association studies, *Nat Rev Genet* **10**(10), 681–90.
- Steuer R.: 2006, Review: on the analysis and interpretation of correlations in metabolomic data, *Brief Bioinform* **7**(2), 151–8.
- Storey J. D. and Tibshirani R.: 2003, Statistical significance for genomewide studies, *Proc Natl Acad Sci U S A* **100**(16), 9440–5.
- Stranger B. E., Forrest M. S., Clark A. G., Minichiello M. J., Deutsch S., Lyle R., Hunt S., Kahl B., Antonarakis S. E., Tavaré S., Deloukas P. and Dermitzakis E. T.: 2005, Genome-wide associations of gene expression variation in humans, *PLoS Genet* **1**(6), e78.
- Stylianou I. M., Affourtit J. P., Shockley K. R., Wilpan R. Y., Abdi F. A., Bhardwaj S., Rollins J., Churchill G. A. and Paigen B.: 2008, Applying gene expression, proteomics and single-nucleotide polymorphism analysis for complex trait gene identification, *Genetics* **178**(3), 1795–805.
- Swertz M. A., De Brock E. O., Van Hijum S. A., De Jong A., Buist G., Baerends R. J., Kok J., Kuipers O. P. and Jansen R. C.: 2004, Molecular genetics information system (molgenis): alternatives in developing local experimental genomics databases, *Bioinformatics* **20**(13), 2075–83.
- Swertz M. A. and Jansen R. C.: 2007, Beyond standardization: dynamic software infrastructures for systems biology, *Nat Rev Genet* **8**(3), 235–43.

- Tesson B. M., Breitling R. and Jansen R. C.: 2010, Diffcoex: a simple and sensitive method to find differentially coexpressed gene modules, *submitted*.
- Visscher P. M., Hill W. G. and Wray N. R.: 2008, Heritability in the genomics era—concepts and misconceptions, *Nat Rev Genet* **9**(4), 255–66.
- Wang B., Xiao Y., Ding B. B., Zhang N., Yuan X., Gui L., Qian K. X., Duan S., Chen Z., Rao Y. and Geng J. G.: 2003a, Induction of tumor angiogenesis by slit-robo signaling and inhibition of cancer growth by blocking robo activity, *Cancer Cell* **4**(1), 19–29.
- Wang J., Williams R. W. and Manly K. F.: 2003b, Webqtl: web-based complex trait analysis, *Neuroinformatics* **1**(4), 299–308.
- Wang K. H., Brose K., Arnott D., Kidd T., Goodman C. S., Henzel W. and Tessier-Lavigne M.: 1999, Biochemical purification of a mammalian slit protein as a positive regulator of sensory axon elongation and branching, *Cell* **96**(6), 771–84.
- Wang S., Yehya N., Schadt E. E., Wang H., Drake T. A. and Lusis A. J.: 2006, Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity, *PLoS Genet* **2**(2), e15.
- Wang S., Zheng T. and Wang Y.: 2007, Transcription activity hot spot, is it real or an artifact?, *BMC Proc* **1 Suppl 1**, S94.
- webCite: 2010, Pubmed search 17-05-2010 for “eqtl” or “genetical genomics”.
- Wessel J., Zapala M. A. and Schork N. J.: 2007, Accommodating pathway information in expression quantitative trait locus analysis, *Genomics* **90**(1), 132–42.
- West M. A., van Leeuwen H., Kozik A., Kliebenstein D. J., Doerge R. W., St Clair D. A. and Michelmore R. W.: 2006, High-density haplotyping with microarray-based expression and single feature polymorphism markers in arabidopsis, *Genome Res* **16**(6), 787–95.
- Whiteley A. R., Derome N., Rogers S. M., St-Cyr J., Laroche J., Labbe A., Nolte A., Renault S., Jeukens J. and Bernatchez L.: 2008, The phenomics and expression quantitative trait locus mapping of brain transcriptomes regulating adaptive divergence in lake whitefish species pairs (*coregonus* sp.), *Genetics* **180**(1), 147–64.
- Whitlock M. C.: 2005, Combining probability from independent tests: the weighted z-method is superior to fisher’s approach, *J Evol Biol* **18**(5), 1368–73.

- Wicks S. R., Yeh R. T., Gish W. R., Waterston R. H. and Plasterk R. H.: 2001, Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map, *Nat Genet* **28**(2), 160–4.
- Wit E. and McClure J. D.: 2004, *Statistics for Microarrays; Design, Analysis and Inference*, John Wiley Sons, Chichester.
- Wit E., Nobile A. and Khanin R.: 2005a, Near-optimal designs for dual-channel microarray studies, *Applied Statistics* **54**(5), 817–30.
- Wit E., Nobile A. and Khanin R.: 2005b, Simulated annealing for near-optimal dual-channel microarray designs, *Appl Statistics* **54**, 817–830.
- Woo Y., Krueger W., Kaur A. and Churchill G.: 2005, Experimental design for three-color and four-color gene expression microarrays, *Bioinformatics* **21 Suppl 1**, i459–67.
- Wright S.: 1921, Correlation and causation, *J. Agric. Res* **20**, 557–585.
- Wu C., Delano D. L., Mitro N., Su S. V., Janes J., McClurg P., Batalov S., Welch G. L., Zhang J., Orth A. P., Walker J. R., Glynne R. J., Cooke M. P., Takahashi J. S., Shimomura K., Kohsaka A., Bass J., Saez E., Wiltshire T. and Su A. I.: 2008, Gene set enrichment in eQTL data identifies novel annotations and pathway regulators, *PLoS Genet* **4**(5), e1000070.
- Wu J. Y., Feng L., Park H. T., Havlioglu N., Wen L., Tang H., Bacon K. B., Jiang Z., Zhang X. and Rao Y.: 2001, The neuronal repellent slit inhibits leukocyte chemotaxis induced by chemotactic factors, *Nature* **410**(6831), 948–52.
- Xue Y., Haas S. A., Brino L., Gusnanto A., Reimers M., Talibi D., Vingron M., Ekwall K. and Wright A. P.: 2004, A dna microarray for fission yeast: minimal changes in global gene expression after temperature shift, *Yeast* **21**(1), 25–39.
- Yang Y. H., Dudoit S., Luu P., Lin D. M., Peng V., Ngai J. and Speed T. P.: 2002, Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res* **30**(4), e15.
- Yang Y. H. and Speed T.: 2002, Design issues for cDNA microarray experiments, *Nat Rev Genet* **3**(8), 579–88.
- Yu Y., Maroney P. A., Denker J. A., Zhang X. H., Dybkov O., Luhrmann R., Jankowsky E., Chasin L. A. and Nilsen T. W.: 2008, Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition, *Cell* **135**(7), 1224–36.

- Yvert G., Brem R. B., Whittle J., Akey J. M., Foss E., Smith E. N., Mackelprang R. and Kruglyak L.: 2003, Trans-acting regulatory variation in *saccharomyces cerevisiae* and the role of transcription factors, *Nat Genet* **35**(1), 57–64.
- Zhang B. and Horvath S.: 2005, A general framework for weighted gene co-expression network analysis, *Stat Appl Genet Mol Biol* **4**, Article17.
- Zhao Z., Boyle T. J., Bao Z., Murray J. I., Mericle B. and Waterston R. H.: 2008, Comparative analysis of embryonic cell lineage between *caenorhabditis briggsae* and *caenorhabditis elegans*, *Dev Biol* **314**(1), 93–9.
- Zhu J., Lum P. Y., Lamb J., GuhaThakurta D., Edwards S. W., Thieringer R., Berger J. P., Wu M. S., Thompson J., Sachs A. B. and Schadt E. E.: 2004, An integrative genomics approach to the reconstruction of gene networks in segregating populations, *Cytogenet Genome Res* **105**(2-4), 363–74.
- Zhu J., Wiener M. C., Zhang C., Fridman A., Minch E., Lum P. Y., Sachs J. R. and Schadt E. E.: 2007, Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations, *PLoS Comput Biol* **3**(4), e69.
- Zhu J., Zhang B., Smith E. N., Drees B., Brem R. B., Kruglyak L., Bumgarner R. E. and Schadt E. E.: 2008, Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks, *Nat Genet* **40**(7), 854–61.

---

## Samenvatting

Generalized genetical genomics (GGG) is een methode op het gebied van de systeemgenetica. Het combineert de analyse van genetische variatie met variatie in moleculaire eigenschappen op populatieniveau, in verschillende omgevingen, met als doel inzicht te krijgen in de interacties tussen genotype en omgevingsinvloeden.

Dit proefschrift begint met het uiteenzetten van het GGG principe (Hoofdstuk 1). Daarna presenteren wij een nieuw computer programma, designGG, voor het zo goed mogelijk ontwerpen van GGG experimenten (Hoofdstuk 2).

In het vervolg gaan wij in op de meest geopperde kritieken op methoden voor het bepalen van causale relaties met genetische data. Verder hebben we verschillende permutatiemethodieken (voor het bepalen van significantie van 'quantitative trait loci' (QTL) en QTL hotspots) onderzocht (Hoofdstuk 3-4).

Verder hebben we de GGG methode toegepast tijdens drie pilootstudies: In de eerste studie laten we zien dat erfelijke verschillen in de plasticiteit van genexpressie grotendeels gereguleerd worden door genetische factoren *in trans*. In de tweede studie laten we zien dat erfelijke verschillen in genexpressie erg gevoelig zijn ten aanzien van het cellulaire ontwikkelingsstadium. In de derde studie, met behulp van genoom-brede *tiling arrays*, vonden wij dat de alternative splicing mechanisme in *C. elegans* in het algemeen genetisch robuust lijkt te zijn en dat slechts een fractie van de genen erfelijke variatie in hun *splice* patronen laat zien. (Hoofdstuk 5-7).

Ten slotte discussiëren wij enkele fundamentele uitdagingen met betrekking tot databewerking, QTL mapping, en netwerkreconstructie. Hier doen wij ook suggesties voor vervolgonderzoek om de reikwijdte van de systeemgenetica verder te vergroten (Hoofdstuk 8).



---

## Acknowledgements

First of all, I would like to express my sincere gratitude to Prof. Ritsert Jansen. It is a pleasure and an honor to have had the opportunity to work with him. His inspiration, motivation, infectious enthusiasm and immense knowledge guided me throughout my period of research. His understanding, openness and kindness also provided a good basis for the present thesis. I could not have imagined having a better supervisor for my Ph.D. study.

My sincere thanks also go to Prof. Rainer Breitling. He arrived in Groningen at a crucial initial stage in my research and provided tremendously stimulating discussions and insightful suggestions. The door to Prof. Breitlings office has always been open, whenever I ran into difficulties or had a question. Thank you for all the support, challenges and encouragement.

I wish to express my warmest thanks to our collaborators on this project: without them this thesis would not have been possible. They are Prof. Jan Kammenga, Dr. Olga Alvarez, Dr. Basten Snoek, Ana Vinuela, Miriam Rodriguez-Sanchez from Wageningen University; Prof. Gerald de Haan, Dr. Lenja Bystrykh, Alice Gerrits from the University Medical Center Groningen; Prof. Gary Churchill and Dr. Ron Korstanje from the Jackson Laboratory; and Prof. Michael Biehl from the Computing Science Department of the University of Groningen. I am also very grateful to Prof. Siem Heisterkamp, Prof. Ernst Wit, Prof. Harold Snieder and Dr. Peter Terpstra, all at Groningen University, for their valuable advice and comments on my research.

Without the amazing administrative help of Klazien Offens, my GBiC life would have been quite difficult. Thank you very, very much for your efficient work and the true friendship of you and your family.

I am thankful beyond words to everyone in our research group, current and former members with whom I had the pleasure to work: Morris Swertz, Martijn Dijkstra, Rudi Alberts, Jingyuan Fu, Richard Scheltema, Bruno Tesson, Gonzalo Vera, Joeri van de Velde, Frank Johannes, Tauqeer Alam, Anna Kolesnichenko, Elena Merlo, Andris Jankevics, Marnix Medema, Danny Arends, Rene Wardenaar, George

Byelas, Erik Roos, Despiona Antonakaki and Nino Demetrashvili. Your scientific input, but even more importantly the many memorable moments we shared, will always be cherished by me.

I have also been lucky to meet a lot of wonderful friends during my time in Groningen: Gaifen, He Tao and Erwin, Wu Bian and Qiqi, Calvin, Qu Ning and Yijin, Qu Yang, Zhao Bo and Xiaoyan, Ding Ning, Yu Jinyan, Wu Yingxiang and Wouter, Liying, Xiaowen, Du Xiaoguang, Xiao Xiao and Zhang Jinhua, Xie Dan, Chen Yang, Jiang Likun, Yu Zhilin, Wang Renxuan and many other friends. I enjoyed all the time we spent together on badminton, card games, dinners and outings. Your generous help to me and my family will always be appreciated by us.

A special hearty thanks to my parents, for their unending love, support and guidance over the years. I am grateful to my younger sister and brother-in-law for taking good care of my parents during my study abroad. I would like to thank my most beloved daughter, Yuanxin, for granting me joy and happiness of a magnitude I never knew before. I apologize to her for putting her in the after-school care during the endless hours I worked away from her. I hope she can understand and forgive me. Finally, my deepest gratitude goes to my husband, Chengjian. I am endlessly indebted and grateful to you for everything. My love, thank you for helping me in all aspects of my life. Thank you for your unconditional love and being a strong pillar of support in the unavoidable ups and downs of producing this thesis. Thank you also for sharing the moments of happiness and victory with me, which makes them even more memorable.

李扬

Yang Li  
Groningen  
Sept. 24, 2010



---

## Publications

- **Li Y\***, Tesson BM\*, Churchill GA and Jansen RC (2010) *Critical reasoning on causal inference in genome-wide linkage and association studies* **Trend Genet** (in press). \* Equal contribution
- Tesson BM\*, **Li Y\***, Breitling R, and Jansen RC (2010) *Scaling up classical genetics to thousands of molecular traits: promises and challenges* **Proceedings of the 9th World Congress of Genetics Applied to Livestock Production** 0985.  
\* Equal contribution
- **Li Y**, Breitling R, Snoek B, van der Velde KJ, Swertz M, Riksen J, Jansen RC and Kammenga JEC (2010) *Global genetic robustness of the alternative splicing machinery in C. elegans*, **Genetics** 186:405.
- **Li Y**, Swertz M, Vera G, Fu J, Breitling R and Jansen RC (2009) *designGG: an R-package and Web tool for the optimal design of genetical genomics experiments*, **BMC Bioinformatics** 10:188.  
\*\* Highly accessed paper
- Gerrits A\*, **Li Y\***, Tesson BM\*, Bystrykh LV, Weersing E, Dethmers-Ausema B, Dontje B, Wang X, Breitling R, Jansen RC and de Haan G (2009) *Expression quantitative trait loci are highly sensitive to cellular differentiation state*, **PLoS genetics** 5:e1000692.  
\* Equal contribution  
\*\* Highlight in **Nature Rev Genet** 10:819
- **Li Y**, Breitling R and Jansen RC (2008) *Generalized genetical genomics - the added value from experimental perturbation*, **Trend Genet** 24:518–524

- **Li Y\***, Alvarez OA\*, Gutteling EW, Tijsterman M, Fu J, Riksen JAG, Hazendonk E, Prins P, Plasterk RHA, Jansen RC, Breitling R and Kammenga JEC (2006) *Mapping determinants of gene expression plasticity by genetical genomics in C. elegans*, **PLoS Genetics** 2:e222. doi:10.1371/journal.pgen.0020222.  
\* Equal contribution  
\*\* Faculty of 1000 Biology EXCEPTIONAL paper
- Breitling R, **Li Y**, Tesson B, Fu J, Wu C, Wiltshire T, Gerrits A, Bystrykh LV, de Haan G, Su AI and Jansen RC (2008) *Genetical genomics: spotlight on QTL hotspots*, **PLoS Genetics** 4:e1000232. doi:10.1371/journal.pgen.1000232
- Biehl M, Breitling R and **Li Y**, (2007) *Analysis of tiling micorarray data by learning vector quantization and relevance learning*, **IDEAL: Lecture Notes in Computer Science** 4881: 880-889
- Albert R, Terpstra P, **Li Y**, Breitling R, Nap JP and Jansen RC (2007) *Sequence polymorphisms cause many false cis eQTLs*, **PLoS One** 7: e622. doi: 10.1371/journal.pone.0000622
- Korstanje R, Desai J, Lazar G, Rollins J, Spurr M, Joseph J, Kadambi S, **Li Y**, Cherry A, Paigen B and Millonig JH (2008) *Localization of modifier loci of the vacuolated lens mutant, a mouse model for spina bifida and congenital cataracts*, **Physiol Genomics** 35: 296-304

## Awards

- 2008 Chinese Government Award for Outstanding Student Abroad
- 2009 First Prize for Best Poster at 17th Annual GBB Symposium, Groningen, The Netherlands

## Publications before PhD study

- Li Y, Jiang JH, Wu HL, Chen ZP and Yu RQ (2000) *Alternating Coupled Matrices Resolution Method for Three-way Arrays Analysis*, **Chemom. Intell. Lab. Syst.** 52: 33-43
- Li Y, Jiang JH, Chen ZP, Xu CJ, and Yu RQ (1999) *Robust Linear Discriminant Analysis for Pattern Recognition*, **J. Chemom.** 13: 3-13
- Li Y, Jiang JH, Chen ZP, Xu CJ, and Yu RQ (1999) *A New Method Based on Counter-propagation Network Algorithm for Chemical Pattern Recognition*, **Anal. Chimi. Acta** 388: 161-170
- Xu CJ, Liang YZ, Li Y and Du YP (2003) *Chemical Rank Estimation by Noise Perturbation in Functional Principal Component Analysis*. **Analyst** 128: 75-81.
- Chen ZP, Li Y, and Yu RQ (2001) *Pseudo Alternating Least Squares Algorithm for Trilinear Decomposition*, **J. Chemom.** 15: 149-167.
- Chen ZP, Wu HL, Li Y, and Yu RQ (2000) *Novel Constrained PARAFAC Algorithm for Second Order Linear Calibration*, **Anal. Chimi. Acta** 423: 187-196.
- Jiang JH, Wu HL, Li Y, and Yu RQ (2000) *Three-way Data Resolution by Alternating Slice-wise Diagonalization Method*, **J. Chemom.** 14: 15-36.
- Chen ZP, Wu HL, Jiang JH, Li Y, and Yu RQ (2000) *A Novel Trilinear Decomposition Algorithm for Second-order Linear Calibration*, **Chemom. Intell. Lab. Syst.** 52: 75-86.
- Chen ZP, Jiang JH, Li Y and Yu RQ (1999) *Nonlinear Mapping Using Real-valued Genetic Algorithm*, **Chemom. Intell. Lab. Syst.**, 45: 409-418.
- Chen ZP, Jiang JH, Li Y, Shen HL, Liang YZ and Yu RQ (1999) *Smoothed Window Factor Analysis*, **Anal. Chimi. Acta** 381: 233-246
- Jiang JH, Wu HL, Li Y, and Yu RQ (1999) *Alternating Coupled Vectors Resolution For Trilinear Analysis of Three-way Data*, **J. Chemom.** 13: 557-578.
- Chen ZP, Liang YZ, Jiang JH, Li Y, Qian JY and Yu RQ (1999) *Determination of the Number of Components in Mixture Using a New Approach Incorporating Chemical Information*, **J. Chemom.** 13: 15-30.

- Chen ZP, Jiang JH, Li Y, Liang YZ and Yu RQ (1999) *Fuzzy Linear Discriminant Analysis for Chemical Data sets*, **Chemom. Intell. Lab. Syst.** 45: 295-302.
- Wang JH, Jiang JH, Xiong JF, Li Y, Liang YZ and Yu RQ (1998) *Chemical Rank Estimation for Excitation-emission Matrices Using a Morphological Approach*, **J. Chemom.** 12: 95-104.

## Selected conference contributions

- Aug 1-6, 2010 **9th World congress on Genetics applied to livestock production** (Leipzig, Germany), Invited talk: *Scaling up classical genetics to thousands of molecular traits: promises and challenges*
- Nov 2, 2009 **UK Plant QTL Workshop** (Warwick, UK). Invited talk: *eQTLs and QTL network*
- Oct 1–2, 2009 **Systems Genetics: from man to microbe, from genotype to phenotype** (Groningen, NL). Selected talk: *Causal inference in genome-wide association and linkage studies: the road ahead to systems genetics*
- Sept 21–23, 2009 **European Conference on Complex Systems 2009** (Warwick, UK). Selected talk: *Causal inference in genome-wide association and linkage studies: the road ahead to systems genetics*
- May 3–5, 2009 **8th Annual Complex Trait Consortium Conference** (Manchester, UK). Selected talk: *Critical reasoning on causality: extreme accuracy and large sample sizes will be needed in systems genetics*
- June 1–3, 2008 **7th Annual Complex Trait Consortium Conference** (Montreal, Canada). Selected talk: *Generalized genetical genomics – the added value from experimental perturbation*
- May 31, 2008 **Quebec Transgenic Research Network– satellite symposium** (McGill University, Canada) Invited talk: *Generalized genetical genomics*
- March 25–30, 2007 **Similarity-based Clustering Seminar** (Schloss Dagstuhl, Germany). Selected talk: *Analyzing genome tiling microarrays for the detection of novel expressed genes*
- Oct 17–18, 2006 **2nd BENELUX Bioinformatics conference** (Wageningen, The Netherlands). Selected talk: *Mapping Determinants of Gene Expression Plasticity by Genetical Genomics in C. elegans*

---

## Curriculum vitae

Yang Li was born on 4 October 1974 in the Hubei province of China. In 1992 she started to study chemistry at Hunan University and obtained a Bachelor in Science degree in 1996. During her studies, she was awarded as Excellent Student and received a Top Student scholarship. After that she began to study Chemometrics in the group of Prof. Ruqin Yu (member of the Chinese Academy of Science). In 1999 she received her Master of Science degree, with a thesis on “New algorithms in three-way data analysis and chemical pattern recognition and their applications in chemistry”. In the following five years she worked at Hunan University as a lecturer and continued her research. She published more than 15 papers in international journals during this period, and obtained the first-class award for outstanding teachers in 2003. In 2005, she continued her academic career by moving from China to The Netherlands, and from the field of Chemometrics to a new and challenging field: Bioinformatics. She performed a PhD project under the supervision of Prof. Dr. Ritsert C. Jansen at the University of Groningen, working on “Genetical genomics – advanced methods and applications” funded by the Netherlands Organization for Scientific Research (NWO-86504001). Since 2009, she is a post-doctoral researcher in the same group, applying the results of her PhD work in a project funded by EU FP7 PANACEA 222936 on the systems genetics of cancer, using *C. elegans* worms as a model system.

路漫漫其修远兮 吾将上下而求索

