Genetical genomics approaches for systems genetics

Bruno Tesson

The work described in this thesis was carried out at the Groningen Bioinformatics Centre, University of Groningen, The Netherlands. The author was financially supported by a BioRange grant SP1.2.3 from the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

Paranymphs: Yang Li & Tejas Gandhi

Cover generated using www.wordle.net
Printed by: Drukkerij Van Denderen, B.V. Groningen, The Netherlands

ISBN (gedrukte versie): 978-90-367-4912-1
ISBN (digitale versie): 978-90-367-4913-8

RIJKSUNIVERSITEIT GRONINGEN

Genetical genomics approaches for systems genetics

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
vrijdag 20 mei 2011
om 16.15 uur

door

Bruno Marie Tesson
geboren op 16 augustus 1983
te Bourges, Frankrijk

Promotores	Prof. dr. R.C. Jansen Prof. dr. R. Breitling
Beoordelingscommissie	Prof. dr. K. Schughart Prof. dr. C. Wijmenga Prof. dr. B.S. Yandell

ISBN: 978-90-367-4912-1

Contents

Summary

Genetical genomics is an interdisciplinary field concerned with the consequences of natural genetic variation on multiple molecular traits (mRNA expression levels, or protein and metabolite abundance). The goal of genetical genomics is to contribute to the establishment of informative molecular models explaining how DNA variation leads to observable phenotypic differences such as the emergence of a disease. Genetical genomics aims to become a genetic approach to systems biology and can therefore be referred to as a 'systems genetics' approach.

In this thesis, we introduce the principles of genetical genomics for a general readership (**Chapter 1**). The applicability of genetical genomics is illustrated with a screen of gene expression in hematopoietic cells from a population of inbred mice (**Chapter 2**). The results of this experiment demonstrate that the genetic variants controlling gene expression levels are highly sensitive to the differentiation state of the cells.

A computational protocol for mapping of genetic variants underlying variation in gene expression traits in inbred populations is fully developed in **Chapter 3**, addressing both theoretical and practical aspects of the implementation of the genetical genomics approach.

Such genetic variants are referred to as eQTL (expression quantitative trait loci). **Chapter 4** is concerned with a particularly controversial issue in genetical genomics: the relevance of eQTL hotspots on the genome. Those hotspot regions harbor genetic variation that seems to affect the expression of a (very) large number of genes (sometimes thousands). They could therefore reveal the presence of major biological regulators. However, because of limitations of the statistical methods commonly used, some studies have questioned the biological significance of hotspots. Here, we propose a permutation strategy that allows us to discard numerous hotspots due to statistical artifacts induced by widespread coexpression.

In **Chapter 5** we present DifCoEx a new bioinformatics method for differential coexpression analysis. We illustrate the use of DiffCoEx by applying it to the analysis of a publicly available microarray study of the effect of carcinogenic products on mutant tumor-prone rats. We show that the method is able to reveal meaningful groups of genes which do not show differential expression patterns, but are differentially correlated.

Chapter 6 and **Chapter 7** are concerned with statistical inference of causal relationships between phenotypes using genetic data. This topic has great significance for the of field biomedical research because it has been presented as a way to identify drug targets for the treatment of diseases and metabolic conditions with complex genetic inheritance. In **Chapter 6**, we review the different methods that have been used in genetics studies to try to connect phenotypes in functional networks. Subsequently, in **Chapter 7** we focus more specifically on a popular method

based on co-mapping of phenotypes and we delineate the critical conditions required for its proper use. In particular, we show that this method cannot produce reliable results within the settings of most current genetic experiments (including genetical genomics) because of the limited population sizes, the limited effect size of most genetic variants and the omnipresence of noise in high-throughput biological technologies.

The last part of this thesis is devoted to a discussion of when and how genetical genomics can successfully contribute to a systems genetics approach of biology (**Chapter 8**).

Chapter 1
Introduction

Genetical genomics is a computational biology strategy that applies concepts of quantitative genetics to the analysis of high-throughput data from modern molecular profiling technologies such as microarrays, mass spectrometers or next generation sequencers. The principle of genetical genomics is to exploit the genome-wide genetic perturbation arising from natural variation in a population or induced by experimental crosses to study the phenotypic response at all intermediate molecular levels such as in mRNA expression, or protein and metabolite abundance. Using this strategy, one is able to perturb (and expose) virtually any molecular pathway while keeping the organism under study in a functioning natural state (as opposed to the more radical disruptions induced by gene knock-out or knock-down experiments for example). This property places genetical genomics at the forefront of systems genetics: systems genetics aims at constructing a holistic view of biological processes by integrating data from multiple molecular levels and from different tissues into explanatory models. In this chapter, we introduce the basic principles of genetics and how they are applied in genetical genomics. In the end, we outline the contents of this thesis.

1.1 Introduction to classical genetics principles and QTL mapping

Genetics is the science concerned with the mechanisms that underlie heredity. The origin of genetics is often traced back to the middle of the 19th century and Abbott Gregor Mendel's careful observations on the transmission of certain traits in pea plants from parents to offspring, from which he derived the fundamental principles nowadays known as Mendel's Laws [1]. In one of his experiments, Mendel for example crossed two self-pollinating plants: one with yellow seeds and one with green seeds. He then observed that all the first generation offspring plants had yellow seeds, while when those offspring plants were self-pollinated, one out of four plants had green seeds. Mendel deduced that the seed color trait was carried by discrete units that are transmitted from parents to offspring unchanged, and that every offspring individual received one such unit from each parent. He also concluded that the unit responsible for the yellow color was dominant in that, when a plant inherited the units for both colors, the seeds were yellow. He additionally observed that the traits he studied were inherited independently and therefore concluded that the units behind different traits are passed independently. Those observations are known as Mendel's Laws and describe patterns of inheritance for some traits controlled by single loci (known as Mendelian traits). Mendel's observations would later be biologically explained by the biological process of meiosis, in which gametes receive one copy of each chromosome, and the units Mendel described became known as genes. Different versions of the genes coding for different values of the corresponding trait later became known as alleles.

At first, Mendel's visionary work did not receive the attention it deserved, and it was only at the dawn of the 20th century that Hugo de Vries, William Bateson and others pursued his profound insights further [2-4]. The theoretical and statistical basis of quantitative genetics was then laid down: the focus of genetics was expanded from the study of traits with discrete observable properties (for example seed color) to the study of quantitative traits with continuous measurable values (for example plant height) [5].

When Thomas Hunt Morgan discovered that some of the traits (eye color and wing size) that he observed on his mutant *Drosophila melanogaster* flies tended to be inherited jointly, he proposed the concept of linkage and hypothesized that this joint inheritance was related to the proximity of genes on chromosomes [6]. These fundamental insights opened the way to the establishment of genetic maps that positioned the genes coding for studied traits onto linear chromosomes. The first genetic map was proposed by Sturtevant, one of Morgan's students [7]. Until the middle of the 20th century, the molecular nature of chromosomes (and genes) was unknown and it was therefore impossible to observe the actual differences that encoded different phenotypes. For that reason, geneticists had to rely on indirect manifestations of genetic information in the form of morphological markers: new

traits were mapped relatively to other previously studied traits that were easy to observe and used as markers.

After DNA was revealed as the molecular incarnation of genes [8], and after Watson and Crick resolved its molecular structure [9], the discovery of the first restriction enzyme in 1968 [10] opened the way to the first genotyping technologies. In those pioneering technologies, genetic differences in the lengths of different sequence repeats at specific DNA markers were visualized as bands in electrophoresis gels. As those technologies gradually matured, molecular markers replaced morphological markers and denser genetic maps became available. Nowadays many different cost-efficient genotyping solutions (including sequencing and Single Nucleotide Polymorphisms arrays) have opened the way to systematic genome-wide fine mapping of quantitative traits (Quantitative Trait Locus or QTL mapping).

The process of QTL mapping (**Figure 1**) consists in searching for genome regions that influence the value of a given trait. For example, identifying a QTL for plant height means finding a DNA region at which the plants that carry a certain allele tend to be significantly higher or lower than those carrying another allele. This can be done by simply comparing for each available genotyped marker along the genome, the distribution of trait values associated with different alleles.

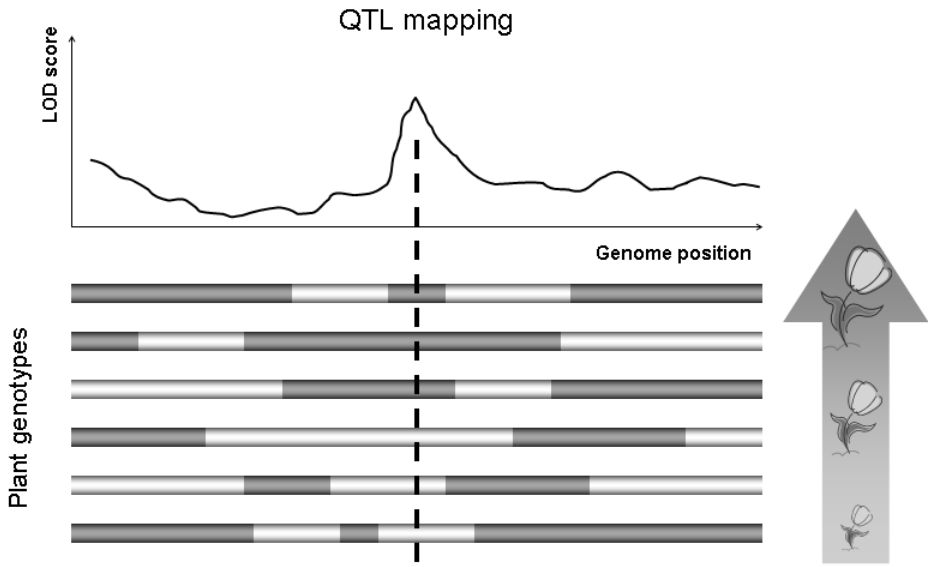


Figure 1 - Explanatory schema of QTL mapping of plant size. In the lower part of the figure, schematic representations of individual plant genotypes are shown. We assume the mapping is done in recombinant inbred lines which are therefore homozygous mosaic of two possible parental genotypes (represented in dark grey and in white). The genotypes have been sorted from bottom to top by increasing size of the corresponding plant. QTL mapping scans the genome for a location at which the genotype distribution explains the difference in plant sizes. In this example, the dotted line is one such position as all plants with white genotypes are smaller than those with grey genotypes. This is reflected in the schematic QTL profile plot on top of the figure by a significant peak.

ANOVA (Analysis of Variance) is a suitable statistical framework for such an analysis but suffers from missing data in sparse genetic maps. In 1989, Lander and Botstein proposed Interval Mapping [11]: a novel QTL mapping methodology offering solutions to these two shortcomings and pinpointing more precisely the actual QTL positions.

While Mendel's laws can explain inheritance patterns for traits controlled by single loci, most phenotypes including diseases tend to be controlled by multiple genes and fall into the category of "complex traits". Such complex traits necessitate more complicated models that include more QTL as co-factors. Multiple QTL Mapping and Composite Interval Mapping were therefore proposed to uncover the complex genetic makeup behind complex phenotypes [12, 13].

QTL mapping experiments are often divided in two major classes: linkage mapping studies and association mapping studies. Linkage mapping studies track the joint inheritance of certain phenotypes and chromosomal regions amongst related individuals (families or experimental crosses such as backcrosses, F2 or Recombinant Inbred Lines) to infer close physical proximity (linkage) between those phenotypes and those regions. Association mapping on the other hand, studies the correlation between marker genotypes and phenotypes in a population of mostly unrelated individuals.

1.2 Molecular phenotyping

New high throughput profiling technologies have revolutionized modern biology and at the same time changed the way genetics is performed. They have greatly expanded the collection of phenotypes that can be studied; adding to the classical and directly observable classical phenotypes such as weight or color, the abundances and states of a very wide variety of bio-molecules (for example, mRNA transcripts, proteins and metabolites). These new molecular traits can inform us on the inner mechanisms that underlie biological processes. A list of molecular profiling technologies that can be used to study the genetics of molecular traits is given in **Table 1**.

In this thesis, the data that are analyzed are primarily from gene expression microarrays, therefore in this introduction we will present in more details this technology. Microarrays are chips of glass or silicon on to which short DNA sequences known as probes are bound at spots that are spread along the surface. These probes are designed so that they are complementary to specific gene sequences. During a microarray experiment, RNA is extracted, and then complementary DNA (cDNA) is produced and amplified through the Polymerase Chain Reaction [14]. Following these steps, the amplified cDNA is applied onto the chip where it can hybridize to the complementary probes. The amount of cDNA that has bound to a specific probe is then read using a scanner which quantifies the intensity of a fluorescent tag that has been incorporated to the cDNA. The scan of a microarray

provides a simultaneous measurement of thousands of RNA signals, for example mRNAs which can be used to infer the activity of genes.

Molecular level	Technologies	References
Genome	SNP microarray DNA-Seq	[15] [16]
Epigenome	ChIP-on-chip ChIP-Seq	[17] [18]
Transcriptome	Microarray RNA-Seq	[19] [20]
Proteome	2D gel electrophoresis Mass Spectrometry Antibody-based protein chip	[21] [22] [23]
Metabolome	Mass spectrometry NMR	[24] [25]

Table 1 – Technologies for profiling of different molecular levels.

Similar technologies have been added to the collection of tools available to biologists. Beadarrays, commercialized by Illumina, work in a similar fashion as microarrays, except the probes are attached onto beads rather than on a chip. More recently, Next Generation Sequencing (NGS) technologies allow to sequence and count a growing fraction of all of the RNA or DNA sequences present in a sample.
Microarrays are not restricted to mRNA measurements. In particular, tiling arrays have been developed to survey the entire genome and discover any transcribed region, including those with non-coding RNAs. SNP arrays on the other hand are used to assay DNA, and identify sequence variants, which has made genotyping fast and cost-efficient. Other application of microarrays include Comparative Genome Hybridization which is used to identify copy number variation and ChIP-on-chip (Chromatin Immuno-precipitation on chip) that allows one to identify the binding sites of given proteins.

In combination, all those technologies allow biologists to gain insights into the molecular networks that drive physiological processes with an unprecedented level of details.

1.3 From the mapping of molecular traits to Systems Genetics

Nowadays genome-wide association studies or linkage studies allow pinpointing with relative precision the location of genes which play even relatively minor role in the development of complex phenotypes such as many human diseases. In addition to identifying the genes in which the exact mutations are located that are responsible for a specific phenotype, the challenge is to identify the processes through which

variation in these genes leads to disease. Answering this question requires delving deeper into the biology and therefore studying more intrinsic traits such as cholesterol levels, or the abundance of proteins within relevant organs, the level of expression of genes, or the activity of metabolic reactions. As we have discussed in the previous section, the technologies allowing an exploration of such molecular traits have matured in recent years, making the prospect of simultaneously mapping virtually all traits from a molecular level realistic [26]. A new global strategy was therefore envisioned for the study of complex traits: systems genetics [27], as it became known, would allow researchers to track the biological flow of information from the original DNA mutation to the observable phenotypic variation, by exposing the molecular mechanisms involved such as transcriptional networks and metabolic pathways (**Figure 2**).

The first studies in model organisms focused on large-scale mapping of traits from a single molecular level. These studies revealed that molecular traits such as gene expression levels, protein and metabolite abundances are highly heritable and therefore confirmed the relevance of applying genetics approaches to their study [28]. Furthermore, for gene expression traits, a large part of the underlying genetic variation could be tracked back to the chromosomal proximity of the genes themselves, forming what became known as local or *cis*-eQTLs, in contrast with distant or *trans*-eQTLs. Another striking finding has been the revelation of the existence of genome regions to which variation in large number of traits can be mapped [29]; such regions have been designated as "QTL hotspots". This genetic information was then used to try to infer biological relationships between those traits and to connect them into networks [30] (for example transcriptional networks). In more recent studies, efforts have been devoted to the integration of phenotypes from different levels, jointly studying gene expression, proteome, metabolome and sometimes classical traits such as diseases [31, 32]. Moreover a complete understanding of physiological processes requires studying different molecular levels across different types of organs, tissues, cell-types, developmental time points, and perhaps even in different populations. Because variation in traits is the result of a complex interplay between genetic and environmental factors, much can be learned by extending genetical genomics experiments with the addition of environmental perturbation to the natural genetic variation [33]. Combining all those dimensions into single explanatory biological models is the aim of systems genetics.

Because systems genetics is changing the scale of biological experiments, it is accompanied by a set of new challenges. The first of these challenges is computational: the simultaneous mapping of tens of thousands of traits and the integration of multiple data types requires adapted hardware and software infrastructures. Next there are methodological challenges: systems genetics calls for new approaches to extract meaningful information from the ever-increasing amounts of data produced. Properly controlling the multiple testing problems, dealing with the systematic noise and artifacts that come with high-throughput data, combining evidence from differ-

ent data and from the scientific literature, connecting molecular traits into relevant biological networks can only be achieved within sound statistical frameworks. This thesis is devoted to those methodological issues.

1.4 Outline of thesis contribution

In addition to the present introductory chapter, this thesis comprises seven chapters (see also **Figure 2**).

Chapter 2 presents a genetical genomics analysis of hematopoietic differentiation in a mouse cross. From a recombinant inbred line panel of about 24 strains, samples from four hematopoietic cell-types were collected and expression profiled using Illumina bead arrays. The chapter presents a report of the eQTLs (expression Quantitative Trait Loci) that were identified and argues that those are highly sensitive to the cellular differentiation state. This finding highlights the importance of targeting relevant tissues and cell-types in systems genetics experiments.

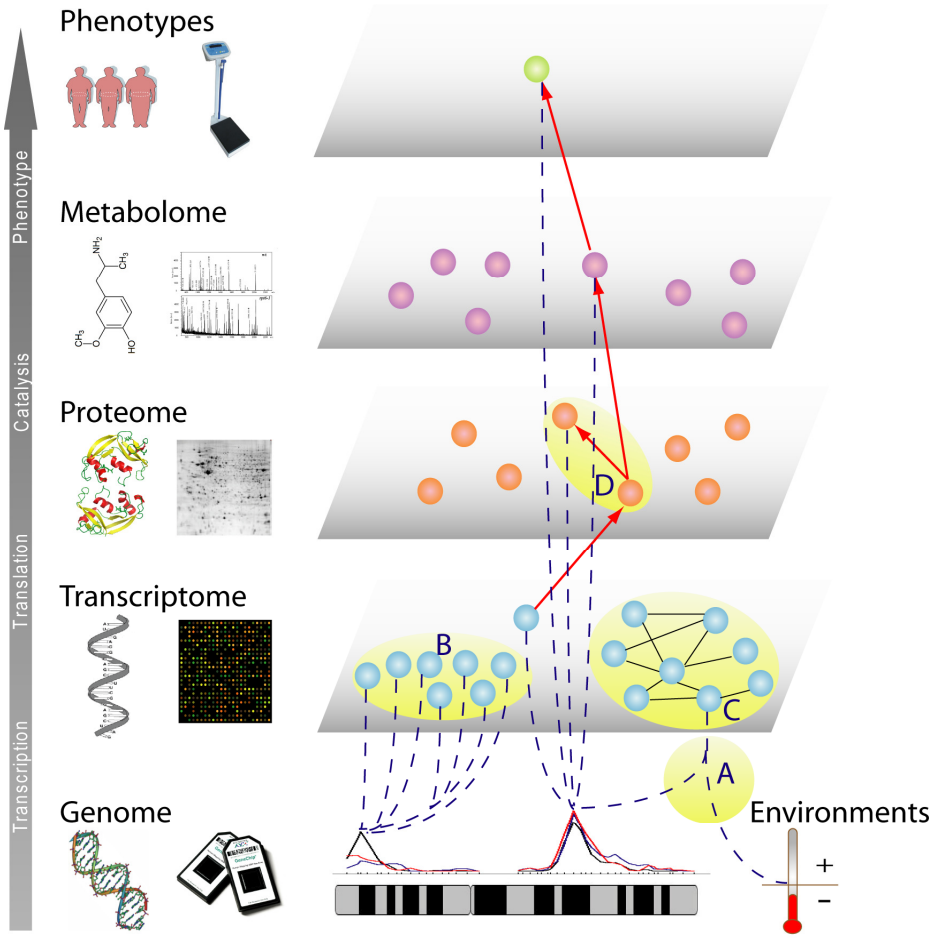
In **Chapter 3**, all the steps involved in an eQTL mapping experiment are detailed in a computational protocol that includes R scripts. The focus is primarily on linkage analysis in mouse or rat experimental crosses such as recombinant inbred lines. A number of technical issues (and some solutions to them) are discussed.

Chapter 4 addresses the puzzling discrepancy in the reporting of eQTL hotspots in the literature and argues that many hotspots are actually caused by confounding correlation. An appropriate permutation procedure that allows to discard many spurious eQTL hotspots is advocated.

Chapter 5 introduces and showcases a new method for Differential Coexpression analysis called DiffCoEx. DiffCoEx is a simple and sensitive method that can identify groups of genes that are differentially correlated between different conditions. Such a method may potentially identify molecular pathways that are active specifically in one condition. The method is applied to a published cancer-related rat dataset, and it is shown that the differential coexpression analysis identifies genes that would not otherwise have been picked up by classical differential expression methods.

The sixth and seventh chapters address emerging methods in causal inference with genetic data which are shifting the paradigm of network inferences by providing statistical evidence to support directed links between genes, proteins, metabolites or diseases. In **Chapter 6**, different approaches using genetic data for gene network inference that have been proposed are reviewed. **Chapter 7** examines the statistical potential of such methods under different realistic settings: varying population sizes and in the presence or absence of hidden factor variation and suggests ways to overcome some of the limitations.

Finally, **Chapter 8** discusses current issues that will benefit from future research in genetical genomics.



A: eQTL mapping and eQTL by environment/tissue interactions Chapter 2-3
B: eQTL hotspots Chapter 4
C: (Differential) coexpression networks Chapter 5
D: QTL based causal inference between traits Chapter 6-7

Figure 2 - Systems genetics: an integrative strategy.

1.5 References

1. Mendel G: **Versuche über Pflanzen-Hybriden.** In: *Verhand- lungen des naturforschenden Vereines in Brünn, 1865*, IV:3–47.
2. de Vries H: **Intracellular pangenesis: including a paper on fertilization and hybridization**: The Open Court Publishing Co.; 1910.
3. Bateson W: **The progress of genetic research**. In: *Third Conference on Hybridization and Plant Breeding: 1906*; 1906: 90-97.
4. Bateson W: **Mendel's principles of heredity a defence, with a translation of Mendel's original papers on hybridisation**. Cambridge [u.a.]: Cambridge Univ. Press; 2009.
5. Fisher RA: **The correlation between relatives on the supposition of Mendelian inheritance**. *Transactions of the Royal Society of Edinburgh* 1918, **52**:399-433.
6. Morgan TH: **The mechanism of Mendelian heredity**: Holt; 1915.
7. Sturtevant AH: **The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association**. *Journal of Experimental Zoology* 1913, **14**(1):43-59.
8. Avery OT, Macleod CM, McCarty M: **Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii**. *Journal of Experimental Medicine* 1944, **79**(2):137-158.
9. Watson JD, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid**. *Nature* 1953, **171**(4356):737-738.
10. Meselson M, Yuan R: **DNA restriction enzyme from E. coli**. *Nature* 1968, **217**(5134):1110-1114.
11. Lander ES, Botstein D: **Mapping mendelian factors underlying quantitative traits using RFLP linkage maps**. *Genetics* 1989, **121**(1):185-199.
12. Jansen RC: **Controlling the type I and type II errors in mapping quantitative trait loci**. *Genetics* 1994, **138**(3):871-881.
13. Zeng ZB: **Precision mapping of quantitative trait loci**. *Genetics* 1994, **136**(4):1457-1457.
14. Mullis KB, Faloona FA: **Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction**. *Methods in Enzymology* 1987, **155**:335-350.
15. Hacia JG, Fan JB, Ryder O, Jin L, Edgemon K, Ghandour G, Mayer RA, Sun B, Hsie L, Robbins CM *et al*: **Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays**. *Nature Genetics* 1999, **22**(2):164-167.

16. Schuster SC: **Next-generation sequencing transforms today's biology**. *Nature Methods* 2008, **5**(1):16-18.
17. Ballestar E, Paz MF, Valle L, Wei S, Fraga MF, Espada J, Cigudosa JC, Huang TH, Esteller M: **Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer**. *The EMBO Journal* 2003, **22**(23):6335-6345.
18. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A *et al*: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing**. *Nature Methods* 2007, **4**(8):651-657.
19. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW: **Yeast microarrays for genome wide parallel genetic and gene expression analysis**. *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94**(24):13057-13062.
20. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M: **Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing**. *BioTechniques* 2008, **45**(1):81-94.
21. O'Farrell PH: **High resolution two-dimensional electrophoresis of proteins**. *The Journal of Biological Chemistry* 1975, **250**(10):4007-4021.
22. Washburn MP, Wolters D, Yates JR, 3rd: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology**. *Nature Biotechnology* 2001, **19**(3):242-247.
23. MacBeath G, Schreiber SL: **Printing proteins as microarrays for high-throughput function determination**. *Science (New York, NY* 2000, **289**(5485):1760-1763.
24. Fiehn O: **Metabolomics--the link between genotypes and phenotypes**. *Plant Molecular Biology* 2002, **48**(1-2):155-171.
25. Nicholson JK, Lindon JC, Holmes E: **'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data**. *Xenobiotica; the fate of foreign compounds in biological systems* 1999, **29**(11):1181-1189.
26. Jansen RC, Nap JP: **Genetical genomics: the added value from segregation**. *Trends in Genetics* 2001, **17**(7):388-391.
27. Threadgill DW: **Meeting report for the 4th annual Complex Trait Consortium meeting: from QTLs to systems genetics**. *Mammalian Genome* 2006, **17**(1):2-4.
28. Brem RB, Yvert G, Clinton R, Kruglyak L: **Genetic dissection of transcriptional regulation in budding yeast**. *Science (New York, NY* 2002, **296**(5568):752-755.

29. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L: ***Trans*-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors**. *Nature Genetics* 2003, **35**(1):57-64.
30. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF *et al*: **Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'**. *Nature Genetics* 2005, **37**(3):225-232.
31. Fu J, Keurentjes JJ, Bouwmeester H, America T, Verstappen FW, Ward JL, Beale MH, de Vos RC, Dijkstra M, Scheltema RA *et al*: **System-wide molecular evidence for phenotypic buffering in Arabidopsis**. *Nature Genetics* 2009, **41**(2):166-167.
32. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S *et al*: **Genetics of gene expression and its effect on disease**. *Nature* 2008, **452**(7186):423-428.
33. Li Y, Breitling R, Jansen RC: **Generalizing genetical genomics: getting added value from environmental perturbation**. *Trends Genet* 2008, **24**(10):518-524.

Chapter 2
Expression quantitative trait loci are highly sensitive to cellular differentiation state

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 study 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.

Originally published as:
Expression quantitative trait loci are highly sensitive to cellular differentiation state.
Gerrits A*, Li Y*, **Tesson BM***, Bystrykh LV, Weersing E, Ausema A, Dontje B, Wang X, Breitling R, Jansen RC, de Haan G.
PLoS Genetics 2009 Oct;5(10):e1000692.
*equal contributions

2.1 Introduction

Genetical genomics uses quantitative genetics on a panel of densely genotyped individuals to map genomic loci that modulate gene expression [1]. The quantitative trait loci identified in this manner are referred to as expression quantitative trait loci, or eQTLs [2]. Most genetical genomics studies that have thus far been reported have analyzed single cell types or compared developmentally unrelated and distant cell types [3-8]. 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 [9-12]. 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.

2.2 Results

2.2.1 Genetic regulation of gene expression

We evaluated genome-wide RNA transcript expression levels in purified $Lin^-Sca\text{-}1^+c\text{-}Kit^+$ multi-lineage cells, committed $Lin^-Sca\text{-}1^-c\text{-}Kit^+$ progenitor cells, erythroid $TER\text{-}119^+$ cells, and myeloid $Gr\text{-}1^+$ 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 [13]. 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 $Lin^-Sca\text{-}1^+c\text{-}Kit^+$ 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 $Lin^-Sca\text{-}1^+c\text{-}Kit^+$ population, for

simplicity we will refer to this population as stem cells. The Lin$^-$Sca-1$^-$c-Kit$^+$ cell population does not contain stem cells and lymphoid precursors, but does include common progenitors of the myeloid and erythroid lineages [14]. Finally, TER-119$^+$ cells and Gr-1$^+$ 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 1** and **Table S1**).

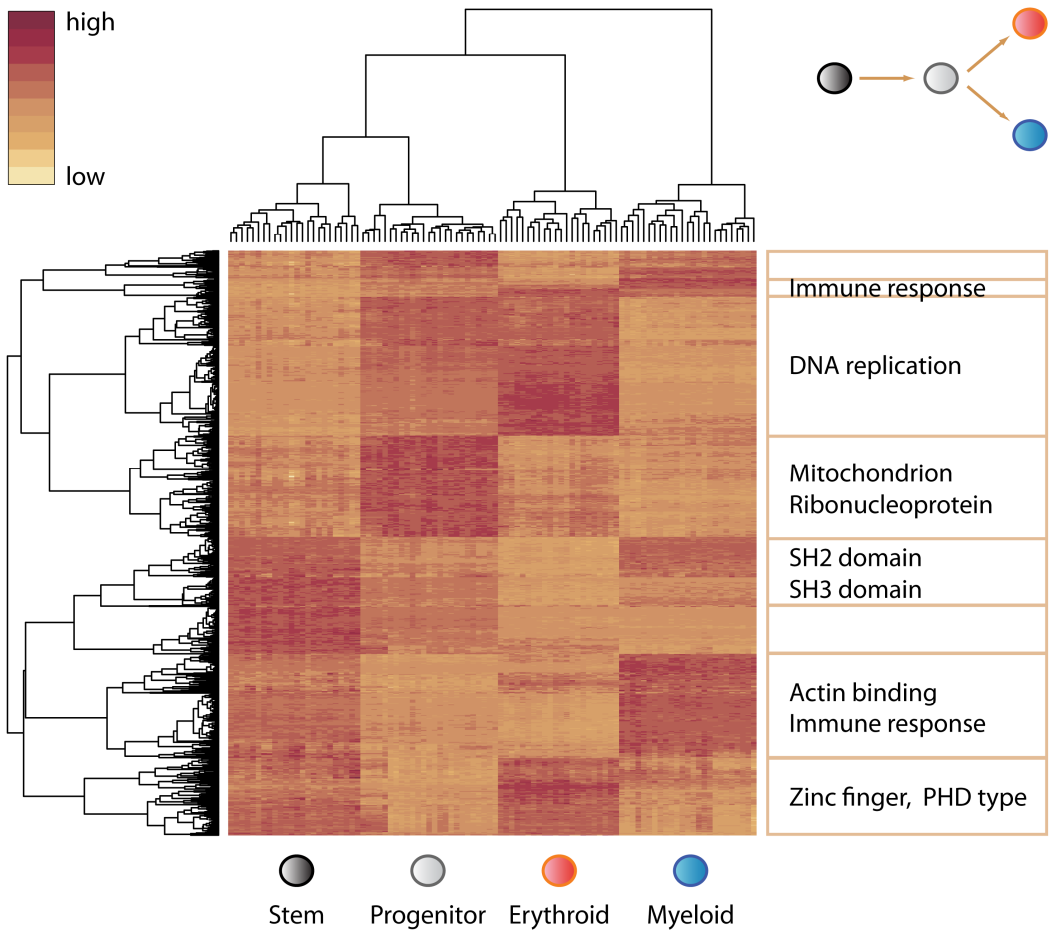


Figure 1 - Mean expression levels for all probes in the four cell types. Unsupervised clustering including all probes for the 96 RNA samples follows cell-type (top hierarchical tree), while clustering of the 876 most varying probes reveals distinct categories of genes that show cell-type-specific expression (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*) [15]. 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.

2.2.2 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 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 [16]. 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 $p < 10^{-6}$ (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 1**).

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 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**.

2.2.3 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 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 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 2B**, left panel). *Slit2* encodes an excreted chemorepellent molecule that is known to be expressed in embryonic stem cells [17], to be involved in neurogenesis [18] and angiogenesis [19], and to inhibit leukocyte chemotaxis [20]. We found a total of 140 genes that have eQTLs that are preferentially/selectively active in stem cells (**Figure 2B**, right panel, largest symbols, **Table 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 1**). Very distinct patterns of cell-type-specific gene regulation emerged when these eQTLs were visualized in genome-wide dot plots (**Figures 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.

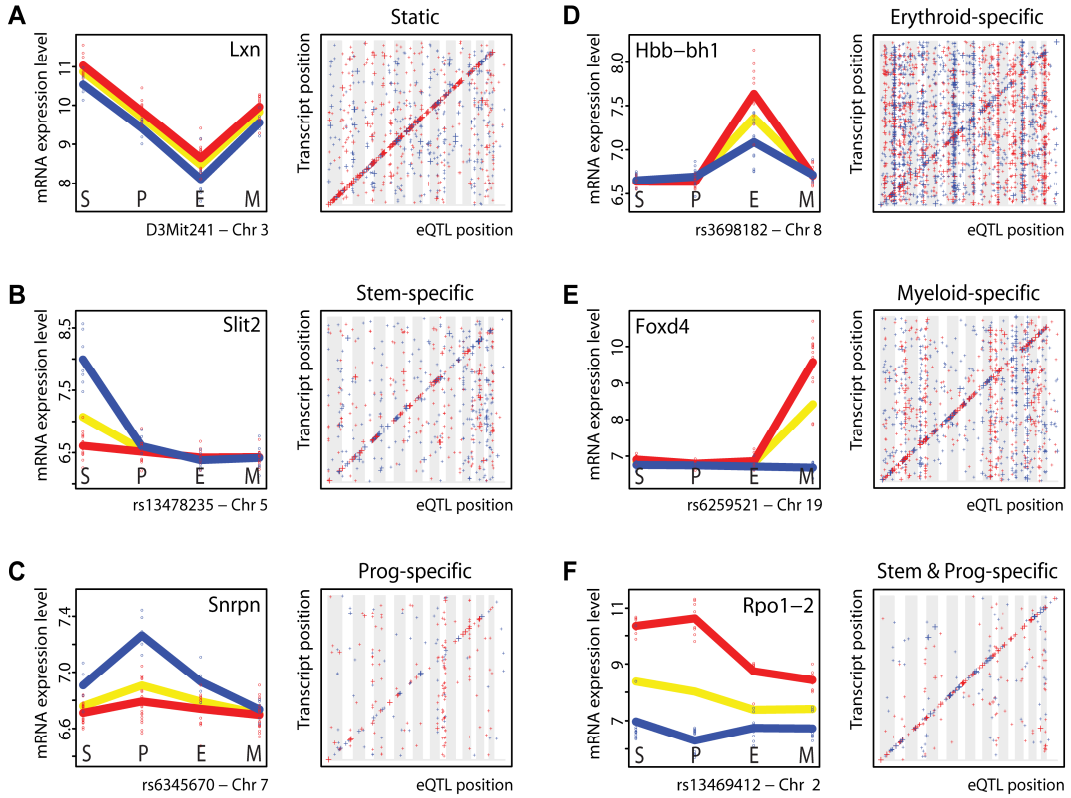


Figure 2 - Identification of *static* and *dynamic* eQTLs. (A) Genome-wide identification of cell-type-independent *static* eQTLs. (Left panel) *Lxn* mRNA levels were analyzed in all 4 cell types. Each circle represents an individual sample (strain). The yellow line shows mean expression levels across 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**.

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) [21]. 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 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**.

	eQTL subcategory		***# probes***	***# markers***	***# probes / # marker***
Static		*Local*	268	161	1.66
		Distant	97	76	1.28
		Total	365	213	1.71
Dynamic	All	*Local*	642	282	2.28
		Distant	641	276	2.32
		Total	1283	445	2.88
	Stem cells	*Local*	87	66	1.32
		Distant	53	42	1.26
		Total	140	105	1.33
	Progenitor	*Local*	32	27	1.19
		Distant	13	12	1.08
		Total	45	39	1.15
	Erythroid	*Local*	131	90	1.46
		Distant	400	164	2.44
		Total	531	223	2.38
	Myeloid	*Local*	163	121	1.35
		Distant	132	72	1.83
		Total	295	179	1.65

Table 1 - Number of probes with eQTLs ($p < 10^{-6}$) and the number of associated markers.

2.2.4 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 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 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 [22]. 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 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.

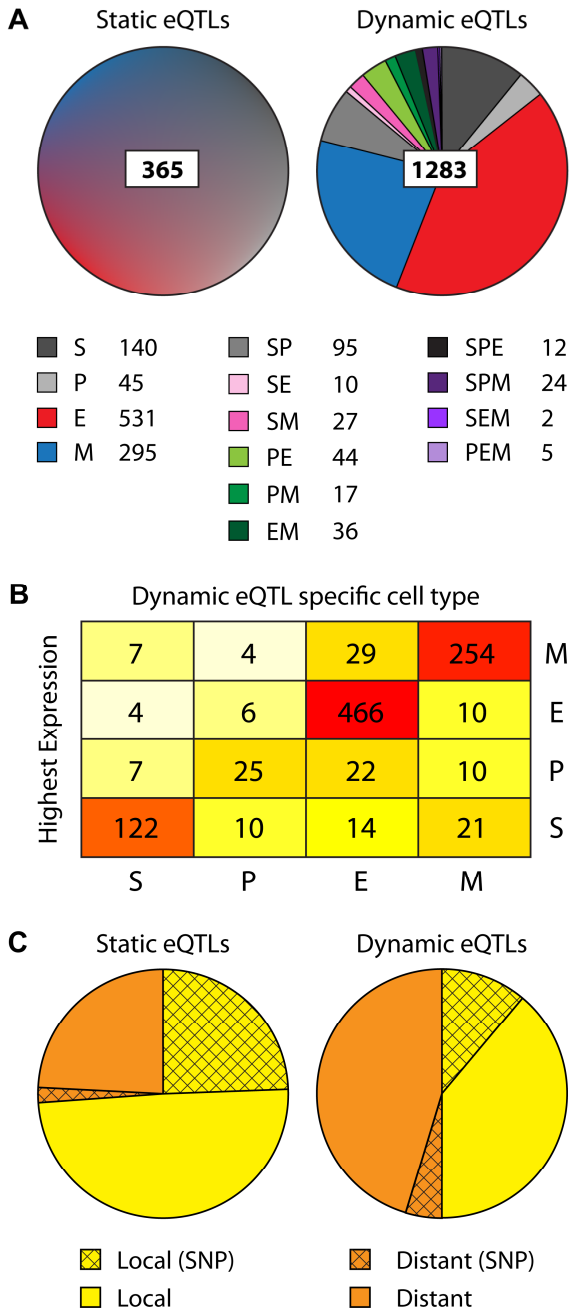


Figure 3 - Quantitative overview of *static* and *dynamic* eQTLs. (A) Pie charts presenting all 365 *static* and 1283 *dynamic* eQTLs that were detected with $p < 10^{-6}$. *Dynamic* eQTLs are subdivided in all 14 categories of interaction eQTLs. **(B)** Matrix showing the four cell-type-dependent *dynamic* 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.

2.3 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 [23-26] 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 [27]. 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 [4, 5, 28-31]. 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 paper were deposited in the online database *GeneNetwork (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 [32, 33]. 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 [34], and as such can serve as a starting point for hypothesis-driven research in the fields of stem cell biology and hematology.

2.4 Methods

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

2.4.2 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.

2.4.3 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 TotalPrep RNA Amplification Kit (Ambion) according to the manufacturer's specifications starting with 100 ng total RNA. Per sample, 1.5 µg 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 log2 transformed and then quantile normalized as a single group.

2.4.4 Clustering of genes

For cluster analysis we retained only genes having a minimal fold change of 2 (difference of 1 in log2 scale) in either direction in mean expression on the transition from $Lin^-Sca\text{-}1^+c\text{-}Kit^+$ to $Lin^-Sca\text{-}1^-c\text{-}Kit^+$ and on the transition from $Lin^-Sca\text{-}1^-c\text{-}Kit^+$ to $TER\text{-}119^+$ or to $Gr\text{-}1^+$. 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 [35].

2.4.5 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 ith microarray; μ 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 ith microarray ($i = 1,\ldots n$) and jth cell type; CT_j is the jth cell type effect; $(Q\times CT)_{ij}$ is the interaction effect between the ith eQTL genotype and jth 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.

2.4.6 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*.

2.4.7 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 includes 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.

2.4.8 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 [36] 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.

2.4.9 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 [37]. Then we permuted the strain labels and applied the ANOVA model $y = Q + CT + Q\times CT + e$ 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.

2.4.10 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 ith microarray ($i = 1,\ldots n$) and jth cell type; CT_j is the jth cell type effect; Q^* and $(Q^* \times CT)_{ij}$ 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 ith genotype and jth cell type, $Q_i^* Q_i$ is the epistasis effect and e_{ij} is the residual error.

URLs

All raw data were deposited at GEO (http://www.ncbi.nlm.nih.gov/geo/). All processed data presented in this paper were deposited at *GeneNetwork* (www.genenetwork.org) [32, 33]. Additional files are available at: http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000692

2.5 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.

2.6 References

1. Jansen RC, Nap JP: **Genetical genomics: the added value from segregation**. *Trends Genet* 2001, **17**(7):388-391.
2. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G *et al*: **Genetics of gene expression surveyed in maize, mouse and man**. *Nature* 2003, **422**(6929):297-302.
3. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF *et al*: **Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'**. *NatGenet)*2005, **37**(3):225-232.
4. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA *et al*: **Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function**. *NatGenet* 2005, **37**(3):233-242.
5. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V *et al*: **Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease**. *NatGenet* 2005, **37**(3):243-253.
6. Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, Fischer J, Maatz H, Kren V, Pravenec M *et al*: **Heritability and tissue specificity of expression quantitative trait loci**. *PLoSGenet* 2006, **2**(10):e172.
7. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE: **Genetic inheritance of gene expression in human cell lines**. *AmJHumGenet* 2004, **75**(6):1094-1105.
8. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG: **Genetic analysis of genome-wide variation in human gene expression**. *Nature* 2004, **430**(7001):743-747.
9. Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA, Lemischka IR: **A stem cell molecular signature**. *Science* 2002, **298**(5593):601-604.
10. Chambers SM, Boles NC, Lin KY, Tierney MP, Bowman TV, Bradfute SB, Chen AJ, Merchant AA, Sirin O, Weksberg DC *et al*: **Hematopoietic Fingerprints: An Expression Database of Stem Cells and Their Progeny**. *Cell Stem Cell* 2007, **1**(5):578-591.
11. Kiel MJ, Yilmaz OH, Iwashita T, Yilmaz OH, Terhorst C, Morrison SJ: **SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells**. *Cell* 2005, **121**(7):1109-1121.
12. Forsberg EC, Prohaska SS, Katzman S, Heffner GC, Stuart JM, Weissman IL: **Differential expression of novel potential regulators in hematopoietic stem cells**. *PLoSGenet* 2005, **1**(3):e28.

13. Peirce JL, Lu L, Gu J, Silver LM, Williams RW: **A new set of BXD recombinant inbred lines from advanced intercross populations in mice**. *BMCGenet* 2004, **5**:7.
14. Bryder D, Rossi DJ, Weissman IL: **Hematopoietic stem cells: the paradigmatic tissue-specific stem cell**. *AmJPathol* 2006, **169**(2):338-346.
15. Rockman MV, Kruglyak L: **Genetics of global gene expression**. *NatRevGenet* 2006, **7**(11):862-872.
16. Liang Y, Jansen M, Aronow B, Geiger H, Van Zant G: **The quantitative trait gene latexin influences the size of the hematopoietic stem cell population in mice**. *NatGenet* 2007, **39**(2):178-188.
17. Katoh Y, Katoh M: **Comparative genomics on SLIT1, SLIT2, and SLIT3 orthologs**. *OncolRep* 2005, **14**(5):1351-1355.
18. Wang KH, Brose K, Arnott D, Kidd T, Goodman CS, Henzel W, Tessier-Lavigne M: **Biochemical purification of a mammalian slit protein as a positive regulator of sensory axon elongation and branching**. *Cell* 1999, **96**(6):771-784.
19. Wang B, Xiao Y, Ding BB, Zhang N, Yuan X, Gui L, Qian KX, Duan S, Chen Z, Rao Y *et al*: **Induction of tumor angiogenesis by Slit-Robo signaling and inhibition of cancer growth by blocking Robo activity**. *Cancer Cell* 2003, **4**(1):19-29.
20. Wu JY, Feng L, Park HT, Havlioglu N, Wen L, Tang H, Bacon KB, Jiang Z, Zhang X, Rao Y: **The neuronal repellent Slit inhibits leukocyte chemotaxis induced by chemotactic factors**. *Nature* 2001, **410**(6831):948-952.
21. Breitling R, Li Y, Tesson BM, Fu J, Wu C, Wiltshire T, Gerrits A, Bystrykh LV, De Haan G, Su AI *et al*: **Genetical genomics: spotlight on QTL hotspots**. *PLoSGenet* 2008, **4**(10):e1000232.
22. Alberts R, Terpstra P, Li Y, Breitling R, Nap JP, Jansen RC: **Sequence polymorphisms cause many false *cis* eQTLs**. *PLoSONE* 2007, **2**(7):e622.
23. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C *et al*: **An integrative genomics approach to infer causal associations between gene expression and disease**. *NatGenet* 2005, **37**(7):710-717.
24. Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG *et al*: **Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes**. *NatGenet* 2007, **39**(10):1208-1216.
25. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S *et al*: **Genetics of gene expression and its effect on disease**. *Nature* 2008, **452**(7186):423-428.
26. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK *et al*: **Variations in DNA elucidate molecular networks that cause disease**. *Nature* 2008, **452**(7186):429-435.

27. Li Y, Breitling R, Jansen RC: **Generalizing genetical genomics: getting added value from environmental perturbation**. *Trends Genet* 2008, **24**(10):518-524.
28. Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JA, Hazendonk E, Prins P, Plasterk RH, Jansen RC *et al*: **Mapping determinants of gene expression plasticity by genetical genomics in C. elegans**. *PLoSGenet* 2006, **2**(12):e222.
29. West MA, Van Leeuwen H, Kozik A, Kliebenstein DJ, Doerge RW, St Clair DA, Michelmore RW: **High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis**. *Genome Res* 2006, **16**(6):787-795.
30. Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, Van den Ackerveken G, Snoek LB, Peeters AJ, Vreugdenhil D, Koornneef M, Jansen RC: **Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci**. *ProcNatlAcadSciUSA* 2007, **104**(5):1708-1713.
31. Whiteley AR, Derome N, Rogers SM, St-Cyr J, Laroche J, Labbe A, Nolte A, Renaut S, Jeukens J, Bernatchez L: **The phenomics and expression quantitative trait locus mapping of brain transcriptomes regulating adaptive divergence in lake whitefish species pairs (Coregonus sp.)**. *Genetics* 2008, **180**(1):147-164.
32. Chesler EJ, Lu L, Wang J, Williams RW, Manly KF: **WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior**. *NatNeurosci* 2004, **7**(5):485-486.
33. Wang J, Williams RW, Manly KF: **WebQTL: web-based complex trait analysis**. *Neuroinformatics* 2003, **1**(4):299-308.
34. Gerrits A, Dykstra B, Otten M, Bystrykh L, De Haan G: **Combining transcriptional profiling and genetic linkage analysis to uncover gene networks operating in hematopoietic stem cells and their progeny**. *Immunogenetics* 2008, **60**(8):411-422.
35. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):3.
36. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *ProcNatlAcadSciUSA* 2003, **100**(16):9440-9445.
37. Anderson M, Braak CT: **Permutation tests for multi-factorial analysis of variance**. *Journal of Statistical Computation and Simulation* 2003, **73**(2):85-113.

Chapter 3
eQTL analysis in mice and rats

Since the introduction of Genetical Genomics in 2001, many studies have been published on various organisms, including mouse and rat. Genetical genomics makes use of the latest microarray profiling technologies and combines vast amounts of genotype and gene expression information, a strategy that has proven very successful in inbred line crosses. The data are analyzed using standard tools for linkage analysis to map the genetic determinants of gene expression variation. Typically, studies have singled out hundreds of genomic loci regulating the expression of nearby and distant genes (called local and distant expression quantitative trait loci, respectively; eQTLs). In this chapter, we provide a step-by-step guide to performing genome-wide linkage analysis in an eQTL mapping experiment by using the R statistical software framework.

Originally published as:
eQTL analysis in mice and rats.
Tesson BM, Jansen RC
Methods in Molecular Biology 2009;573:285-309.

3.1 Introduction

A genetical genomics [1] study involves the perturbation of thousands of genes at the same time through genetic mechanisms of recombination and segregation to create genome-wide "mosaics" of naturally occurring gene variants. Genetical genomics experiments then correlate gene expression variation with DNA variation for tens of thousands of genes, performing tens of thousands times an analysis similar to traditional QTL analysis of a classical phenotypic trait. The analysis of variance (ANOVA) methods offer a framework well suited for such QTL analyses.

Over the past few years, a large number of mouse recombinant inbred populations (RILs, e.g. the BXD or BXA panels) and tissues have been studied in eQTL screens [2-9]. The field is now expanding with the study of outbred mice [10]. Many of these data have been uploaded to the GeneNetwork database [11], which have made this the central repository for mouse and rat eQTL data. While eQTL publications on rats have been scarcer, there have been a few studies, for example using the BXH/HXB panel of recombinant inbred strains [12, 13].

This chapter provides a computational protocol for eQTL analysis on RIL crosses in mice and rats. The protocol can easily be adapted to suit other genetic populations, such as backcrosses or intercrosses [14].

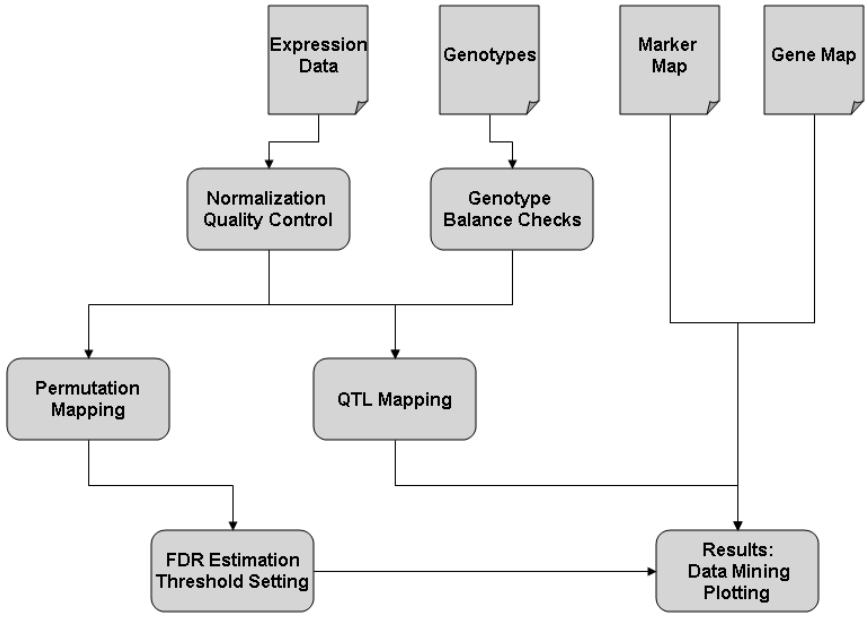


Figure 1 - Flowchart of eQTL mapping protocol

3.2 Materials

3.2.1 Hardware and software requirements

The protocol requires:
- R (www.r-project.org): R is a programming environment for statistical computing and graphics. It is available under the GNU General Public License on Windows, Linux/Unix and Mac systems. R has a command line-based interface and is widely used in the field of biostatistics thanks to the availability of multiple add-on packages designed to address specific biological analyses. All the code lines and functions presented in courier font are written in R language. Detailed knowledge of R programming is not required but the interested reader can go to the R tutorial: http://cran.r-project.org/doc/manuals/R-intro.pdf.
- CPU/memory requirements: this protocol is illustrated with a sample dataset of 100 genes, so that the protocol will run well on a regular desktop computer. For a real genome-wide experiment, you are strongly advised to use multiple-core machines (*see* **Note 1** on parallel computation).

3.2.2 Dataset

The methods we describe here are showcased on Illumina BeadArray data. Illumina is an increasingly popular technology for gene expression profiling and uses arrays containing multiple beads with 50-mer probes attached. Illumina bead arrays have been developed for a number of species including humans, mice, and rats. The protocol described in this chapter is not however specific to Illumina data and can be applied to virtually any technology. Some adjustments will need to be made in the particular case of Affymetrix arrays (*see* **Note 2**) due to specificities of this technology (i.e. multiple probes per gene).

For this protocol, we use a small sample dataset of 100 expression traits for efficiency purposes. This dataset was extracted from a survey of hematopoietic stem cells in a population of 24 mouse recombinant inbred strains (BXD). This dataset and an electronic version of the code presented in this chapter are available at the following URL http://gbic.biol.rug.nl/supplementary/2008/linkageGG.

Genotype data

The genotype data should be prepared as a tab-delimited file: each column represents one individual, each row a different marker (*see* **Table 1**). Values are either 1 for the first parental strain, 2 for the second parental strain, or 1.5 in the case of heterozygote individuals (these should be rare in the case of RILs). The markers are ordered by genomic location as in the marker map file (see below).

Genetical genomics screens can be very expensive; the costs of sample preparation, microarrays and genotyping must be multiplied by the population size. Selective genotyping is usually not a realistic option to reduce the costs in this context because the samples with the most informative genotypes depend on which gene is being considered. We therefore assume here that the genotype data is complete as is usually the case in genome-wide eQTL studies on recombinant inbred lines. The interested reader may want to refer to **Note 3** for software to handle missing information or sparse marker maps.

	BXD6	BXD28	BXD19	BXD15	BXD40	BXD12	BXD31
rs6376963	2	2	2	2	1	2	1
rs6298633	2	2	2	2	1	2	1
D1Mit1	2	2	2	2	1	2	1
rs3654866	2	1	2	2	1	2	1
rs3088964	2	1	2	2	1	2	1

Table 1 - Example of genotype data in tab-delimited format. The columns represent the different recombinant inbred lines (here from the BXD cross). The rows are different markers. RILS are homozygous 1 for the B6 allele or homozygous 2 for the DBA2 allele.

Expression data

BeadStudio, the standard Illumina software, produces probe data from bead level intensities. The output of BeadStudio contains several columns per sample. In this chapter, we use the raw bead summary data as output by BeadStudio. From the BeadStudio output files, we extract the AVG_SIGNAL columns per sample. These columns contain raw averaged bead intensities for each probe. The expression data are stored in a tab-delimited file, where each column refers to one individual and each row to a probe, as shown in **Table 2**.

Some authors have suggested alternative pre-processing methods for Illumina data (*see* **Note 4** for references).

	BXD6	BXD28	BXD19	BXD15	BXD40
GI_84579826-I	341.668	349.7453	509.0667	495.4675	591.0002
GI_84579830-A	105.3439	113.3545	117.8497	111.6411	109.7728
GI_84579883-I	121.9119	126.5275	126.1814	132.6144	119.5611
GI_84579884-A	138.155	138.7963	158.4077	150.2157	133.7268
GI_84579905-A	189.2942	180.8074	274.7991	367.868	204.4543
GI_84662726-I	148.5721	147.1926	153.2858	145.0625	135.5658
GI_84662775-S	132.148	125.3375	139.3298	136.605	130.0435

Table 2 - Example of raw expression data in tab-delimited format. The first column shows the unique probe IDs, the other columns refer to the samples denoted here by their RIL numbers.

Marker map

We also need a genetic map with the genomic positions of the markers in the genotype file. These positions can be specified in centimorgans (cM) or in mega base pairs (as one or both present). The marker map should be a text file in tab-delimited format.

	Marker_Chr	Marker_cM	Marker_Mb
rs6376963	1	0.895	5.008089
rs6298633	1	2.367	6.820241
D1Mit1	1	3.549	11.50072
rs3654866	1	5.797	13.69223
rs3088964	1	6.962	15.19202

Table 3 - Example of marker data in tab-delimited format. The first column contains marker IDs, the second column contains chromosome numbers, the third column contains centimorgan positions, and the last column base pair positions.

Probe and gene annotation data

We finally need a file containing all the relevant probe information, including the genes targeted by the probes and genomic positions of the probes. This file should also be in tab-delimited format. Annotations provided by microarray manufacturers are often not complete or up-to-date. It is sometimes necessary to re-annotate the probes based on a BLAT search of probe to genome sequence. See **Note 5** for some tools which enable such re-annotation.

	Probe_Chr	Probe_Mb	Gene_Symbol	Gene_Description	Gene_ID
GI_84579826-I	10	87.87682	*Gnptab*	N-acetylglucosamine-1-phosphate transferase, alpha and beta subunits	432486
GI_84579830-A	12	8.560399	*Slc7a15*	solute carrier family 7 (cationic amino acid transporter, y+ system), member 15	328059
GI_84579883-I	11	120.7646	*Slc16a3*	solute carrier family 16 (monocarboxylic acid transporters), member 3	80879

Table 4 - Probe annotations. The first column contains the probe IDs, the second column contains chromosome numbers, the third column base pair positions and the last columns give gene information.

3.3 Methods

3.3.1 Experimental design

When profiling many samples with microarrays, it is often necessary to divide the

samples into batches, which may then be profiled at different times or even dates. Attention should be paid to the random assignment of samples to batches in order to minimize the influence of confounding factors. The number of batches should be small (so that it won't take too many degrees of freedom in the analysis) and the batches should preferably be of equal size. Obviously, you should keep track of this batch organization and of any relevant information possibly associated with the profiling process. The eQTL analysis procedure can be adapted to take into account batch effects as described in **Note 6**.

Special considerations apply to sex, treatment, or environmental factors. If sex is not a factor of interest in the study, it is safer to limit the experiment to males or females only. Alternatively, if individuals of different sexes or different treatments or conditions are used, the model used for the eQTL analysis should account for these as additional factors (*see* **Section 3.3.6**). Strategies have been developed to optimize the power of the eQTL study with multiple conditions (e.g. *see* [15]).

The population size is obviously a critical choice. While more is always better in terms of statistical power, you must find the right balance between the costs incurred by microarray screens and the numbers of degrees of freedom necessary to fit the models you are using in your study. The relationship between population size and power in classical QTL analysis is discussed and illustrated in [16].

3.3.2 Loading the data into R

The following commands will import the data into the R workspace:

```
> rawExpr <-
as.matrix(read.csv(file="raw_data.txt",row.names=1,header=TRUE,sep="\t"))
> genotypes <-
as.matrix(read.csv(file="genotypes.txt",row.names=1,header=TRUE,sep="\t"))
> markerMap <-
as.matrix(read.csv(file="markerMap.txt",row.names=1,header=TRUE,sep="\t"))
> geneMap <-
as.matrix(read.csv(file="geneMap.txt",row.names=1,header=TRUE,sep="\t"))
```

We convert the data to logarithmic scale with:

```
> log2expr <- log2(rawExpr);
```

3.3.3 Useful checks on the data

Clustering of the expression data

A rapid clustering of the samples can detect major correlation structure such as caused by batch effects (see **Note 6** on how to deal with these artifacts).

```
>sample_clustering <- hclust(dist(t(log2expr)));
>plot(sample_clustering);
```

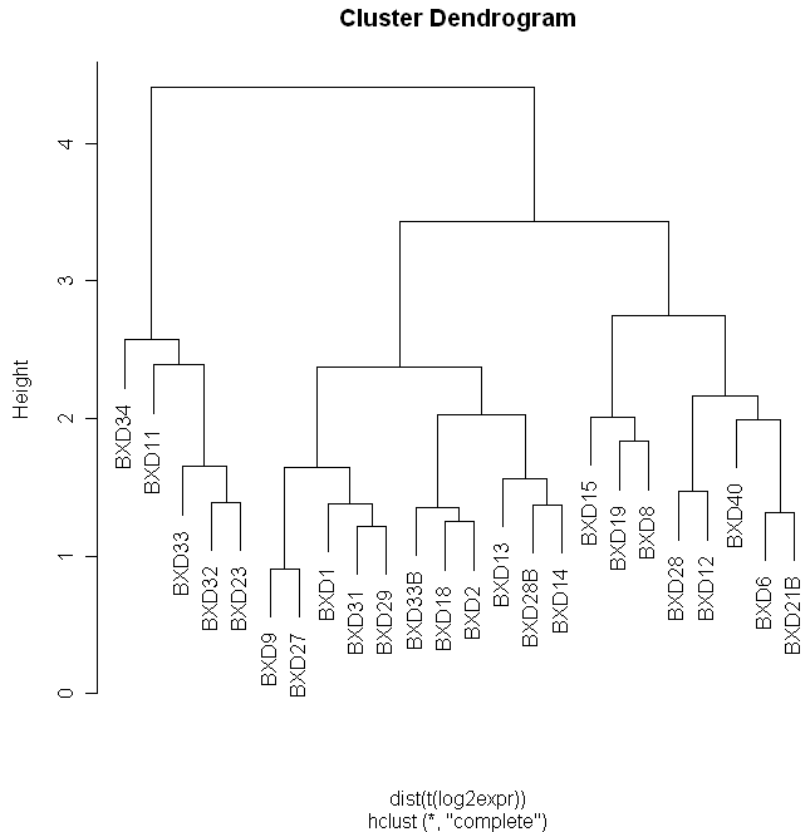


Figure 2 - Hierarchical clustering of samples using the raw expression data. From this plot, the samples may appear to be divided into three separate clusters. If these correspond to experimental batches, it would be wise to include them in the model (see Note 6).

At this stage, it is advisable to go back to the information collected during the wet lab process (*see* **Section 3.3.1**) to try to match that information with possible clusters.

Genotype imbalance

It may happen that one of the parental genotypes is very poorly represented at some markers, especially with a small population size. Such imbalance may be caused at random or by segregation distortion, and can lead to an acute sensitivity to outliers; it should therefore be watched carefully. **Figure 3** illustrates the genotype distribution across the markers. The code for plotting the genotypes diagnosis graph is:

```
#The following vector contains all the chromosome lengths
#in Mb for plotting purposes
> chr.lengths<-c(198,182,160,156,152,150,146,133,125,130,
122,121,121,124,104,99,96,91,62,166);
> names(chr.lengths)<- c("1","2","3","4","5","6","7","8",
"9","10","11","12","13","14","15","16","17","18","19","X");
```

```
#Check of segregation distortion
> plotGenotypeBalance <- function(genotypes,markerMap,chr.lengths)
{
  op <- par()
  mk_pos<-as.numeric(markerMap[,"Marker_Mb"]) +
       diffinv(chr.lengths)[match(markerMap[,"Marker_Chr"],names(chr.lengths))];
  breaks<- c(0,
        apply(cbind(mk_pos[1:length(mk_pos)-1],mk_pos[2:length(mk_pos)]),1,mean),
        mk_pos[length(mk_pos)])
  pos<-rep(mk_pos,ncol(genotypes))
  geno_fac<-factor(c(as.numeric(genotypes)))
  par(fig=c(0.001, 0.999, 0.28, 1),mai=c(0,0,1,0))
  spineplot(pos,geno_fac,breaks=breaks,xaxlabels='',
            border=NA,col=c("white","black","grey"),
            yaxlabels='',main="Genotype balance")
  chr_col<-
c("GREY","WHITE")[match(markerMap[,"Marker_Chr"],names(chr.lengths))%%2+1]
  par(new=T, fig=c(0.001, 0.999, 0.1, 0.28),mai=c(1,0,0,0))
  spineplot(mk_pos,factor(chr_col),breaks=breaks,
            border=NA,xaxlabels="",yaxlabels='',
            ylab="Chr",xlab="Marker Positions")
  par(op)
}
> plotGenotypeBalance(genotypes,markerMap,chr.lengths)
```

If this “information content” plot reveals a region with such imbalance, QTLs mapped in this region should be carefully scrutinized, since the minor genotype group will be extremely sensitive to outlier samples. The superimposition of the information content on the QTL profiles can provide additional insight into local variations in the statistical power available to detect eQTLs: the regions with the best power being those where the genotypes are perfectly balanced (50% for both parental genotypes).

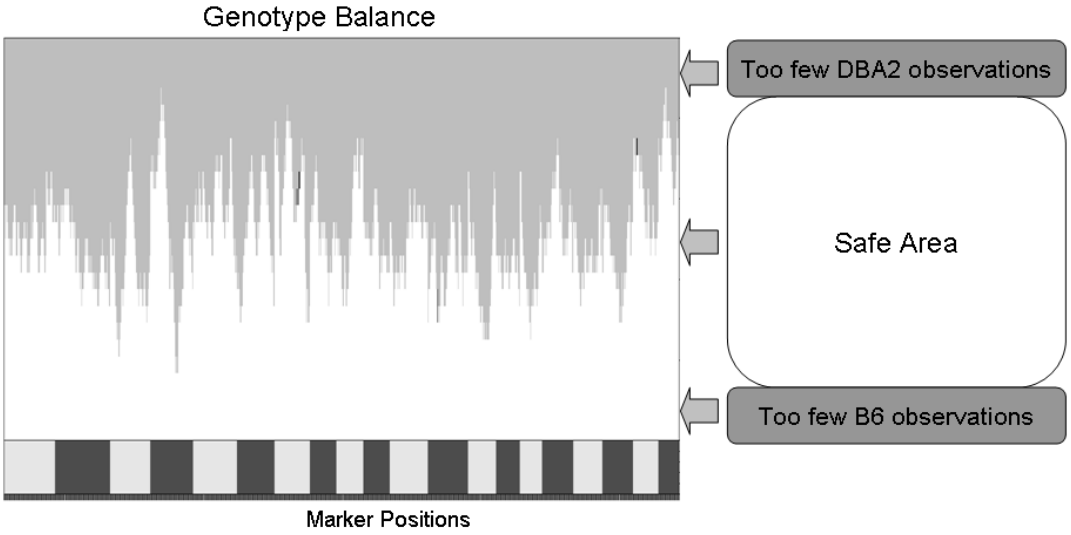


Figure 3 - Genotype balance plot. This plot represents the proportions of individuals with either genotype: white and grey denote the two parental genotypes; heterozygotes appear in black in the middle.

3.3.4 Normalization of the expression data

Microarray data of multiple samples need to be normalized (i.e. converted to the same scale) to allow them to be compared across samples. Normalization removes some of the between-array technical variation. A robust, simple, and efficient method is quantile normalization [17], which is widely used and has been shown to be one of the most appropriate methods in the context of eQTL mapping [18]. Quantile normalization orders intensities per sample and then replaces the intensity by the mean of the measurement at that rank in all the samples. An implementation of this normalization method is available in the Bioconductor Affy package [19]. Our sample dataset has already been normalized, and should therefore not be re-normalized.

The commands to normalize a complete microarray dataset are given below.

Installing and loading Affy R library:

```
> source("http://bioconductor.org/biocLite.R");
> biocLite("affy");
> library(affy);
```

These are the commands that apply to quantile normalization:

```
>normExpr <- normalize.quantiles(log2expr);
>dimnames(normExpr)<-dimnames(rawExpr);
```

It is advisable to perform similar checks to those described in **Section 3.3.3** on the normalized data to control how the normalization procedure affects the data structure.

```
#Our sample dataset was extracted from a complete dataset which was already norma-
lized.
> normExpr<-log2expr;
```

3.3.5 Mapping

Definition of the model

The first step of the actual eQTL analysis is to define the relevant model to use. In the simplest case, namely single-marker mapping without batches and without different environments, the model only includes the genotype effect:

$$\mathbf{Y_i = m_i + G_j + e_{ij}}$$

where Y_i is the expression measurement for probe i, m_i is the mean intensity of probe i over all samples, G_j is a factor containing the genotypes at marker j, and e_{ij} is the error term.

Fitting the model

The following commands are used to first fit the model using the lm() function and then retrieve significances (p-values) at each marker along the genome. It is as-

sumed that the order of the columns (samples) in the normExpr matrix matches the order of the columns in the `genotypes` matrix.
Single Marker Mapping function:

```
> singleMarkerMapping<-function(traits,genotypes)
{
  qtl_profiles <- NULL;
  for (i in 1:nrow(traits))
  {
    current_profile<-NULL;
    for (j in 1:nrow(genotypes))
    {
      model <- traits[i,] ~ genotypes[j,];
      anova_table<-anova(lm(model));
      current_profile<-c(current_profile, -log10(anova_table[1,5]));
    }
    qtl_profiles<-rbind(qtl_profiles,current_profile);
  }
  rownames(qtl_profiles) <- rownames(traits);
  colnames(qtl_profiles) <- rownames(genotypes);
  qtl_profiles ;
}
```

Then applying the single-marker mapping function to the expression values and the genotypes:

```
> qtlProfiles<-singleMarkerMapping(
     traits = normExpr, genotypes = genotypes);
```

Warning: this step can be computationally very intensive (*see* **Note 1**).

Processing and visualizing the results

Using this single-marker mapping approach, we obtain p-values for linkage for each gene with each of the markers on our genetic map. The p-value distribution across the genome for a given gene is termed the QTL profile of that gene and can be plotted as shown on **Figure 4** using the following function:

```
> plotQTLProfile<-
function(qtl_profile,markerMap,chr.lengths)
{
  chrStrips<-seq(0,0,length=sum(chr.lengths))
  for(i in 2*0:as.integer((length(chr.lengths)-1)/2)+1)
  {
    for (j in
        (diffinv(chr.lengths)[i]:diffinv(chr.lengths)[i+1]))
    {
      chrStrips[j]<-1;
    }
  }
  plot(chrStrips,type='h',col="#ECECEC",xlab='',
       ylab='',axes=F,ylim=c(0,1));
  par(new=TRUE);
  marker_x_positions <-
    as.numeric(markerMap[,"Marker_Mb"]) +
    diffinv(chr.lengths)[
  match(markerMap[,"Marker_Chr"],names(chr.lengths))
];
```

```
  plot(
        y=qtl_profile,x=marker_x_positions,
        xlim=c(0,sum(chr.lengths)),
        ylim=c(0,max(max(qtl_profile)+1,6)),
        type='l',xlab="Marker Position", ylab="-log(p)"
      );
}
```

We use this function to plot the QTL profile of our first probe, which targets the *Gnptab* gene.

```
> plotQTLProfile(qtlProfiles[1,],markerMap,chr.lengths)
```

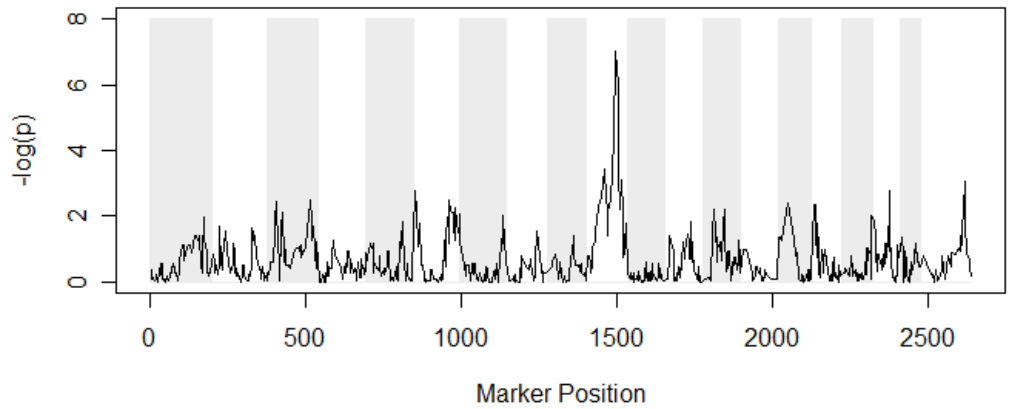


Figure 4 - Example of a QTL profile plot. This profile shows a QTL peak for the *Gnptab* gene on chromosome 10.

Here we set the significance threshold for detection of an eQTL to –log10(p-value) > 6 (*see* **Section 3.3.7** on how to set significance thresholds). The following function extracts primary QTL peaks from the QTL profiles:

```
> getQTLMaxPeaks <- function(qtl_profiles,threshold)
{
  max_index <- function(v)
  {
    which(v == max(v,na.rm=T))[1];
  }
  maxQTLs <- cbind(
               rownames(qtl_profiles),
               colnames(qtl_profiles)[
apply(qtl_profiles,1,max_index)],
               apply(qtl_profiles,1,max,na.rm=T));
  maxQTLsThreshold <-
  matrix(
maxQTLs[which(as.numeric(maxQTLs[,3])>=threshold),],
ncol=3);
  colnames(maxQTLsThreshold) <- c("Probe","Marker","p");
  maxQTLsThreshold;
}

> QTLPeaksThresh3 <-
getQTLMaxPeaks(qtlProfiles,threshold=3)
```

We now have a list of all the significant primary eQTLs, reported in triplets containing: the probe, the marker with the smallest linkage p-value, and the largest "minus log p-value" of that linkage. It is sometimes useful to report a confidence interval too. Two approaches are commonly employed: bootstrap and 1-lod-score drop off [20, 21]. We can now generate an eQTL dot plot, which provides an informative summary of the mapping results.

```
> qtlDotPlot <- function(QTLPeaks,markerMap,geneMap,chr.lengths)
{
 chrStrips <- seq(0,0,length=sum(chr.lengths));
 for(i in 2*0:as.integer((length(chr.lengths)-1)/2)+1)
 {
  for (j in  (diffinv(chr.lengths)[i]:diffinv(chr.lengths)[i+1]))
  {
   chrStrips[j] <- 1;
  }
 }
 plot(chrStrips, type='h', col="#ECECEC", xlab='', ylab='', axes=F,ylim=c(0,1));
 par(new=TRUE);
 QTL_Positions <- as.numeric( markerMap[QTLPeaks[,"Marker"],"Marker_Mb"]) +
   diffinv(chr.lengths)[match(markerMap[QTLPeaks[,"Marker"],"Marker_Chr"],
                              names(chr.lengths))];
 Gene_Positions <- as.numeric(geneMap[QTLPeaks[, "Probe"],"Probe_Mb"]) +
  diffinv(chr.lengths)[match(geneMap[QTLPeaks[,"Probe"],"Probe_Chr"],
                             names(chr.lengths))];
plot(x=QTL_Positions, y=Gene_Positions, xlim=c(0,sum(chr.lengths)),
  ylim=c(0,sum(chr.lengths)), pch=19, xlab="QTL position", ylab="Gene Position");
}
> qtlDotPlot( QTLPeaksThresh3,markerMap,geneMap,chr.lengths);
```

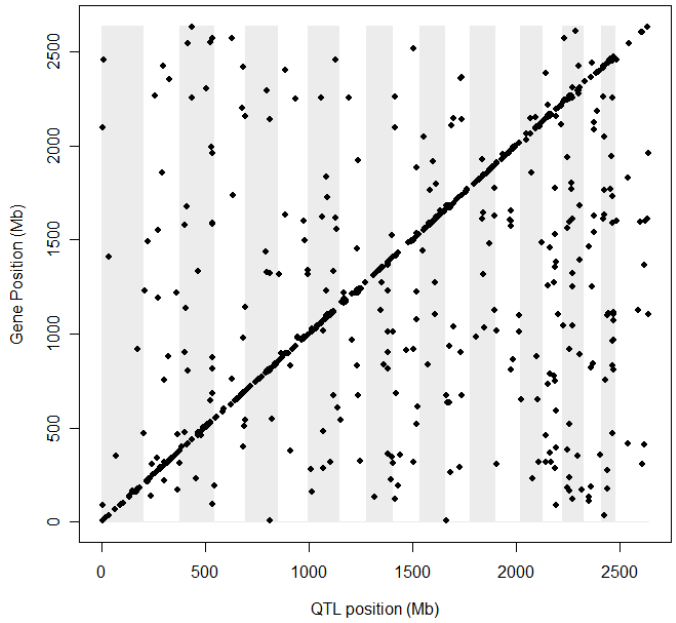


Figure 5 - An example of an eQTL dot plot. Each dot represents a significant eQTL, with the gene position on the Y axis and the QTL position on the X axis. (This plot was obtained using the complete dataset, not just the 100 probe subset we are using as a sample for this chapter.)

Locally acting eQTLs appear on the diagonal and are often over-represented. You can also sometimes see the presence of vertical bands (typically between 0-8), which have been suggested to reflect the presence of regulation hotspots: a distant eQTL controlling or regulating many genes [22].

3.3.6 More elaborate models

Multiple QTL mapping

To potentially improve statistical power for eQTL detection, it can be worthwhile fitting multiple QTL models, for example, by a stepwise procedure: correct for the first (most significant) eQTL effect found, and then map the corrected data to detect a second eQTL. The two-eQTL model is as follows:

$$\mathbf{Y_i = m_i + G_k + G_j + e_{ij}}$$

where G_k is the genotype vector at the first QTL position.

In this code example, we look for secondary eQTLs for the probes, for which a primary QTL has already been identified in **Section 3.3.5**: we use `maxQTLPeaksThresh3`.

```
> secondaryMarkerMapping <- function(traits,genotypes,primaryQTLs)
{
 if (paste(rownames(traits),collapse='') !=
paste(primaryQTLs[,"Probe"],collapse=''))
 {
  print("Error: Traits submitted do not match traits with primary eQTLs.");
  return;
 }
 qtl_profiles <- NULL;
 for (i in 1:nrow(traits))
 {
   current_profile <- NULL;
   for (j in 1:nrow(genotypes))
   {
    model <- traits[i,] ~ genotypes[primaryQTLs[i,"Marker"],]+ genotypes[j,];
    anova_table <- anova(lm(model));
    current_profile <- c(current_profile, -log10(anova_table[2,5]));
   }
   qtl_profiles <- rbind(qtl_profiles,current_profile);
  }
  rownames(qtl_profiles) <- rownames(traits);
  colnames(qtl_profiles) <- rownames(genotypes);
  qtl_profiles
}

> secondaryQTLProfiles <- secondaryMarkerMapping(
normExpr[QTLPeaksThresh3[,"Probe"],],
genotypes,QTLPeaksThresh3 );
```

Using the function defined in **Section 3.3.5** to extract QTL peaks, we can

now create a list of the significant secondary eQTLs for a given threshold:

```
> secondaryQTLPeaksThresh3 <- getQTLMaxPeaks(secondaryQTLProfiles, threshold=3);
```

It is, of course, possible to include three or more QTLs per gene by extending the model. However, you should be cautious because of overfitting issues. This sequential way of defining the co-factors to include in the model may not be optimal, and there are a number of more advanced strategies which address the problem of model selection (*see* **Note 7**).

Epistasis

In the previous step we have presented how to detect multiple QTLs per gene. However, we have not tested for interaction between the eQTLs (i.e. if the effect of one eQTL is modulated by the effect of the second one). Such complex mechanisms are common in gene regulation and are termed epistasis. Our eQTL analysis model can again be extended to take such epistasis effects into account.

$$\mathbf{Y_i = m_i + G_k + G_j + G_j * G_k + e_{ij}}$$

The code below tests for epistasis between two eQTLs that we identified for the gene in **Section 3.3.6**:

```
> my_probe<-secondaryQTLPeaksThresh3[1,1];
#Probe of the gene we will test for epistasis
> G1 <-genotypes[QTLPeaksThresh3[
 which(QTLPeaksThresh3[,"Probe"]==my_probe), "Marker"],];
> G2 <- genotypes[secondaryQTLPeaksThresh3[1, "Marker"],];
> model_epistasis<- normExpr[my_probe,] ~ G1 + G2 + G1:G2;
> anova_table <- anova(lm(model_epistasis));
> interaction_p_value <- anova_table[3,5];
```

In this example, the *p*-value is insignificant and there is no evidence for epistasis. Interactions also occur between two loci whose main effects (terms G1 and G2 in the model) may not be significant on their own. It can therefore be relevant to screen for interactions for any possible pairs of loci, but this can sometimes be computationally unrealistic (a two-dimensional genome scan leads to a huge multiple-testing problem). For more guidance on strategies for epistasis testing, *see* **Note 7**.

Adding environments / treatments

Genetical genomics studies can provide insights into the way different environments or treatments affect the regulation of gene expression. When combining the genetic perturbation naturally present in inbred populations with the effect of different environments, the study of the interaction between those two causes of variation can teach us about the plasticity of eQTLs [15, 23]. We can illustrate this with the

example of the study of gene expression regulation across several cell types. In the example below, expression profiles were collected from four distinct cell types. The following model can be used:

$$\mathbf{Y_i = m_i + CT + G_j + CT * G_j + e_{ij}}$$ where CT is the cell type factor.

For this example, we need to load new data files:

```
>genotypes4CT <- as.matrix(read.csv(file="genotypes4ct.txt",sep="\t",row.names=1));
>expr4CT <-as.matrix(read.csv(file="expr4ct.txt", sep="\t", row.names=1));
#the cell types are coded as “1”, “2”, “3”and “4”
>CT.factor <- factor(c(rep(1,24), rep(2,25), rep(3,22), rep(4,25)));
```

The mapping function therefore becomes:

```
>singleMarkerMappingWithEnv <-
function (traits, genotypes, env.factor)
{
  P1 <- P2 <- P3 <- P4 <- NULL;
  for (i in 1:nrow(traits))
 {
   p1 <- p2 <- p3 <- NULL;
   for (j in 1:nrow(genotypes))
   {
     model_environment<- traits[i,] ~ factor(env.factor)+
         genotypes[j,] + factor(env.factor):genotypes[j,];
     anova_table <- anova(lm(model_environment));
     p1 <- c(p1,-log10(anova_table[[5]][1]));
     p2 <- c(p2,-log10(anova_table[[5]][2]));
     p3 <- c(p3,-log10(anova_table[[5]][3]));
   }
   P1 <- rbind(P1,p1);     # Env
   P2 <- rbind(P2,p2);     #qtl
   P3 <- rbind(P3,p3);     #qtlxEnv
  }
  dimnames(P1) <- list(rownames(traits),rownames(genotypes));
  dimnames(P2) <- list(rownames(traits),rownames(genotypes));
  dimnames(P3) <- list(rownames(traits),rownames(genotypes));
  results<-list();
  results$Profiles_Environment <- P1;
  results$Profiles_QTL <- P2;
  results$Profiles_QTLxEnvironment <- P3;
  results;
}
```

This function outputs three *p*-values for each trait-marker pair: the first p-value indicates the significance level of the environment term (a low *p*-value indicates a clear overall influence of the environment on the trait; this *p*-value is *not* valid if the environment has not been randomly allocated to samples). The second *p*-value is the significance of the main genotype effect at that marker, while the third *p*-value reflects the significance of the genotype by environment interaction term.

```
>results4CT <-singleMarkerMappingWithEnv(expr4CT,genotypes4CT,env.factor=CT.factor);
>interactionQTLsThresh3<-
getQTLMaxPeaks(results4CT$Profiles_QTLxEnvironment,threshold=3);
```

A norm of reaction plot (**Figure 6**) can show how the eQTL effect is modulated by the environment. It can be obtained using the following function:

```
>plotNormOfReaction<-function(trait,genotype,env.factor)
{
 env.factor <- as.numeric(env.factor);
 plottingColors <- c("black","black","lightgrey","grey");
 yrange <- range(trait) + c((range(trait)[1] - range(trait)[2])/5,
   (range(trait)[2] - range(trait)[1])/5);
 plot(y=trait,x=env.factor,xlim=range(env.factor), ylim=yrange,
   col=plottingColors[2*genotype], xaxt='n', pch=19, xlab="Environment",
   ylab="expression for individuals");
 meanGroupValues <- matrix(nrow=2,ncol=length(unique(env.factor))) ;
 for (env in 1:length(unique(env.factor)))
 {
  meanGroupValues[1,env] <-
     mean(trait[intersect(which(genotype == 1), which(env.factor == env))]);
  meanGroupValues[2,env] <-
       mean(trait[intersect(which(genotype == 2), which(env.factor == env))]);
  text(env,yrange[1],env,cex=1.5,col="black");
 }
 par(new=T);
 matplot( y=t(meanGroupValues),xlim=range(env.factor), ylim=yrange, xlab='',
   ylab='', xaxt='n',yaxt='n', type='l',lty=1,lwd=4,col=c("black","grey"));
}
>plotNormOfReaction( expr4CT[interactionQTLsThresh3[1,"Probe"],],
  genotypes4CT[interactionQTLsThresh3[1,"Marker"],], CT.factor);
```

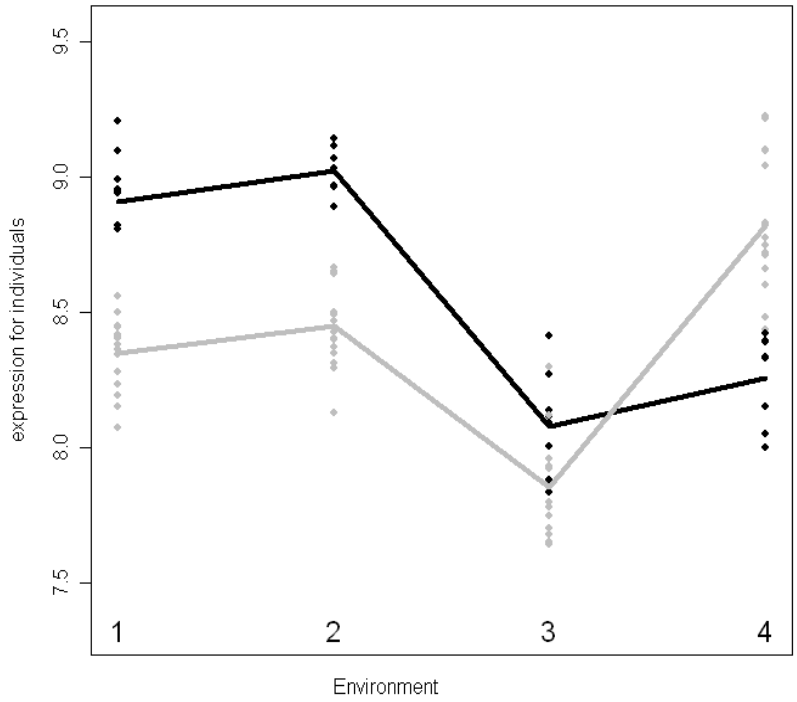


Figure 6 - Norm of reaction plot: eQTL by environment interaction, with the cell types on the X axis and the gene expression on the Y axis. Each dot is an individual sample measurement (black = B6, grey = DBA2). The lines represent mean values. The effect of the QTL is here reversed in cell type 4 compared with the other three cell types.

3.3.7 Determining the significance threshold

In eQTL analysis, the determination of the threshold for statistical significance is a critical aspect since multiple testing issues arise from both the high number of genes studied and the high number of genomic loci at which linkage is tested. The p-values yielded by the ANOVA must be adjusted to take into account these multiple testing issues. Bonferroni correction is somewhat too drastic here since the tests are not independent: firstly, the markers tested are intrinsically linked and thus correlated to their neighbors on each chromosome; and, secondly, large families of genes are known to be co-regulated, so there is also a correlation structure in that dimension.

A more appropriate approach is to estimate a False Discovery Rate (FDR) based on a permutation strategy [24]. A carefully designed permutation procedure will make it possible to estimate the null distribution. The principle is to apply the exact same analysis protocol to permuted datasets, calculate the average number of rejected null hypotheses for a certain *p*-value threshold in those permuted datasets, and then derive an FDR estimate, at that *p*-value threshold, as the average number of rejected hypotheses in the permuted datasets divided by the number of rejected hypotheses at the same threshold in the true data.

Different permutation strategies are possible: we advise permuting only the genotypes of the individuals, while conserving trait values (gene expression measurements). This ensures that the permutation procedure does not break the internal correlation structure of the data (within markers and within genes), but that any linkage detected between a marker and gene expression in a permuted dataset is a false-positive [25].

The following function here estimates the number of false-positives obtained with a permuted dataset for a range of p-value thresholds in the single-marker mapping case discussed in **Section 3.3.5**.

```
>estimateFalsePositives <- function(traits,genotypes,threshold_range,nperm)
{
  permuteGenotypes <- function(geno)
  {
   geno[,sample(1:ncol(geno),ncol(geno),replace=F)];
  }
  Counters <- NULL
  for (i in 1:nperm)
  {
    permGenotypes <- permuteGenotypes(genotypes);
    current_profiles <- singleMarkerMapping(traits,permGenotypes);
    current_counters <- NULL;
    for (thresh in threshold_range)
    {
      current_counters <- c(current_counters,
         length(which(apply(current_profiles,1,max)>=thresh)));
    }
    Counters <- rbind(counters,current_counters);
  }
  colnames(counters) <- apply(matrix(threshold_range),1,as.character);
```

```
for (j in 1:nrow(counters))
  {
   rownames(counters)[j] <- paste("permutation round",j);
  }
  Counters;
}
> false_positive_estimates <-
estimateFalsePositives(normExpr, genotypes, threshold_range=c(3,4,5,6,7), nperm=10);
#In this example, for efficiency purpose we run only 10 permutations round. For a
#reliable estimation, 100 permutations would be a minimum.
```

We can derive an estimate of the FDR, *e.g.* for a p-value threshold of 3:

```
#number of rejected null hypothesis:
positives_thresh3 <- nrow(QTLPeaksThresh3)
#number of false positives
false_positives_thresh3 <- mean(false_positive_estimates[,"3"])
#FDR
FDR_thresh3 <- false_positives_thresh3/positives_thresh3;
```

The result here gives a very high FDR (>50%) which means we need to use a more stringent threshold than –logp> 3. This code can easily be adapted to estimate the FDRs for other mapping procedures, the principle being that the permuted data should be analyzed with the same model and the same procedure as the real data. The cases of complex models for stratified data or interacting factors require adapted permutation procedures [26].

Another advantage of this permutation procedure is that it allows an unbiased estimation of the significance of the number and size of eQTL hotspots. There is some speculation that some hotspots may be the result of false-positive linkage of groups of correlated genes to random genome positions (with no regulatory connection) [27, 28]. Calculating the size and the number of the hotspots obtained with permuted datasets that have retained the correlation between genes is a straightforward manner of testing the significance of hotspots [25].

Some authors have suggested using different thresholds for local and distant eQTLs: detecting local effects does not involve genome-wide testing of loci and can therefore be controlled with relaxed thresholds [29]. Finally, it is important to take into account the fact that sex chromosomes have specific properties which require different thresholds for sex and autosomal chromosomes (*see* **Note 8**).

3.3.8 Interpretation of the results

A typical eQTL analysis will yield hundreds or thousands of genetic linkages. Extracting meaningful biological information from the results can prove challenging.

Local eQTLs typically offer insights into possible *cis*-regulatory differences between the two alleles. Inspection of the polymorphisms in the regulatory regions of the gene can provide insight into the possible molecular mechanism (e.g. a SNP located in a transcription factor binding site located in the promoter region of the

gene). Polymorphisms located within the probe target regions can also create a technically false-positive eQTL (*see* **Note 9**)

A distant eQTL indicates the presence of a distant regulator (e.g. a transcription factor or a miRNA gene) at the QTL location. This regulator may either be locally regulated or contain a non-synonymous polymorphism affecting its function. It is, however, usually difficult to directly pinpoint the regulator because of the relatively poor mapping resolution (a QTL typically spans several Mb and contains tens to hundreds of candidate genes).

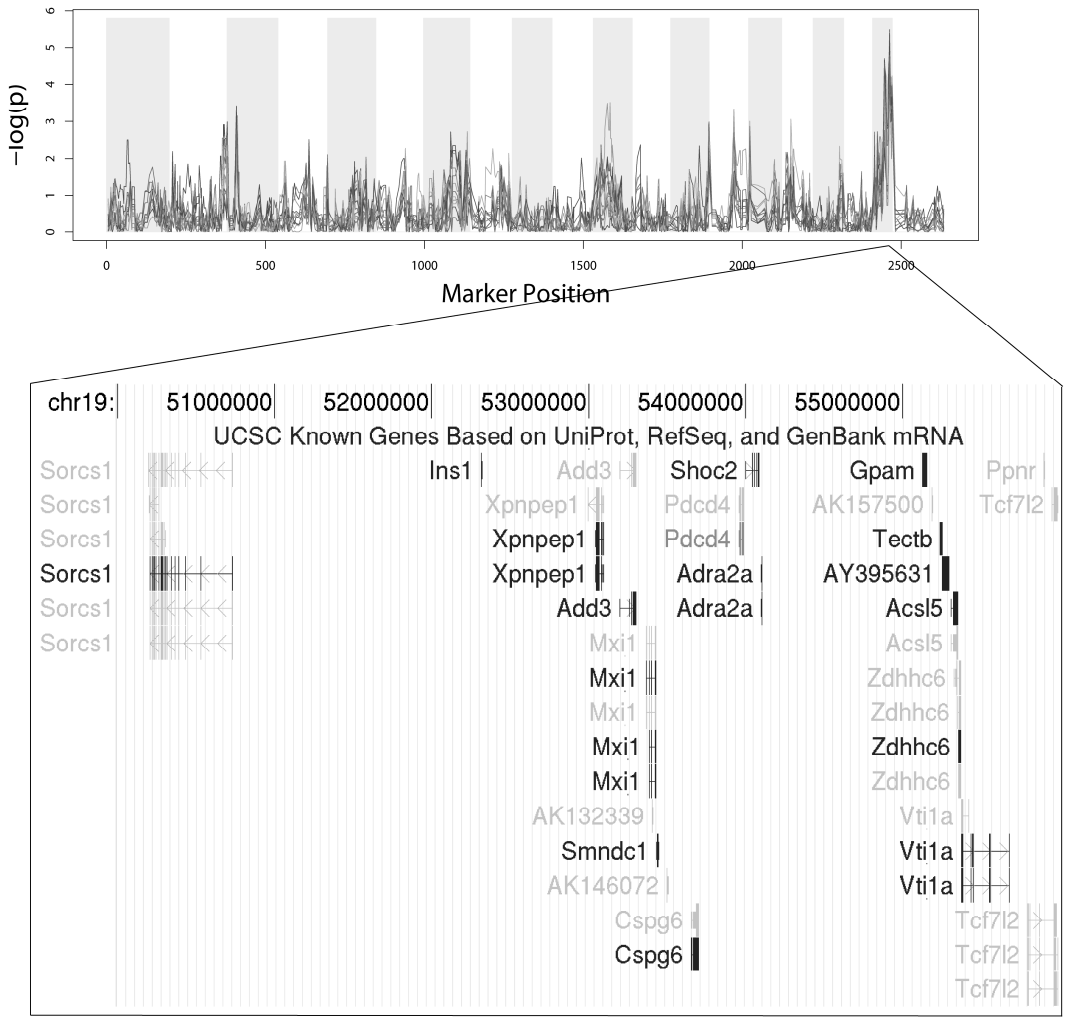


Figure 7 - Multiple possible candidate regulators: on the top panel the QTL profiles of 16 genes show a common peak. A large number of genes, illustrated by a UCSC genome browser screenshot **(lower panel)**, lie within the confidence interval of that eQTL.

Hotspots, which are large groups of genes having co-localizing eQTLs, may reveal the action of master regulators (i.e. genes controlling many others). It is possible to design strategies to reduce the large number of candidate regulators (typically hundreds) that fall into the hotspot QTL region. We present here one small sample hotspot, illustrated in **Figure 7**. Sixteen genes were found to share a QTL. We investigate the possibility of a common regulator located in that QTL region. Different candidates (genes physically located within the QTL) are ranked according to their correlation with each of the hotspot genes. Using RankProduct [30] it is then possible to prioritize the candidates (**Table 5**).

Hotspot elucidation, and more generally QTL gene candidate search are data-driven research processes which integrate heterogeneous types of information [31]: we have illustrated the use of correlation measurements. Other data types can include Transcription Factor Binding Site (TFBS) modules investigation, gene annotations, and databases of known protein-protein interactions [32]. Methods such as the Rank Product can be used to prioritize candidates based on these criteria.

Candidate Regulators	Rank Product of Correlation with Hotspot Genes	*p*-Value
Mxi1	2.412555	<0.00001
Add3	2.641117	<0.00001
Smndc1	3.808562	0.0012
Shoc2	3.883223	0.00135
Gpam	4.179971	0.0027
Sorcs1	4.571672	0.00825
5830416P10Rik	5.282127	0.03055
Adra2a	7.849839	0.38
1700001K23Rik	9.409524	0.69665
Pdcd4	10.60708	0.8643
Gucy2g	10.74565	0.878475
Dusp5	11.54332	0.93845
Tcf7l2	11.95649	0.9578
Tectb	13.5352	0.99195
Zdhhc6	14.56703	0.99825
Ins1	15.25569	0.99965
Vti1a	15.5439	0.99975
Rbm20	16.02744	0.99995
Acsl5	16.25449	0.99995
Xpnpep1	18.07143	1

Table 5 - Prioritization of candidate regulators based on Rank Product of correlation with hotspot genes. The genes with the lowest p-values are those that correlate best with the hotspot genes and are therefore given top priority. *Mxi1* and *Add3* are the most likely candidates according to this correlation criterion.

3.4 Notes

1. Computational capacity issues
Some of the protocol steps (mapping, permutation procedure) can be computationally very intensive. If available, it is advisable to use a multi-core machine or a cluster of computers to perform these steps. The jobs can easily be separated by groups of probes, since every probe is here mapped separately. R/Parallel [33] is a useful R package, which allows R to run iterative tasks in parallel on multiple processors. Another trick that can be used to reduce the computing time is to drop redundant markers (neighboring markers with identical genotypes for all samples) at the start of the analysis.

Memory issues can also arise when huge matrices are created within R. The amount of memory used by R is, for example, limited to 1 GB in the default windows setup of the program. This memory limit can be extended using the command memory.limit(size_MB). However, the maximum memory cannot exceed the physical memory available in the computer. A possible workaround is to divide the traits into smaller entities and to write intermediary results to files.

2. Affymetrix-related issues
Affymetrix arrays differ from alternative expression profiling technologies by their use of probe sets made of multiple (10-16) probes targeting one gene. While a number of studies have focused on probe set summarized data to perform eQTL mapping, we suggest using a more subtle approach taking into account probe level intensities. This approach has been extensively described in [34].

3. Alternative software solutions
R/QTL [35] is an R package which includes many functions for mapping, including an algorithm to infer missing genotype data using Hidden Markov Models. GeneNetwork (www.genenetwork.org [11]) also offers eQTL analysis for user uploaded data, one trait at a time, and genome-wide analysis tools for a number of published datasets.

4. Alternative Illumina data pre-processing
Compared with Affymetrix for example, Illumina is a relatively new technology and standard analysis guidelines have yet to emerge. While in this chapter we illustrate our eQTL analysis with raw probe summarized data as output by BeadStudio, there are alternative possibilities which make use of bead level intensity data. See, for example, work by Dunning et al. [36, 37] for some methods and software.

5. Probe (re-)annotations
The correct annotation of probes is a critical aspect of any microarray analysis. It is especially crucial in the case of eQTL studies since the presence of subtle differenc-

es in the probe target sequences (SNPs) between parental lines can produce technical false-positive eQTLs (*see* **Note 9**). Annotation files provided by array manufacturers tend to be incomplete and outdated and do not include genetic variation information across different strains.

The strategy of probe re-annotation should therefore comprise a systematic BLAT search of each of the probe sequences against the latest genome assembly build, combined with a mining of polymorphism databases. This is obviously a gruesome task for which there are a few software tools available [38-40].

6. Batches and hidden factors

Microarray data are known to be very sensitive to the effect of batches, which can create artificial correlation. In the context of eQTL mapping, these effects can act as confounding factors and cause multiple spurious genetic linkages, often forming apparent hotspots: if the confounding factor influences many genes (as is the case with microarray batches) and if there is, by chance, correlation of that factor with the genotypes at a certain genomic locus, then all genes will artificially map to that region, misleadingly suggesting the presence of a master regulator [34]. If the confounding factor is known, it is possible to correct for its effect by adapting the mapping model. In the single marker mapping case:

$$\mathbf{Y_i = m_i + B + G_j + e_{ij}}$$ where B is the batch factor.

Caution: if multiple environments are used, it is usually required to account for the batch effects in an environment specific fashion. In this case, an appropriate model would be:

$$\mathbf{Y_i = m_i + B + T + B*T + G_j + G*T + e_{ij}}$$

The B * T allows for a more careful batch effect correction: for example if one gene was only expressed in one of the environments (in our example one tissue), then the batch effect could affect only the gene in that tissue, and should not be corrected in the samples belonging to the other environments.

7. Advanced model selection procedures

The selection of relevant co-factors and interaction terms in generalized models, and particularly in the context of QTL mapping, has been widely discussed in the scientific literature. Mapping of multiple QTLs and epistasis testing can be seen as model selection problems. For examples, see [41, 42].

8. The case of sex chromosomes

While most of the Y chromosome does not undergo recombination, the recombination rate of the X chromosome is slower than that of the autosomes. This has important consequences on the detection of significant QTLs. For a comprehensive view of these issues, *see* [43].

9. Probe hybridization artifacts

When several probes are available for the same gene, it is not uncommon to observe a difference in the mapping results of those probes: "the probes tell different stories" or statistically there is eQTL by probe interaction [34]. This can be explained either by biological mechanisms (alternative splicing) or by technical artifacts. Such technical artifacts may arise when a polymorphism is present in the sequence targeted by a probe [44]. If the probe was designed specifically based on the genome sequence of one of the parental strains, it is possible that some polymorphism causes the other genotype mRNA products to have a weaker binding affinity and thus a lower signal. Such effects will yield spurious local eQTL linkages. If the probes have been designed specifically based on the sequence of one of the two parental strains (say, strain A, and not strain B), it is possible to estimate roughly the number of local eQTLs affected by this issue. For example, if 65% of local eQTLs are linked with a higher expression of the gene for the A allele, while for the other 35% local eQTLs the B allele is more highly expressed. This contrasts with the 50%-50% expected without hybridization effect. In this case, we would expect 65-35 = 30% of eQTLs to be caused by this hybridization difference rather than by a real differential expression effect.

3.5 References

1. Jansen RC, Nap JP: **Genetical genomics: the added value from segregation**. *Trends Genet* 2001, **17**(7):388-391.
2. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G *et al*: **Genetics of gene expression surveyed in maize, mouse and man**. *Nature* 2003, **422**(6929):297-302.
3. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF *et al*: **Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'**. *Nat Genet* 2005, **37**(3):225-232.
4. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA *et al*: **Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function**. *Nat Genet* 2005, **37**(3):233-242.
5. Mehrabian M, Allayee H, Stockton J, Lum PY, Drake TA, Castellani LW, Suh M, Armour C, Edwards S, Lamb J *et al*: **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* 2005, **37**(11):1224-1233.
6. Lan H, Chen M, Flowers JB, Yandell BS, Stapleton DS, Mata CM, Mui ET, Flowers MT, Schueler KL, Manly KF *et al*: **Combined expression trait correlations and expression quantitative trait locus mapping**. *PLoS Genet* 2006, **2**(1):e6.
7. Wang S, Yehya N, Schadt EE, Wang H, Drake TA, Lusis AJ: **Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity**. *PLoS Genet* 2006, **2**(2):e15.
8. McClurg P, Janes J, Wu C, Delano DL, Walker JR, Batalov S, Takahashi JS, Shimomura K, Kohsaka A, Bass J *et al*: **Genomewide association analysis in diverse inbred mice: power and population structure**. *Genetics* 2007, **176**(1):675-683.
9. Wu C, Delano DL, Mitro N, Su SV, Janes J, McClurg P, Batalov S, Welch GL, Zhang J, Orth AP *et al*: **Gene set enrichment in eQTL data identifies novel annotations and pathway regulators**. *PLoS Genet* 2008, **4**(5):e1000070.
10. Ghazalpour A, Doss S, Kang H, Farber C, Wen PZ, Brozell A, Castellanos R, Eskin E, Smith DJ, Drake TA *et al*: **High-resolution mapping of gene expression using association in an outbred mouse stock**. *PLoS Genet* 2008, **4**(8):e1000149.
11. Wang J, Williams RW, Manly KF: **WebQTL: web-based complex trait analysis**. *Neuroinformatics* 2003, **1**(4):299-308.

12. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V *et al*: **Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease**. *Nat Genet* 2005, **37**(3):243-253.
13. Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, Fischer J, Maatz H, Kren V, Pravenec M *et al*: **Heritability and tissue specificity of expression quantitative trait loci**. *PLoS genetics* 2006, **2**(10):e172.
14. Darvasi A, Soller M: **Advanced intercross lines, an experimental population for fine genetic mapping**. *Genetics* 1995, **141**(3):1199-1207.
15. Li Y, Breitling R, Jansen RC: **Generalizing genetical genomics: getting added value from environmental perturbation**. *Trends Genet* 2008.
16. Van Ooijen JW: **LOD significance thresholds for QTL analysis in experimental populations of diploid species**. *Heredity* 1999, **83 (Pt 5)**:613-624.
17. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics (Oxford, England)* 2003, **19**(2):185-193.
18. Williams RBH, Cotsapas CJ, Cowley MJ, Chan E, Nott DJ, Little PFR: **Normalization procedures and detection of linkage signal in genetical-genomics experiments**. *Nat Genet* 2006, **38**(8):855-856.
19. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level**. *Bioinformatics (Oxford, England)* 2004, **20**(3):307-315.
20. Visscher PM, Thompson R, Haley CS: **Confidence intervals in QTL mapping by bootstrapping**. *Genetics* 1996, **143**(2):1013-1020.
21. Lander ES, Botstein D: **Mapping mendelian factors underlying quantitative traits using RFLP linkage maps**. *Genetics* 1989, **121**(1):185-199.
22. Darvasi A: **Genomics: Gene expression meets genetics**. *Nature* 2003, **422**(6929):269-270.
23. Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JA, Hazendonk E, Prins P, Plasterk RH, Jansen RC *et al*: **Mapping determinants of gene expression plasticity by genetical genomics in C. elegans**. *PLoS genetics* 2006, **2**(12):e222.
24. Churchill GA, Doerge RW: **Empirical threshold values for quantitative trait mapping**. *Genetics* 1994, **138**(3):963-971.
25. Breitling R, Li Y, Tesson BM, Fu J, Wu C, Wiltshire T, Gerrits A, Bystrikh LV, de Haan G, Su AI *et al*: **Genetical Genomics: Spotlight on QTL Hotspots.** *PLoS genetics* 2008, **(in press)**.
26. Churchill GA, Doerge RW: **Naive application of permutation testing leads to inflated type I error rates**. *Genetics* 2008, **178**(1):609-610.

27. de Koning DJ, Haley CS: **Genetical genomics in humans and model organisms**. *Trends Genet* 2005, **21**(7):377-381.
28. Perez-Enciso M: **In silico study of transcriptome genetic variation in outbred populations**. *Genetics* 2004, **166**(1):547-554.
29. Doss S, Schadt EE, Drake TA, Lusis AJ: ***Cis*-acting expression quantitative trait loci in mice**. *Genome research* 2005, **15**(5):681-691.
30. Breitling R, Armengaud P, Amtmann A, Herzyk P: **Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments**. *FEBS letters* 2004, **573**(1-3):83-92.
31. Stylianou IM, Affourtit JP, Shockley KR, Wilpan RY, Abdi FA, Bhardwaj S, Rollins J, Churchill GA, Paigen B: **Applying gene expression, proteomics and single-nucleotide polymorphism analysis for complex trait gene identification**. *Genetics* 2008, **178**(3):1795-1805.
32. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE: **Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks**. *Nat Genet* 2008, **40**(7):854-861.
33. Vera G, Jansen RC, Suppi RL: **R/parallel--speeding up bioinformatics analysis with R**. *BMC bioinformatics* 2008, **9**:390.
34. Alberts R, Terpstra P, Bystrykh LV, de Haan G, Jansen RC: **A statistical multiprobe model for analyzing *cis* and *trans* genes in genetical genomics experiments with short-oligonucleotide arrays**. *Genetics* 2005, **171**(3):1437-1439.
35. Broman KW, Wu H, Sen S, Churchill GA: **R/qtl: QTL mapping in experimental crosses**. *Bioinformatics (Oxford, England)* 2003, **19**(7):889-890.
36. Dunning MJ, Smith ML, Ritchie ME, Tavare S: **beadarray: R classes and methods for Illumina bead-based data**. *Bioinformatics (Oxford, England)* 2007, **23**(16):2183-2184.
37. Dunning MJ, Barbosa-Morais NL, Lynch AG, Tavare S, Ritchie ME: **Statistical issues in the analysis of Illumina data**. *BMC bioinformatics* 2008, **9**:85.
38. Verdugo RA, Medrano JF: **Comparison of gene coverage of mouse oligonucleotide microarray platforms**. *BMC genomics* 2006, **7**:58.
39. Alberts R, Vera G, Jansen RC: **affyGG: computational protocols for genetical genomics with Affymetrix arrays**. *Bioinformatics (Oxford, England)* 2008, **24**(3):433-434.
40. Alberts R, Terpstra P, Hardonk M, Bystrykh LV, de Haan G, Breitling R, Nap JP, Jansen RC: **A verification protocol for the probe sequences of Affymetrix genome arrays reveals high probe accuracy for studies in mouse, human and rat**. *BMC bioinformatics* 2007, **8**:132.
41. Jansen R: **Quantitative trait loci in inbred lines**. In: *Handbook of statistical Genetics.* Edited by Balding D, Bishop M, Cannings C, vol. 1, 3rd edn:

Wiley; 2007: 616-617.
42. Aylor DL, Zeng ZB: **From classical genetics to quantitative genetics to systems biology: modeling epistasis**. *PLoS genetics* 2008, **4**(3):e1000029.
43. Broman KW, Sen S, Owens SE, Manichaikul A, Southard-Smith EM, Churchill GA: **The X chromosome in quantitative trait locus mapping**. *Genetics* 2006, **174**(4):2151-2158.
44. Alberts R, Terpstra P, Li Y, Breitling R, Nap JP, Jansen RC: **Sequence polymorphisms cause many false cis eQTLs**. *PLoS ONE* 2007, **2**(7):e622.

Chapter 4
Genetical genomics: Spotlight on QTL hotspots

QTL hotspots are regions of the genome that harbor genetic variation that seems to affect the expression of a large to very large number of genes (sometimes thousands). QTL hotspots could therefore reveal the presence of major biological regulators. However, striking discrepancy among the results reported in scientific literature has caused concern over their significance. In this chapter, we review published reports about eQTL hotspots and we propose a permutation strategy that allows us to discard numerous hotspots as statistical artifacts.

Originally published as:
Breitling R, Li Y, **Tesson BM**, Fu J, Wu C, Wiltshire T, Gerrits A, Bystrykh LV, de Haan G, Su AI, Jansen RC.
Genetical genomics: spotlight on QTL hotspots.
PLoS Genetics 2008 Oct;4(10):e1000232.

4.1 Introduction

Genetical genomics aims at identifying quantitative trait loci (QTL) 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 [1], 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 and 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 coincide with the position of the gene and are presumably acting *in cis*, for example 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 power to detect them and, consequently, an inability to reliably delimit hotspots [2]. 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 [3]. Other studies have reported up to about 30 hotspots, far less than the number of significant local eQTLs (**Table 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 [4].

Recently, Wu et al. [5] 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 1600 candidate hotspots, each with a minimum hotspot size of 50 target genes. Furthermore, they demonstrate 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 (Wu & Su, unpub-

lished) shows that this results in a false discovery rate of 83%, largely resulting from multiple testing across 157,000 SNPs and 6601 probe sets. This relatively permissive threshold was chosen since 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 [6]. 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 [2, 7-9]. There are many non-genetic mechanisms which 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 [5] permuted the observed eQTLs among genes (**Figure 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 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 1A**, maintaining the correlation of the expression traits while destroying any genetic association [2, 10]. 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 [5] 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 1**), especially where no appropriate permutation tests were applied to

determine the statistical significance of hotspots [2].

Of course, this does not imply that all hotspots are necessarily false positives. As described above, about 5% of the co-mapping clusters in [5] 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 enrichment of functional associations (p-value < 0.16, FDR = 67%; Wu & Su, unpublished). 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 comprising a hotspot [5].

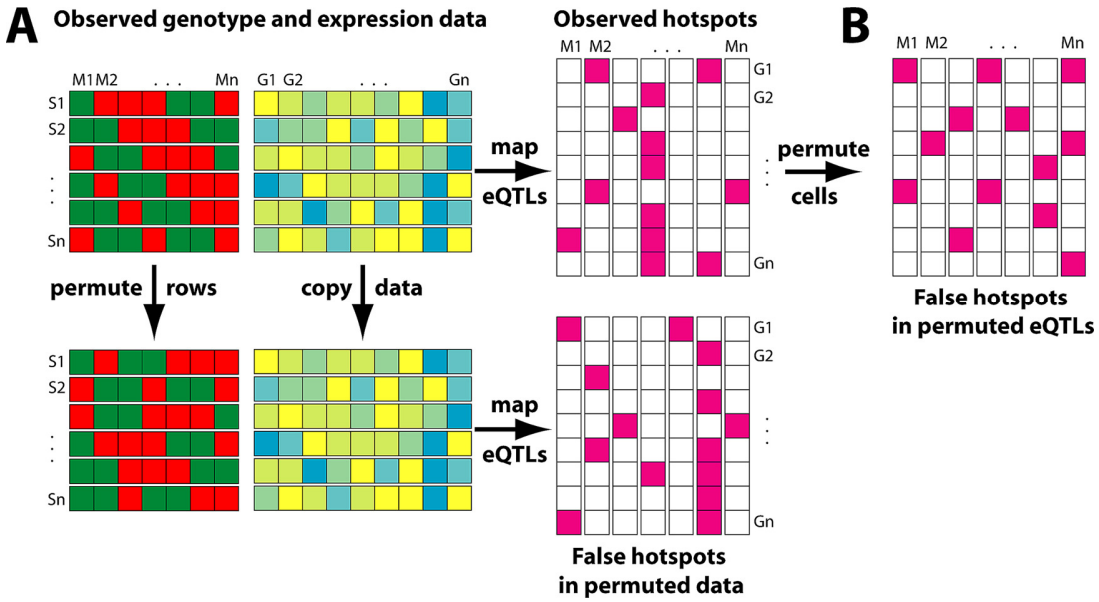


Figure 1 - Alternative Permutation Strategies for Determining the Significance of eQTL Hotspots in Linkage and Association Studies. (A) The top panel shows the original data. The genotype matrix contains information about the genotype of each strain (S1…Sn) at each marker position along the genome (M1…Mn). For each strain, the expression of genes G1…Gn 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 [5], 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.

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 [11, 12]. An example is the recent work of Zhu et al. [12]: 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 9 out of 13 hotspots (69%) originally reported in [13]. 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 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 [14-17]. Experimental evidence for phenotypic buffering of protein coding polymorphisms is well established [18, 19].

In fact, it has been shown that phenotypic buffering is a general property of complex gene-regulatory networks [20]. 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 [21, 22]. Future studies in genetical genomics should aim at further elucidating the striking rarity of eQTL hotspots.

Table 1. eQTL hotspots reported in selected genetical genomics studies*.

Paper	Organism	Population size	Number of local eQTLs	Number of distant eQTLs	Threshold for eQTLs	Number of hotspots
Brem et al., Science, 2002[23]	yeast	40	185	385	$p<5x10^{-5}$	8
Yvert et al., Nat Genet, 2003[13]	yeast	86	578	1716	$p<3.4x10^{-5}$	13
Schadt et al., Nature, 2003[1]	mouse	111	1022	1985	LOD>4.3	7
Kirst et al., Plant Physiol, 2004[24]	eucalyptus	91	1	8	experiment-wise α=0.10	2
Monks et al., AJHG, 2004[25]	human	15 CEPH families (167)	13	20	$p<5x10^{-5}$	0
Morley et al., Nature, 2004[26]	human	14 CEPH families	29	118	$p<4.3x10^{-7}$	2
Cheung et al., Nature, 2005[27]	human	57	65	0	p<0.001	0
Stranger et al., PLoS Genet, 2005[28]	human	60	10–40	3	corrected p-value 0.05	0
Chesler et al., Nat Genet, 2005[29]	mouse	35	83	5	FDR=0.05	7
Bystrykh et al., Nat Genet, 2005[30]	mouse	30	478	136	genome-wide p<0.005	"multiple"
Hubner et al., Nat Genet, 2005[31]	rat	259	622	1211	p<0.05	2
Mehrabian et al., Nat Genet, 2005[32]	mouse	111	20107 total		LOD>2	1
DeCook et al., Genetics, 2006[33]	*Arabidopsis*	30	3525 total		FDR=2.3%	5
Lan et al., PLoS Genet, 2006[34]	mouse	60	723	5293	LOD>3.4	15
Wang et al., PLoS Genet, 2006[35]	mouse	312	2118	4556	$P<5x10^{-5}$	7
Li et al., PLoS Genetics, 2006[36]	*C. elegans*	80	414	308	p<0.001 FDR=0.04	1
Keurentjes et al., PNAS, 2007[4]	*Arabidopsis*	160	1875	1958	FDR=0.05	~29
McClurg et al., Genetics, 2007[37]	mouse	32	N.A.	N.A.	N.A.	25
Emilsson et al., Nature, 2008[3]	human	470	1970	52	FDR=0.05	0
Schadt et al., PLoS Biol, 2008[38]	human	427	3210	242	$p<1.6x10^{-12}$	23
Ghazalpour et al., PLoS Genet, 2008[39]	mouse	110	471	701	FDR=0.1	4
Wu et al., PLoS Genet, 2008[5]	mouse	28	600	885,840 (Wu & Su, unpub-lished)	p<0.003	1659

*** 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.**

4.3 References

1. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G *et al*: **Genetics of gene expression surveyed in maize, mouse and man**. *Nature* 2003, **422**(6929):297-302.
2. de Koning DJ, Haley CS: **Genetical genomics in humans and model organisms**. *Trends Genet* 2005, **21**(7):377-381.
3. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S *et al*: **Genetics of gene expression and its effect on disease**. *Nature* 2008, **452**(7186):423-428.
4. Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, Van den Ackerveken G, Snoek LB, Peeters AJ, Vreugdenhil D, Koornneef M, Jansen RC: **Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci**. *PNAS USA* 2007, **104**(5):1708-1713.
5. Wu C, Delano DL, Mitro N, Su SV, Janes J, McClurg P, Batalov S, Welch GL, Zhang J, Orth AP *et al*: **Gene set enrichment in eQTL data identifies novel annotations and pathway regulators**. *PLoS Genet* 2008, **4**(5):e1000070.
6. Wessel J, Zapala MA, Schork NJ: **Accommodating pathway information in expression quantitative trait locus analysis**. *Genomics* 2007, **90**(1):132-142.
7. Peng J, Wang P, Tang H: **Controlling for false positive findings of *trans*-hubs in expression quantitative trait loci mapping**. *BMC Proc* 2007, **1 Suppl 1**:S157.
8. Perez-Enciso M: **In silico study of transcriptome genetic variation in outbred populations**. *Genetics* 2004, **166**(1):547-554.
9. Wang S, Zheng T, Wang Y: **Transcription activity hot spot, is it real or an artifact?** *BMC Proc* 2007, **1 Suppl 1**:S94.
10. Churchill GA, Doerge RW: **Naive application of permutation testing leads to inflated type I error rates**. *Genetics* 2008, **178**(1):609-610.
11. Stylianou IM, Affourtit JP, Shockley KR, Wilpan RY, Abdi FA, Bhardwaj S, Rollins J, Churchill GA, Paigen B: **Applying gene expression, proteomics and single-nucleotide polymorphism analysis for complex trait gene identification**. *Genetics* 2008, **178**(3):1795-1805.
12. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE: **Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks**. *Nat Genet* 2008.
13. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L: ***Trans*-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors**. *Nat Genet* 2003, **35**(1):57-64.
14. Le Rouzic A, Carlborg O: **Evolutionary potential of hidden genetic**

variation. *Trends Ecol Evol* 2008, **23**(1):33-37.
15. Gibson G, Wagner G: **Canalization in evolutionary genetics: a stabilizing theory?** *Bioessays* 2000, **22**(4):372-380.
16. Gibson G, Dworkin I: **Uncovering cryptic genetic variation**. *Nat Rev Genet* 2004, **5**(9):681-690.
17. Carlborg O, Haley CS: **Epistasis: too often neglected in complex trait studies?** *Nat Rev Genet* 2004, **5**(8):618-625.
18. Queitsch C, Sangster TA, Lindquist S: **Hsp90 as a capacitor of phenotypic variation**. *Nature* 2002, **417**(6889):618-624.
19. Rutherford SL, Lindquist S: **Hsp90 as a capacitor for morphological evolution**. *Nature* 1998, **396**(6709):336-342.
20. Bergman A, Siegal ML: **Evolutionary capacitance as a general feature of complex gene networks**. *Nature* 2003, **424**(6948):549-552.
21. Iyengar SK, Elston RC: **The genetic basis of complex traits: rare variants or "common gene, common disease"?** *Methods Mol Biol* 2007, **376**:71-84.
22. Bodmer W, Bonilla C: **Common and rare variants in multifactorial susceptibility to common diseases**. *Nat Genet* 2008, **40**:695-701.
23. Brem RB, Yvert G, Clinton R, Kruglyak L: **Genetic dissection of transcriptional regulation in budding yeast**. *Science* 2002, **296**(5568):752-755.
24. Kirst M, Myburg AA, De Leon JP, Kirst ME, Scott J, Sederoff R: **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* 2004, **135**(4):2368-2378.
25. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE: **Genetic inheritance of gene expression in human cell lines**. *AmJHumGenet* 2004, **75**(6):1094-1105.
26. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG: **Genetic analysis of genome-wide variation in human gene expression**. *Nature* 2004, **430**(7001):743-747.
27. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT: **Mapping determinants of human gene expression by regional and genome-wide association**. *Nature* 2005, **437**(7063):1365-1369.
28. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S *et al*: **Genome-wide associations of gene expression variation in humans**. *PLoS Genet* 2005, **1**(6):e78.
29. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA *et al*: **Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function**. *NatGenet* 2005, **37**(3):233-242.
30. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su

AI, Vellenga E, Wang J, Manly KF *et al*: **Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'**. *NatGenet* 2005, **37**(3):225-232.

31. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V *et al*: **Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease**. *Nat Genet* 2005, **37**(3):243-253.
32. Mehrabian M, Allayee H, Stockton J, Lum PY, Drake TA, Castellani LW, Suh M, Armour C, Edwards S, Lamb J *et al*: **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* 2005, **37**(11):1224-1233.
33. DeCook R, Lall S, Nettleton D, Howell SH: **Genetic regulation of gene expression during shoot development in Arabidopsis**. *Genetics* 2006, **172**(2):1155-1164.
34. Lan H, Chen M, Flowers JB, Yandell BS, Stapleton DS, Mata CM, Mui ET, Flowers MT, Schueler KL, Manly KF *et al*: **Combined expression trait correlations and expression quantitative trait locus mapping**. *PLoS Genet* 2006, **2**(1):e6.
35. Wang S, Yehya N, Schadt EE, Wang H, Drake TA, Lusis AJ: **Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity**. *PLoS Genet* 2006, **2**(2):e15.
36. Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JA, Hazendonk E, Prins P, Plasterk RH, Jansen RC *et al*: **Mapping determinants of gene expression plasticity by genetical genomics in C. elegans**. *PLoSGenet* 2006, **2**(12):e222.
37. McClurg P, Janes J, Wu C, Delano DL, Walker JR, Batalov S, Takahashi JS, Shimomura K, Kohsaka A, Bass J *et al*: **Genomewide association analysis in diverse inbred mice: power and population structure**. *Genetics* 2007, **176**(1):675-683.
38. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C *et al*: **Mapping the genetic architecture of gene expression in human liver**. *PLoS Biol* 2008, **6**(5):e107.
39. Ghazalpour A, Doss S, Kang H, Farber C, Wen PZ, Brozell A, Castellanos R, Eskin E, Smith DJ, Drake TA *et al*: **High-resolution mapping of gene expression using association in an outbred mouse stock**. *PLoS Genet* 2008, **4**(8):e1000149.

Chapter 5
DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules

Large microarray datasets have enabled gene regulation to be studied through coexpression analysis. While numerous methods have been developed for identifying differentially expressed genes between two conditions, the field of differential coexpression analysis is still relatively new. More specifically, there is so far no sensitive and untargeted method to identify gene modules (also known as gene sets or clusters) that are differentially coexpressed between two conditions. Here, sensitive and untargeted means that the method should be able to construct de novo modules by grouping genes based on shared, but subtle, differential correlation patterns. We present DiffCoEx, a novel method for identifying correlation pattern changes, which builds on the commonly used Weighted Gene Coexpression Network Analysis (WGCNA) framework for coexpression analysis. We demonstrate its usefulness by identifying biologically relevant, differentially coexpressed modules in a rat cancer dataset. DiffCoEx is a simple and sensitive method to identify gene coexpression differences between multiple conditions.

Originally published as:
Tesson BM, Breitling R, Jansen RC.
DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules.
BMC Bioinformatics. 2010 Oct 6;11:497

5.1 Background

There are two major classes of approach to the analysis of gene expression data collected in microarray studies: either one can identify genes that are differentially expressed in different conditions, or the patterns of correlated gene expression (coexpression). Coexpression analysis identifies sets of genes that are expressed in a coordinated fashion, i.e. respond in a similar fashion to the controlled or uncontrolled perturbation present in the experiment. Such coexpression is considered as evidence for possible co-regulation and for membership to common biological processes under the principle of guilt–by-association [1]. When comparing the transcriptome between two conditions, it is a natural step to identify differential coexpression to get an even more informative picture of the dynamic changes in the gene regulatory networks. Changes in the differential coexpression structure of the genes are, for example, a group of genes strongly correlated in one condition but not in the other, or one module correlating to another module in one condition, whereas they are no longer correlated in the other condition. Differential coexpression may indicate rewiring of transcriptional networks in response to disease or adaptation to different environments.

Differential coexpression has been reported in diverse organisms and across various conditions. For example, Fuller et al. [2] reported a differentially coexpressed module in obese mice compared to lean mice; Van Nas et al. [3] found gender-specific coexpression modules; Oldham et al. [4] identified gene modules that were differentially coexpressed between humans and chimpanzees; and Southworth et al. [5] found that aging in mice was associated with a general decrease in coexpression. Differential coexpression patterns associated with diseases have been an important focus of research, see review by De la Fuente et al. [6].

Differential coexpression methods can be divided into two categories that serve distinct purposes: on the one hand, targeted approaches study gene modules that are defined *a priori*, while, on the other hand, untargeted approaches aim at grouping genes into modules on the basis of their differential coexpression status.

A suitable untargeted method for differential coexpression analysis should satisfy the following criteria:

(i) Sensitively detect groups of genes in which the correlation of gene pairs within the group is significantly different between conditions.

(ii) Sensitively detect changes in correlations between two groups of genes even when the within-group correlation is conserved across conditions.

(iii) Allow for simple comparison of more than two conditions.

Criteria (i) and (ii) are illustrated in **Figure 1**, which schematically depicts biological scenarios that can give rise to differential coexpression.

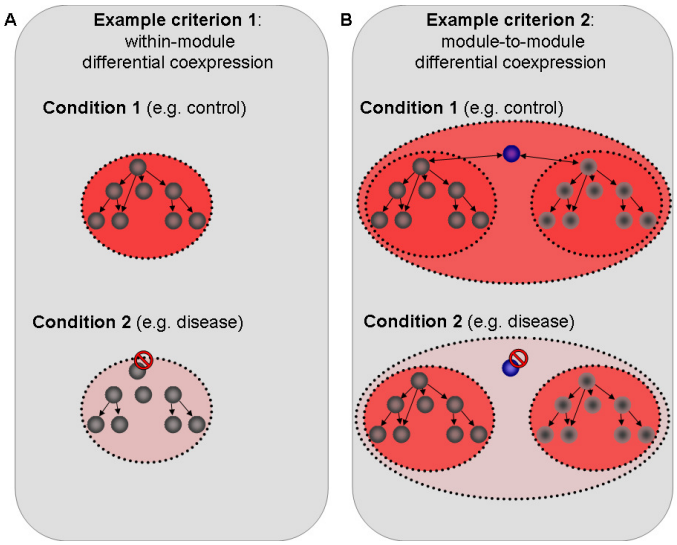


Figure 3 - Illustration of differential coexpression scenarios.

Panel A: A gene network is in a coexpressed state in condition 1 as shown by the red background. In condition 2 an important regulator of that network is now inactive and the module is no longer coexpressed. This scenario is an example of the differential coexpression type described by criterion (i). **Panel B:** Two pathways are coordinated in condition 1 via an important hub gene (shown in blue) whose inactivity in condition 2 means the two pathways are no longer coexpressed. This exemplifies the module-to-module differential coexpression described by criterion (ii).

Multiple methods have been proposed to identify such large-scale correlation patterns [5, 7-12]. However, this early work provided only partial solutions to the problem of differential coexpression since, with one recent exception [5], none of the proposed methods were entirely untargeted. Instead, existing methods can be divided into two categories: targeted and "semi-targeted" approaches. In targeted approaches, pre-defined modules are surveyed for correlation changes between two conditions. For example, Choi et al. [9] proposed a method that focuses on the analysis of modules based on known gene annotations, such as GO categories, and tests the significance of the coexpression changes using a statistical measure known as dispersion. This has the advantage of not requiring the gene sets to be highly correlated in one of the two conditions. However, this method is targeted in that it relies on the study of known functional gene sets and is not able to identify novel, non-annotated modules or modules that would only partially match annotated categories. "Semi-targeted" approaches use classical coexpression methods in one of the conditions to define modules and study whether these modules are also coexpressed in the second condition. DCA (differential clustering analysis) [10] is an example of a method using one of the two conditions as reference, meaning the clusters under consideration are obtained from one condition and then studied in the other condition. In order to avoid bias towards one of the conditions, Ihmels et al.

suggested doing a reciprocal analysis, switching the reference and target conditions, while Southworth et al. used a third dataset as reference [5]. A drawback of such "semi-targeted" methods is that the analysis will only focus on groups of genes that emerge as clusters in at least one of the conditions, and will therefore potentially miss more subtle cases. As an example, a weak but significant condition-dependent correlation structure between a group of genes that otherwise belong to distinct, strongly coexpressed and conserved clusters would not be detected by this approach. A first attempt at an untargeted approach was introduced by Southworth et al. [5], who proposed applying hierarchical clustering using the difference in pairwise correlations between both conditions as a similarity metric for two genes. This approach is therefore suited to identifying groups in which the within-group correlation changes (first criterion), but it cannot be applied to the detection of module-to-module correlation differences (second criterion). The field of differential coexpression analysis would therefore benefit from a new, truly untargeted and sensitive method for identifying differentially correlated modules that would satisfy all three criteria.

Here we present a solution to this problem in the form of the DiffCoEx approach for untargeted differential coexpression analysis: a method which applies the powerful tools of Weighted Gene Coexpression Network Analysis (WGCNA) to differential network analysis. We first describe the five steps involved in DiffCoEx and then, to illustrate the method's effectiveness, we present the results of an analysis performed on a publicly available dataset generated by Stemmer et al. [13].

5.2 Algorithm

Our method builds on WGCNA [14, 15], which is a framework for coexpression analysis. Identification of coexpression modules with WGCNA follows three steps: first an adjacency matrix is defined between all the genes under consideration based on pair-wise correlations. Then the generalized topological overlap measure [16] is computed from the adjacency matrix and converted into a dissimilarity measure. Finally, using this dissimilarity measure, hierarchical clustering is applied, followed by tree cutting using either a static or a dynamic height cut. The resulting clusters form modules of genes in which all members are strongly inter-correlated.

The principle of DiffCoEx is to apply WGCNA to an adjacency matrix representing the correlation changes between conditions. DiffCoEx clusters genes using a novel dissimilarity measure computed from the topological overlap [16] of the correlation changes between conditions. Intuitively, the method groups two genes together when their correlations to the same sets of genes change between the different conditions. The complete process of our differential coexpression analysis comprises five steps, described below. The notation X designates a square matrix

with the dimension of the number of genes considered and x_{ij} is used to define the element of X at row i and column j.

Step 1: Build adjacency matrix $C^{[k]}$ within each condition k as the correlation for all pair of genes (i,j):

$$C^{[k]}: \quad c_{ij}^{[k]} = \mathrm{cor}(gene_i, gene_j)$$

In this step, different correlation measures can be used, such as the Pearson or Spearman coefficient.

Step 2: Compute matrix of adjacency difference:

$$D: \quad d_{ij} = \left(\sqrt{\frac{1}{2} \left| \mathrm{sign}(c_{ij}^{[1]}) * (c_{ij}^{[1]})^2 - \mathrm{sign}(c_{ij}^{[2]}) * (c_{ij}^{[2]})^2 \right|} \right)^{\beta}$$

In this matrix, high values of d_{ij} indicate that the coexpression status of $gene_i$ and $gene_j$ changes significantly between the two conditions. The correlation change is quantified as the difference between signed squared correlation coefficients so that changes in correlation which are identical in terms of explained variance (r^2) are given the same weight. This adjacency matrix is defined such that it only takes values between 0 and 1. The soft threshold parameter β is taken as a positive integer and is used to transform the correlation values so that the weight of large correlation differences is emphasized compared to lower, less meaningful, differences. β should be regarded as a tuning parameter, and in practice it is advisable to try different values of β. In WGCNA, it is recommended to choose β so that the resulting coexpression network follows an approximate scale-free topology [14]. However the "scale-free" topology nature of biological networks has been disputed [17], and another way is to consider the soft threshold parameter as a stringency parameter: using high values of β means putting less emphasis on smaller changes in correlation, and therefore being more statistically stringent. Accordingly, since larger sample sizes come with higher statistical significance of small correlation changes, smaller values of the soft threshold can be used as the sample size increases. In practice, we view the soft threshold parameter as a tuning parameter, and we always check the significance of the result afterwards, both statistically and using biological criteria relevant in each specific study.

Step 3: Derive the Topological Overlap [16] based dissimilarity matrix T from the adjacency change matrix D.

$$T: \quad t_{ij} = 1 - \frac{\sum_k \left(d_{ik}\, d_{kj} \right) + d_{ij}}{\min\left(\sum_k d_{ik}, \sum_k d_{jk} \right) + 1 - d_{ij}}$$

The use of the topological overlap measure to construct a dissimilarity metrics allows the identification of genes that share the same neighbors in the graph formed by the differential correlation network as defined by the adjacency matrix created in Step 2. Intuitively, a low value of t_{ij} (high similarity) means that $gene_i$ and $gene_j$ both have significant correlation changes with the same large group of genes. This group of genes constitutes their "topological overlap" in the differential correlation network and may, or may not, include $gene_i$ and $gene_j$. This property allows DiffCoEx to satisfy both criteria (i) and (ii) as stated earlier. On the one hand, if $gene_i$ and $gene_j$ are part of a module of genes coexpressed in only one condition (criterion (i), illustrated in **Figure 1A**), then the topological overlap between $gene_i$ and $gene_j$ in the difference network consists of all the genes within that module. On the other hand, if $gene_i$ and $gene_j$ are equally inter-correlated in both conditions but correlate with the genes in a distinct module in only one condition (criterion (ii), illustrated in **Figure 1B**), then the topological overlap between $gene_i$ and $gene_j$ in the difference network consists of the genes in that other module. In both cases $gene_i$ and $gene_j$ will therefore be grouped together: in the first case forming a differentially correlated module, and in the second case forming a module with differential module-to-module correlation with another group of genes.
We note that since the adjacency matrix takes values between 0 and 1, the dissimilarity matrix computed here also takes values between 0 and 1, as shown in [14].

Step 4: The dissimilarity matrix T is used as input for clustering and modules are identified.
The clustering can be done using standard hierarchical clustering with average linkage, followed by module extraction from the resulting dendrogram, either using a fixed cut height or with more elaborate algorithms such as the dynamicTreeCut [18]. Alternative clustering techniques, such as Partitioning Around Medoids (PAM) [19], may be used in this step.

Step 5: Assess the statistical significance of coexpression changes.
This is necessary because DiffCoEx uses user-defined parameters: the soft threshold β used to transform the adjacency matrix in Step 2 and the clustering parameters in Step 4 (tree cutting settings, for example). Unsuitable settings may lead to the detection of clusters with non-significant differential coexpression.

The statistical significance of differential coexpression can be assessed using a measure of the module-wise correlation changes such as the dispersion statistic [9], the t-statistic [12], or the average absolute correlation. Permutations or simulations of the data can be used to generate a null distribution of those statistics by providing estimates of the extent of differential correlation that can be expected to occur by chance. An example of implementing a permutation procedure to assess the significance of differential coexpression using the dispersion statistics is presented in **Additional file 1**.

Variants

Extending the DiffCoEx method to multiple conditions

This method can easily be extended to the study of differential coexpression over more than two conditions. The only required change is in Step 2, where the matrix of adjacency differences should be replaced with the following: supposing we have calculated $C^{[1]},\ldots,C^{[k]},\ldots,C^{[n]}$ the correlation matrices for gene pairs in each of the n different conditions:

$$D:\ d_{ij} = \left(\sqrt{\frac{1}{n-1}\sum_k \frac{\left|\text{sign}(c_{ij}^{[k]})*(c_{ij}^{[k]})^2 - c_{ij}^{[0]}\right|}{2}} \right)^{\beta}$$

$$\text{where} \qquad c_{ij}^{[0]} = \frac{1}{n}\sum_k \left(\text{sign}(c_{ij}^{[k]}) * (c_{ij}^{[k]})^2\right)$$

For two conditions, one can verify that this formulation is equivalent to that proposed earlier in Step 2.

A less sensitive variant to detect more striking patterns

If one is interested in picking up only coexpression changes that affect genes forming highly coexpressed modules in at least one of the conditions, the formula in Step 2 can be adapted so that the method uses the difference between the two transformed correlation matrices (with the soft threshold parameter β) as shown below:

$$D \ : \ d_{ij} = \frac{1}{2}\left|\text{sign}(c_{ij}^{[1]}) * (c_{ij}^{[1]})^{\beta} - \text{sign}(c_{ij}^{[2]}) * (c_{ij}^{[2]})^{\beta}\right|$$

This will make the method less sensitive to subtle coexpression changes, but may help in extracting more strikingly differentially coexpressed modules.

Variant without the topological overlap

As with WGCNA, the use of a topological overlap-based metrics makes the approach very sensitive, since it considers the correlation changes to all other genes to determine the similarity between two genes. The method can be simplified by replacing the dissimilarity matrix T of Step 4 by a dissimilarity measure derived directly from the adjacency matrix D:

$$T_{alt} = 1 - D$$

This will make DiffCoEx focus only on within-module differential coexpression (criteria (i)) and not on module-to-module differential coexpression (criteria (ii)). This variant is computationally more efficient since the topological overlap computation is omitted.

5.3 Results

We present here the results of our method as used on a previously published dataset. We identify modules of genes that are differentially coexpressed and, by using gene set enrichment analysis, we provide evidence for their biological relevance.

5.3.1 Dataset

Our dataset (Gene Expression Omnibus GEO GSE5923) contains Affymetrix gene expression profiles of renal cortex outer medulla in wild-type- and Eker rats treated with carcinogens. The dataset is a time course as the rats were treated with Aristolochic Acid (AA) or Ochratoxin A (OTA), respectively, for 1, 3, 7 or 14 days. In total, the dataset consists of 84 arrays measuring 15,923 probe sets. Details about the experimental settings are available in the original paper [13].

Eker rats are predisposed to renal tumor because they are heterozygous for a loss-of-function mutation in the tuberous sclerosis 2 (*Tsc2*) tumor suppressor gene. Stemmer et al. [13] compared the transcriptional responses of the rats to the carcinogens and found that the expression levels of genes belonging to a number of cancer-related pathways were affected differently in the mutant compared to the wild-type rats. In our re-analysis of the data, we switched the focus from differential expression to differential coexpression in an attempt to identify functional modules responding to carcinogen treatment with a different coexpression signature in mutant Eker rats compared to wild type rats.

5.3.2 Analysis

We applied the DiffCoEx method to the quantile normalized data [20]. A duplicate set of 12 controls present only for Eker rats was discarded in order to have a symmetric experimental setting among wild-type- and Eker rats. We used the Spearman rank correlation in order to reduce sensitivity to outliers, and the hierarchical clustering and module assignment was performed using dynamicTreeCut [18]. The detailed algorithm and R code used in this analysis are given in **Additional file 1**.

5.3.3 Findings

The results of the analysis are summarized in **Figure 2A.** We identified a total of 8 differentially coexpressed modules comprising a total of close to 1800 genes (1887 probe sets, 1796 unique genes). The modules were given color names as indicated in **Figure 2A**. Four of these modules (totaling 1361 genes) were significantly more highly correlated in the mutant Eker rats than in the wild-type rats, while only the red module (36 genes) and, to a lesser extent, the green module (116 genes) follow the opposite pattern. This striking asymmetry might reflect the greater fragility of

the Eker rats to carcinogens: in Eker rats, treatment with carcinogens leads to much more coordinated perturbation of the transcriptome than in wild-type rats.

The cases of the black, orange and green modules illustrate an interesting characteristic of DiffCoEx: the method is able to identify module-to-module correlation changes. Interestingly, the black module is not differentially correlated in the wild-type rats compared to the Eker rats. Instead, what qualifies the black module as a differentially coexpressed module is its very significant drop in correlation with the genes in the blue and purple modules in the wild-type rats compared to the Eker mutants (see **Figure 2A**). Similar patterns can be observed for the orange and green modules. This property makes DiffCoEx a sensitive approach for detecting any type of large-scale correlation change.

Following Choi et al. [9], significance of the coexpression differences was assessed by comparing the dispersion index values of each module in the data with the null distribution obtained from permuted (scaled) data (see **Additional file 1** for details and **Additional file 2: Figure S1** for an overview of the permutation results). In 1000 permutations, none of the blue, brown, purple, red or yellow modules obtained as high a dispersion value as that obtained from the non-permuted data, indicating a significance p-value < 0.001. Module-to-module coexpression changes were tested by assessing the significance of the correlation changes between the genes from each possible module pair, using a similar "module-to-module" dispersion measure and generating null distributions from the same permutation approach. **Additional file 2: Figure S1** shows that the coexpression change between the black and blue modules, for example, is highly significant since no permutation yielded as high a dispersion value.

In the next step, the biological significance of the modules was surveyed using gene-set enrichment analysis. We submitted each of the modules to GeneTrail [21] and identified many significantly over-represented GO or KEGG terms among the gene annotations. A subset of some of the most interesting findings is presented in **Table 1**, while complete lists are available as **Additional file 3**. In **Figure 2B,** the expression data for the 13 genes of the yellow module, which were associated with the "pancreatic cancer" KEGG annotation, illustrate what differential coexpression is: a difference in the coordination of the variation of a group of genes between two conditions. In the Eker rats, these cancer genes show coordinated variation, whereas in the wild-type rats this coordination is absent.

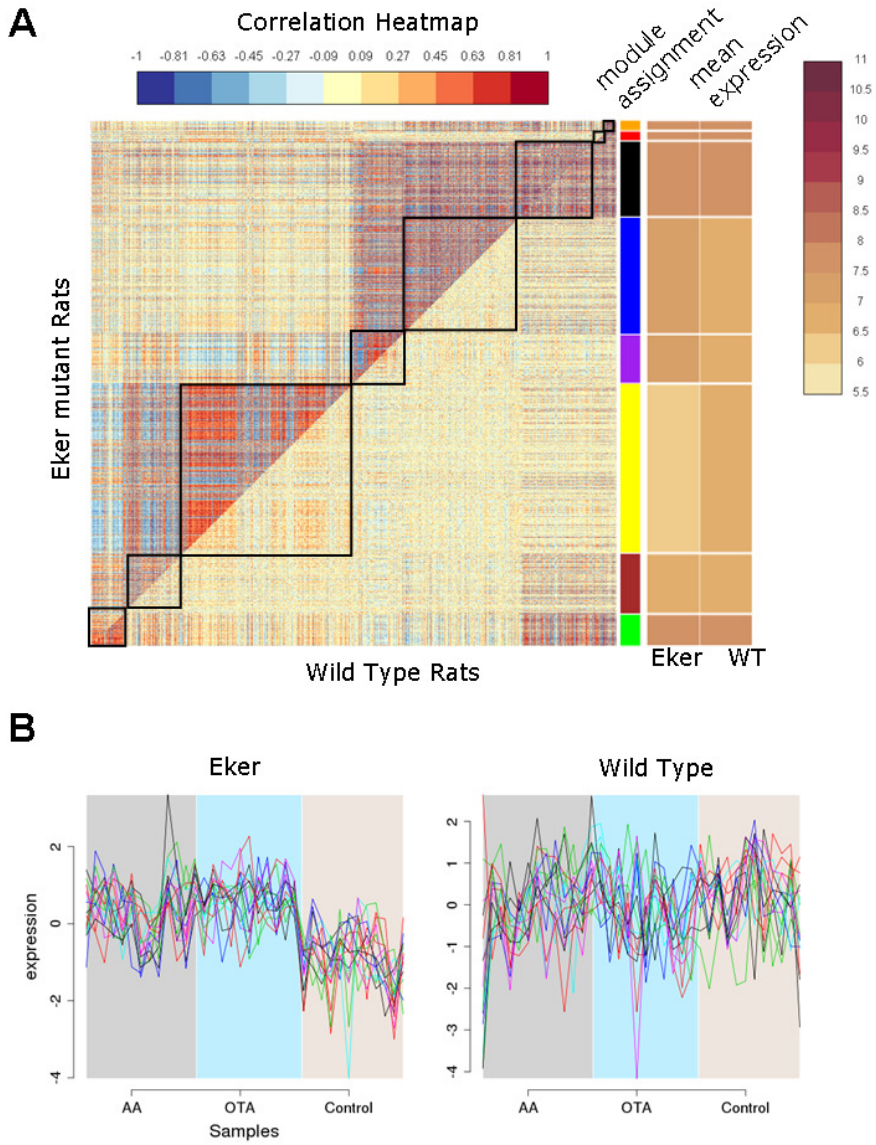


Figure 2 - Differentially coexpressed modules between carcinogen-treated Eker rats and wild-type rats

Panel A: Comparative correlation heat map. The upper diagonal of the main matrix shows a correlation between pairs of genes among the Eker mutant rats (the red color corresponds to positive correlations, blue to negative correlations). The lower diagonal of the heat map shows a correlation between the same gene pairs in the wild-type controls. Modules are identified in the heat map by black squares and on the right side of the heat map by a color bar. The brown bands on the right side indicate the mean expression of the modules in the Eker rats (first column) and the wild-type rats (second column); darker colors indicate higher mean expression levels.

Panel B: Expression variation (scaled) in the Eker mutants (left) and the wild-type rats (right) of the genes in the yellow module which are annotated in KEGG with "pancreatic cancer". In the Eker rats the variation of these genes is tightly correlated, whereas for the wild-type rats it is much more random.

Module	Category	Subcategory	Expected	Observed	Fdr
Black	KEGG	Metabolism of xenobiotics by cytochrome P450	1.367	12	<0.001
	KEGG	Metabolic pathways	22.494	40	<0.001
	GO	Glutathione transferase activity	0.364	9	<0.001
Blue	KEGG	Lysosome	3.373	12	0.008
	KEGG	Metabolic pathways	31.541	48	0.026
	GO	Mitochondrion	35.764	67	<0.001
Brown	GO	Intracellular transport	8.481	22	0.038
Green	GO	Mitochondrion	10.234	26	0.003
	GO	Oxidation reduction	4.015	15	0.003
Orange	GO	Xenobiotic metabolic process	0.079	5	<0.001
Purple	No significant enrichment				
Red	KEGG	Endometrial cancer	0.201	3	0.015
Yellow	KEGG	Pancreatic cancer	3.344	14	<0.001
	KEGG	Renal cell carcinoma	3.702	10	0.043
	KEGG	Pathways in cancer	14.75	27	0.022
	GO	Protein localization	33.676	64	<0.001
	GO	Melanosome	2.995	11	0.009
	GO	Cell projection	33.886	59	0.002
	GO	Small GTPase mediated signal transduction	14.342	31	0.003

Table 1 - Annotations enriched in differentially coexpressed modules. Selected annotations enriched among the genes of each differentially coexpressed modules and associated false discovery rates (fdr). The over-representation analysis was conducted using GeneTrail. The complete results are available in Additional file 1. Interestingly, the black module was enriched for genes involved in "response to xenobiotics", while the blue module contained many genes associated with "metabolic processes". Finally, the yellow module was strongly enriched for genes known to be involved in cancer pathogenesis.

5.3.4 Implementation

This analysis was carried out using the R statistical package with the WGCNA [15] library, on a Linux computer with 128 GB physical memory. Large memory (around 10 GB) is required to compute correlation matrices for over 10,000 genes. For module definition, hierarchical clustering was combined with dynamicTreeCut [18] using a minimum size of 20 genes. Details of the process and code can be found in **Additional file 1**.

5.4 Discussion and conclusions

The method we present here has the advantage of comparing two (or more) datasets in a global, unbiased and unsupervised manner. It represents a major improvement over earlier two-way comparisons, in which clustering was first performed in one condition and the coexpression of the genes in the resulting clusters was then assessed in the other condition. Moreover, DiffCoEx is very sensitive because (i) it does not require differentially coexpressed modules to be detected as coherent, coexpressed modules in one of the two conditions; instead, only the difference in coexpression is considered to define the module; and (ii) it can identify all types of large-scale correlation changes, including module-to-module correlation changes. Using a simulation study (see **Additional file 4**), we demonstrate examples of differential coexpression patterns that can be uncovered using DiffCoEx but that were missed by existing approaches.

Differential coexpression provides information that would be missed using classical methods focusing on the identification of differentially expressed genes. For example, as **Figure 2A** shows, many of the differentially coexpressed clusters display few differences between the two conditions in terms of mean overall expression. This indicates that the changes in correlation that we observed cannot be explained by the genes being not expressed, and therefore not correlated in one of the two conditions.

Differential coexpression may be caused by different biological mechanisms. For example, a group of genes may be under the control of a common regulator (e.g. a transcription factor or epigenetic modification) that is active in one condition, but absent in the other condition. In such a case, the correlation structure induced by variation in the common regulator would only be present in the first condition. Another possible interpretation relates to the presence or absence of variation in some factors driving a gene module. To observe correlation of a group of genes responding to a common factor, this factor needs to vary. In the absence of variation of the driving factor, no correlation can be observed, even though the actual biological links that form the network are not altered. It is therefore important to ensure that the perturbations which give rise to variation within each condition are: (i) biologi-

cally relevant (as opposed to batch effects, for example) and (ii) comparable in nature and amplitude.

DiffCoEx provides a simple and efficient approach to study how different sample groups respond to the same perturbations. These perturbations can be either well characterized and controlled, or stochastic and unknown. In our example analysis, on top of random physiological fluctuations present in any dataset, there was a controlled perturbation induced by the time-course treatment with different carcinogens present. Since the carcinogen treatment is a controlled experimental factor, it is possible to use classical methods to study the transcriptomic changes it induces rather than using DiffCoEx. However, a fundamental advantage of using DiffCoEx in such a case is that it requires no model assumptions and is a quick and efficient approach. Differential coexpression approaches are even more useful when the variation among the samples in one condition is caused by uncontrolled factors, whose effects cannot easily be dissected. A typical example would be genetic variation present in a natural population or an experimental cross. DiffCoEx constitutes a valuable tool of broad applicability now that such genetic studies are becoming increasingly important for studying gene regulatory networks [22-24].

5.5 Acknowledgements

This work was supported by a BioRange grant SP1.2.3 from the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI). We thank Jackie Senior for editing this article.

5.6 Additional files

Additional files are available online at:
http://www.biomedcentral.com/1471-2105/11/497

Additional file 1. Step-by-step R analysis for applying DiffCoEx. This file contains the documented R source code used to perform the analysis described in the main text as well as the simulation study described in **Additional file 4**.

Additional file 2. Significance assessment of module-to-module coexpression changes using permutations. This figure summarizes the results of the significance analysis. 1000 permutations of the samples between the two conditions were performed, and for each of the permuted datasets, the dispersion value (a measure of correlation change for groups of genes) was computed for each module, and for every possible module pair. The number of permutations yielding a higher disper-

sion value than that of the original data was recorded and is displayed in this figure. The figure, for example, indicates that the within-module dispersion value for the black module reached a higher value with permuted data than with original data 249 times. The within-module coexpression change was therefore not significant (p = 0.249) for the black module and this is indicated with a light grey shading. Similarly, the figure shows that no permutations reached as high a value as the original data for the purple to black dispersion, meaning that the black module was significantly differentially coexpressed with the purple module, and this is indicated with dark grey shading.

Additional file 3. Differentially coexpressed modules and enrichment analysis results. This Excel file has separate sheets for the gene lists for each of the differentially coexpressed modules and the results of the enrichment analysis conducted using GeneTrail.

Additional file 4. Simulation study showing the sensitivity of DiffCoEx. This file details the result of a simulation study performed to illustrate a scenario in which DiffCoEx will outperform other, less sensitive, methods.

5.7 References

1. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I: **The Transcriptional Program of Sporulation in Budding Yeast**. *Science* 1998, **282**(5389):699-705.
2. Fuller TF, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, Horvath S: **Weighted gene coexpression network analysis strategies applied to mouse weight**. *Mamm Genome* 2007, **18**(6-7):463-472.
3. van Nas A, Guhathakurta D, Wang SS, Yehya N, Horvath S, Zhang B, Ingram-Drake L, Chaudhuri G, Schadt EE, Drake TA *et al*: **Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks**. *Endocrinology* 2009, **150**(3):1235-1249.
4. Oldham MC, Horvath S, Geschwind DH: **Conservation and evolution of gene coexpression networks in human and chimpanzee brains**. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(47):17973-17978.
5. Southworth LK, Owen AB, Kim SK: **Aging mice show a decreasing correlation of gene expression within genetic modules**. *PLoS Genet* 2009, **5**(12):e1000776.
6. de la Fuente A: **From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases**. *Trends Genet* 2010, **26**(7):326-333.
7. Cho SB, Kim J, Kim JH: **Identifying set-wise differential co-expression in gene expression microarray data**. *BMC Bioinformatics* 2009, **10**:109-109.
8. Choi JK, Yu U, Yoo OJ, Kim S: **Differential coexpression analysis using microarray data and its application to human cancer**. *Bioinformatics (Oxford, England)* 2005, **21**(24):4348-4355.
9. Choi Y, Kendziorski C: **Statistical methods for gene set co-expression analysis**. *Bioinformatics* 2009, **25**(21):2780-2786.
10. Ihmels J, Bergmann S, Berman J, Barkai N: **Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program**. *PLoS Genetics* 2005, **1**(3):e39-e39.
11. Lai Y, Wu B, Chen L, Zhao H: **A statistical method for identifying differential gene-gene co-expression patterns**. *Bioinformatics (Oxford, England)* 2004, **20**(17):3146-3155.
12. Watson M: **CoXpress: differential co-expression in gene expression data**. *BMC Bioinformatics* 2006, **7**:509-509.
13. Stemmer K, Ellinger-Ziegelbauer H, Ahr H-J, Dietrich DR: **Carcinogen-specific gene expression profiles in short-term treated Eker and wild-type rats indicative of pathways involved in renal tumorigenesis**. *Cancer Research* 2007, **67**(9):4052-4068.

14. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis**. *Statistical Applications in Genetics and Molecular Biology* 2005, **4**(1):1128-1128.
15. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC Bioinformatics* 2008, **9**(1):559-559.
16. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical Organization of Modularity in Metabolic Networks**. *Science* 2002, **297**(5586):1551-1555.
17. Khanin R, Wit E: **How scale-free are biological networks**. *J Comput Biol* 2006, **13**(3):810-818.
18. Langfelder P, Zhang B, Horvath S: **Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R**. *Bioinformatics (Oxford, England)* 2008, **24**(5):719-720.
19. Kaufman L, Rousseeuw PJ: **Finding groups in data. an introduction to cluster analysis**; 1990.
20. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.
21. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Müller R, Meese E, Lenhof H-P: **GeneTrail--advanced gene set enrichment analysis**. *Nucleic Acids Research* 2007, **35**(Web Server issue):W186-192-W186-192.
22. Schadt EE: **Molecular networks as sensors and drivers of common human diseases**. *Nature* 2009, **461**(7261):218-223.
23. Li Y, Breitling R, Jansen RC: **Generalizing genetical genomics: getting added value from environmental perturbation**. *Trends in Genetics: TIG* 2008, **24**(10):518-524.
24. Jansen RC, Tesson BM, Fu J, Yang Y, McIntyre LM: **Defining gene and QTL networks**. *Current Opinion in Plant Biology* 2009, **12**(2):241-246.

Chapter 6
Defining gene and QTL networks

Current technologies for high-throughput molecular profiling of large numbers of genetically different individuals offer great potential for elucidating the genotype-to-phenotype relationship. Variation in molecular and phenotypic traits can be correlated to DNA sequence variation using the methods of quantitative trait locus (QTL) mapping. In addition, the correlation structure in the molecular and phenotypic traits can be informative for inferring the underlying molecular networks. For this, new methods are emerging to distinguish among causality, reactivity or independence of traits based upon logic involving underlying quantitative trait loci (QTL). These methods are becoming increasingly popular in plant genetic studies as well as in studies on many other organisms.

Originally published as:
Defining gene and QTL networks.
Jansen RC, **Tesson BM**, Fu J, Yang Y, McIntyre LM.
Current Opinion in Plant Biology 2009 Apr;12(2):241-6.

6.1 Introduction

Since the rediscovery of Georg Mendel's pioneering work on pea crosses, segregating populations have been used to explore the underlying genetic architecture. Quantitative traits were deconstructed into additive, dominant and epistatic effects, without consideration for the underlying molecular components. Technological advances in the 1980s made comprehensive genotyping affordable and mapping the rough location of the underlying genetic contribution for a quantitative trait (QTL) became feasible [1, 2]. To date, more than 1200 studies in plants have been published on mapping phenotypic QTL (phQTL). A new wave of technological advances makes it possible to profile segregating populations for thousands of gene expression phenotypes and map expression QTL (eQTL)[3]. New technology can be used for the parallel measurement of the abundance of 1000s of proteins and metabolites to map protein QTL (pQTL) and metabolite QTL (mQTL). Deep sequencing, chromatin and methyl-DNA immunoprecipitation are just a few of the newest technologies that add to the impressive arsenal of tools available for the study of the genetic variation underlying quantitative phenotypes[4, 5]. Mapping phenotypes for thousands of traits, 'genetical genomics' [6, 7], is the first step in attempting to reconstruct gene networks. Methods for network reconstruction can be used within a particular level (intra level analysis, *i.e* transcript data only), to explain the relationship among traits[8] at that level. Alternatively, the focus can be on understanding relationships across levels (inter level analysis, integrating transcript, protein, metabolite and morphological phenotypic data). Prior knowledge from other experiments can also be incorporated to further develop the picture of the network. **Figure 1** illustrates the challenges that can be encountered with real data.

6.2 Causal, reactive or independent?

The examination of pairwise correlation between traits, or principal components summaries of these traits, can lead to the hypothesis of a functional relationship if that correlation is high [8-13]. Incorporating QTL information, allows the inference of a functional relationship if two traits share multiple QTL, something that it is unlikely to happen at random. Going beyond the detected QTL, the correlation between residuals among traits, after accounting for QTL effects, or correlations between traits conditional on other traits, is further evidence for a network connection. To infer directional effects, it is necessary to analyze the correlations among pairs of traits in detail. If trait T_1 maps to a subset of the QTL of trait T_2, then the common QTL can be taken as evidence for their network connection, while the distinct QTL can be used to infer the direction [6, 8]. If traits T_1 and T_2 have common QTL, without QTL that are distinct, then the inference is complicated and further analysis is needed to discriminate pleiotropy from any of the possible order-

ings among traits (**Box 1** highlights approaches from [8, 10, 12]). Although these problems are 'modern' the groundwork for such analyses are evident in the earliest days of quantitative analysis [14].

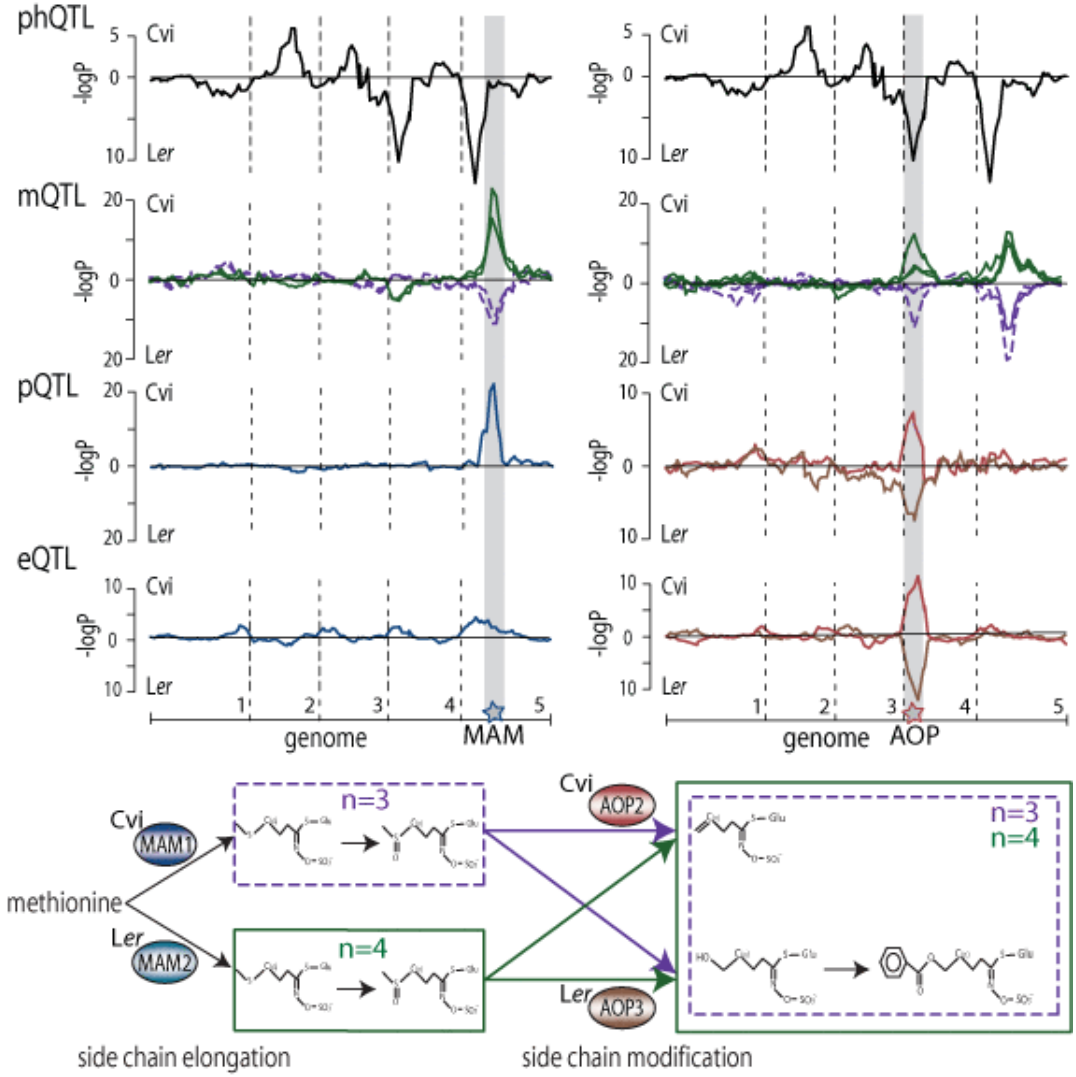


Figure 1 - System-wide QTL analysis for aliphatic glucosinolates.

Data in this example are taken from [41] and demonstrate important network features that reconstruction methods should take into account. The colors in the QTL likelihood graphs (upper panel) and in the pathway (lower panel) correspond. The sign of the QTL effect is shown by plotting the QTL likelihood above the x-axis if the Cvi allele has higher average trait value or below the x-axis if the L*er* allele has higher average trait value. The vertical dashed lines indicate the chromosome borders and the physical gene positions are shown as stars and grey vertical bars.**(continued)**

Figure 1 (continued) - Glucosinolates are important secondary metabolites in plants and are well-known for their toxic effect on insects. The aliphatic glucosinolate biosynthesis pathway is summarized in the lower panel. The MAM genes are involved in side chain elongation process: MAM1 mainly synthesizes C3-glucosinolates (in purple box) and MAM2 mainly synthesizes C4-glucosinolates (in green box). The parents Cvi and L*er* carry different MAM genes. Cvi contains two MAM1 genes and L*er* contains a functional MAM2 in addition to a truncated, non-functional MAM1 gene. The AOP genes are involved in the side-chain modification process: AOP2 and AOP3 generate different types of glucosinolates as described. Both AOP2 and AOP3 are present in L*er* and Cvi. But AOP2 is only expressed in Cvi and AOP3 is only expressed in L*er*. The top panel shows the QTL profiles at different levels (transcripts, proteins, metabolites, disease trait) from a Cvi x L*er* recombinant inbred line population. To clearly demonstrate the QTL effects at different levels and along the pathway, the components are divided into two parts: the left part relates to the MAM1 gene and the metabolites produced by MAM synthesis (MAM2 is not measured); the right part refers to the AOP genes and the metabolites produced by the AOP synthesis. *Cis*-eQTL are detected for AOP2 and AOP3, cis-pQTL for MAM1, AOP2 and AOP3, and mQTL for the various aliphatic glucosinolates. These QTL have the same or opposite sign of QTL effect. To demonstrate whether QTL at molecular levels can propagate to phenotypic level, molecular QTL (eQTL, pQTL and mQTL) are also compared to the QTL of insect susceptibility (phQTL). The disease trait maps to ERECTA, a gene well known for its widespread pleiotropic effect, to AOP, and to third gene, but it does not map to MAM. This can be explained by the fact that the total glucosinolate content maps to AOP only.

6.3 Intra level analysis

In reference organisms, such as Arabidopsis, and in a growing list of plants, the location of the genes producing the transcript or protein studied is known. This added information provides a layer of interpretation for eQTL and pQTL. In Arabidopsis, eQTL and pQTL networks have been defined [15-21]; in barley, eucalyptus and maize eQTL networks have been defined [22-27]. When the eQTL or pQTL co-localize with the gene, this effect may be due to *cis* regulatory effects (Fig 1). The caveat is that the detection of *cis* effects may be an artifact of differential probe hybridization due to sequence polymorphism [28, 29]. If gene expression at a particular locus is regulated by that locus (*cis* effect), and the abundance of the transcript in turn regulates additional loci (*trans* effect) then these expression traits should all map to the same locus. If the number of *trans* loci regulated by a single locus is large, as would be expected from a master regulator, or switch, a *trans* band will be observed at this location. All genes in the QTL are candidate regulators; their partial correlations with the regulated genes can be used to prioritize them [30, 31]. Importantly, genes without *cis*-eQTL can be regulators, manifesting only at the protein level [32]. If the number of transcript or protein traits mapping to a single location exceeds the number expected by chance, then a hotspot has been identified [33, 34].The hotspot can be inferred to represent a possible master regulator or switch. However, as a cautionary note hotspots can be an artifact of improper permutation [34].

Box 1: Advanced causal reasoning

For two traits T_1 and T_2, denote $T_1>T_2$ if T_1 affects T_2; denote $T_1<T_2$ if T_2 affects T_1, and $T_1–T_2$ if there is a correlation but the direction is unknown. If traits T_1 and T_2 have one common QTL, without QTL that are distinct, then the inference of causality is complicated and further analysis needed to discriminate pleiotropy from any of the possible orderings among traits . In this case there are at least three possible models: $QTL>T_1>T_2$; $QTL>T_2>T_1$; $QTL>T_1$ and $QTL>T_2$. If we write the simple regression models $T_2=\alpha_2+\beta_2 QTL+\varepsilon_2$ and $T_1=\alpha_1+\beta_1 QTL+\varepsilon_1$ and if ε_1 and ε_2 are uncorrelated, the QTL may be considered to have pleiotropic effects on the two traits, *i.e.* with no direct link between T_1 and T_2. Alternatively, if there is no evidence for pleiotropy, then the following models can be considered $T_2=\alpha_3+\beta_3 T_1+\varepsilon_3$ and $T_1=\alpha_4+\beta_4 T_2+\varepsilon_4$. The residuals from these models can be used to infer the correct model. If $QTL>T_1>T_2$ is the true relation, then ε_3 will not map to the QTL. In contrast, ε_4 should have a residual signature of the QTL. In cases where ε_3 and ε_4 both map to the QTL, no direction can be given and $T_1–T_2$, while if neither of the two models maps to the QTL, the results are inconclusive [12]. In addition, there are other competitive models such as $QTL>T_1>T_2$ *and* $QTL>T_2$; or a loop $QTL>T_1>T_2$ *and* $T_2>T_1$ that prevent clear (conclusive) inferences about the true network directions [8, 10]. As an important cautionary note, the above conditional models are based on various assumptions, and violation of these assumptions may lead to an increase in error rates for inferences about network structure.

At the metabolite level, mQTL for traits connected in a network may show complex patterns of correlation. For example, the mQTL for the precursor and product of an enzymatic step with differential activity should have opposite signs – indicating that the sign of the mQTL effect also conveys valuable information [35-37]. The effect of an mQTL may be visible on the precursor, the corresponding product and downstream products [Fig 1]. As the number of steps grows, the complexity of the network increases, and network reconstruction based purely on correlation coefficients is challenging. Epistatic interactions among enzymes may further complicate the effort to map and deconstruct their unique patterns, as in the

cases where some allelic combinations can be found in offspring which will then produce metabolites not found in either parent. Although such epistasis may be a rare phenomenon of complex traits [38], it is potentially abundant in secondary metabolism [36, 37].

6.4 Inter level analysis

Inter-level inferences have been made between eQTL and mQTL in Arabidopsis [39] and between mQTL and phenotypic traits in tomato [40], and between eQTL, pQTL, mQTL and phenotypic traits in Arabidopsis [41]. A system-wide analysis can reveal the impact of DNA sequence variation across multiple levels, *i.e*. eQTL at the gene expression level, pQTL for protein abundance or activity traits, mQTL for metabolite abundances and/or phQTL for morphological traits (Fig 1). Some DNA sequence variation will induce strong effects to be detected as hotspots or master regulators of many molecular and phenotypic traits, while others induce effects that are more subtle or are buffered in the network to ensure robustness of the system [41]. Correlations among traits from different levels can be used to generate hypotheses about network connections in inter level analyses. Principal components may be used to summarize a network on one level and the regressed on traits on another level [42]. The complexity of the system is such that two adjacent levels (i.e. transcript and protein) may not be linearly related. For example: DNA sequence variation may not affect expression level (no eQTL) while it does affect protein abundance or activity (pQTL). The "higher" level traits (phQTL) may also be a function of multiple underlying (perhaps interacting) sub networks (see the disease trait in Fig 1). Added complexity may be observed when DNA sequence variation directly affect higher level traits that – through feedback loops – affect other traits at the same or lower levels [39]. These examples indicate that caution is warranted given the intrinsic complexity in real networks.

Correlation analyses will only reveal the linear relationships among levels. Interpreting the correlation structure "beyond" the common and shared QTL, using methods such as those described in **Box 1**, may generate hypotheses about system-wide networks. However, extreme caution is advisable in these interpretations in intra level analyses due to the potential impact of correlated measurement error (leading to false positive connections), and in inter level analyses due to the seeming lack of correlation of between levels (leading to false negative connections) [43].

6.5 Using a priori knowledge

Structural and functional data (gene sequence, gene localization, transcription factor

binding sites (TFBS), Gene Ontology (GO), metabolic pathway, protein-protein interaction (PPI)) as well as independent experimental data gleaned from secondary sources (*i.e.* Gene Expression Omnibus (GEO)), can be used post-hoc to verify the defined gene and QTL networks. For example, if a disease maps to multiple QTL, then the candidate genes in each of the QTL can be analyzed and prioritized using known functional interactions [44]. As another example, particular eQTL *trans* bands may be identified as significantly enriched for a functional GO category [45] or as more likely to represent binding sites for transcription factors [46]. Prior knowledge can also be integrated in analysis. For example, a set of pathway-related genes may not show significant eQTL in gene-by-gene tests, while the set of genes can show such significance in a group-wise test [18].

6.6 Future directions

Genetic variation at multiple loci in combination with environmental factors can induce molecular or phenotypic variation. Variation may manifest itself as linear patterns among traits at different levels that can be deconstructed. Correlations can be attributed to detectable QTL and a logical framework based on common and distinct QTL can be used to infer network causality, reactivity or independence. Unexplained variation can be used to infer direction between traits that share a common QTL and have no distinct QTL. Unexplained variation originates from other minor or modifier QTL, epigenetic factors, and biological, environmental and unfortunately, technical factors. Correlation structure present in the molecular data may reflect technical artifacts, in which case the models used to infer causality are potentially invalid and the inference is potentially erroneous. Additional studies are needed to understand and quantify the level of sensitivity of these network reconstruction methods to technical errors. Further research is also needed to develop and evaluate experimental designs other than the current biparental line crosses: for example, multiple line crosses [46-48], advanced intercrosses [47, 48], or populations of natural ecotypes [49-54]. Prior knowledge and complementary experiments such as deletion mapping followed by independent gene expression studies between parental lines may validate or disprove implicated network connections [55].

The trend of genetic studies to go deeper (more levels) and broader (larger scale and more factors including environmental factors), brings challenges to develop methodology that can reconstruct networks more efficiently and more accurately. Despite the obvious limitations of gene and QTL network reconstruction methods, these and other future developments in biotechnology and genetics hold for sure great promise for the field of quantitative genetics.

6.7 References

1. Doerge RW: **Mapping and analysis of quantitative trait loci in experimental populations**. *Nature reviews* 2002, **3**(1):43-52.
2. Jansen RC: **Studying complex biological systems using multifactorial perturbation**. *Nature reviews* 2003, **4**(2):145-151.
3. Gilad Y, Rifkin SA, Pritchard JK: **Revealing the architecture of gene regulation: the promise of eQTL studies**. *Trends Genet* 2008, **24**(8):408-415.
4. Johannes F, Colot V, Jansen RC: **Epigenome dynamics: a quantitative genetics perspective**. *Nature reviews* 2008, **9**(11):883-890.
5. Martienssen RA, Doerge RW, Colot V: **Epigenomic mapping in Arabidopsis using tiling microarrays**. *Chromosome Res* 2005, **13**(3):299-308.
6. Jansen RC, Nap JP: **Genetical genomics: the added value from segregation**. *Trends Genet* 2001, **17**(7):388-391.
7. Li Y, Breitling R, Jansen RC: **Generalizing genetical genomics: getting added value from environmental perturbation**. *Trends Genet* 2008.
8. Chaibub Neto E, Ferrara CT, Attie AD, Yandell BS: **Inferring causal phenotype networks from segregating populations**. *Genetics* 2008, **179**(2):1089-1100.
9. Adewale AJ, Dinu I, Potter JD, Liu Q, Yasui Y: **Pathway analysis of microarray data via regression**. *J Comput Biol* 2008, **15**(3):269-277.
10. Li R, Tsaih SW, Shockley K, Stylianou IM, Wergedal J, Paigen B, Churchill GA: **Structural model analysis of multiple quantitative traits**. *PLoS genetics* 2006, **2**(7):e114.
11. Liu B, de la Fuente A, Hoeschele I: **Gene network inference via structural equation modeling in genetical genomics experiments**. *Genetics* 2008, **178**(3):1763-1776.
12. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C *et al*: **An integrative genomics approach to infer causal associations between gene expression and disease**. *Nature genetics* 2005, **37**(7):710-717.
13. Zhu J, Wiener MC, Zhang C, Fridman A, Minch E, Lum PY, Sachs JR, Schadt EE: **Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations**. *PLoS computational biology* 2007, **3**(4):e69.
14. Wright S: **Correlation and causation**. *Journal of agricultural research* 1921, **20**(7):557-585.

15. DeCook R, Lall S, Nettleton D, Howell SH: **Genetic regulation of gene expression during shoot development in Arabidopsis**. *Genetics* 2006, **172**(2):1155-1164.
16. Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, Peeters AJ, Vreugdenhil D, Koornneef M, Jansen RC: **Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(5):1708-1713.
17. Wang D, Nettleton D: **Combining classical trait and microarray data to dissect transcriptional regulation: a case study**. *TAG Theoretical and applied genetics* 2008, **116**(5):683-690.
18. Kliebenstein DJ, West MA, van Leeuwen H, Loudet O, Doerge RW, St Clair DA: **Identification of QTLs controlling gene expression networks defined a priori**. *BMC bioinformatics* 2006, **7**:308.
19. Juenger TE, Wayne T, Boles S, Symonds VV, McKay J, Coughlan SJ: **Natural genetic variation in whole-genome expression in Arabidopsis thaliana: the impact of physiological QTL introgression**. *Molecular ecology* 2006, **15**(5):1351-1365.
20. West MA, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, St Clair DA: **Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis**. *Genetics* 2007, **175**(3):1441-1450.
21. Vuylsteke M, Daele H, Vercauteren A, Zabeau M, Kuiper M: **Genetic dissection of transcriptional regulation by cDNA-AFLP**. *Plant J* 2006, **45**(3):439-446.
22. Druka A, Potokina E, Luo Z, Bonar N, Druka I, Zhang L, Marshall DF, Steffenson BJ, Close TJ, Wise RP *et al*: **Exploiting regulatory variation to identify genes underlying quantitative resistance to the wheat stem rust pathogen Puccinia graminis f. sp. tritici in barley**. *TAG Theoretical and applied genetics* 2008, **117**(2):261-272.
23. Kirst M, Basten CJ, Myburg AA, Zeng ZB, Sederoff RR: **Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid**. *Genetics* 2005, **169**(4):2295-2303.
24. Potokina E, Druka A, Luo Z, Moscou M, Wise R, Waugh R, Kearsey M: **Tissue-dependent limited pleiotropy affects gene expression in barley**. *Plant J* 2008, **56**(2):287-296.
25. Potokina E, Druka A, Luo Z, Wise R, Waugh R, Kearsey M: **Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation**. *Plant J* 2008, **53**(1):90-101.
26. Shi C, Uzarowska A, Ouzunova M, Landbeck M, Wenzel G, Lubberstedt T: **Identification of candidate genes associated with cell wall digestibility**

and eQTL (expression quantitative trait loci) analysis in a Flint x Flint maize recombinant inbred line population. *BMC genomics* 2007, **8**:22.

27. Springer NM, Stupar RM: **Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize**. *The Plant cell* 2007, **19**(8):2391-2402.
28. Gilad Y, Borevitz J: **Using DNA microarrays to study natural variation**. *Current opinion in genetics & development* 2006, **16**(6):553-558.
29. Alberts R, Terpstra P, Li Y, Breitling R, Nap JP, Jansen RC: **Sequence polymorphisms cause many false cis eQTLs**. *PLoS ONE* 2007, **2**(7):e622.
30. Bing N, Hoeschele I: **Genetical genomics analysis of a yeast segregant population for transcription network inference**. *Genetics* 2005, **170**(2):533-542.
31. Kulp DC, Jagalur M: **Causal inference of regulator-target pairs by gene mapping of expression phenotypes**. *BMC genomics* 2006, **7**:125.
32. Stylianou IM, Affourtit JP, Shockley KR, Wilpan RY, Abdi FA, Bhardwaj S, Rollins J, Churchill GA, Paigen B: **Applying gene expression, proteomics and single-nucleotide polymorphism analysis for complex trait gene identification**. *Genetics* 2008, **178**(3):1795-1805.
33. de Koning DJ, Haley CS: **Genetical genomics in humans and model organisms**. *Trends Genet* 2005, **21**(7):377-381.
34. Breitling R, Li Y, Tesson BM, Fu J, Wu C, Wiltshire T, Gerrits A, Bystrykh LV, de Haan G, Su AI *et al*: **Genetical genomics: spotlight on QTL hotspots**. *PLoS genetics* 2008, **4**(10):e1000232.
35. Fu J, Swertz MA, Keurentjes JJ, Jansen RC: **MetaNetwork: a computational protocol for the genetic study of metabolic networks**. *Nature protocols* 2007, **2**(3):685-694.
36. Keurentjes JJ, Fu J, de Vos CH, Lommen A, Hall RD, Bino RJ, van der Plas LH, Jansen RC, Vreugdenhil D, Koornneef M: **The genetics of plant metabolism**. *Nature genetics* 2006, **38**(7):842-849.
37. Rowe HC, Hansen BG, Halkier BA, Kliebenstein DJ: **Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome**. *The Plant cell* 2008, **20**(5):1199-1216.
38. Hill WG, Goddard ME, Visscher PM: **Data and theory point to mainly additive genetic variance for complex traits**. *PLoS genetics* 2008, **4**(2):e1000008.
39. Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ: **Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways**. *PLoS genetics* 2007, **3**(9):1687-1701.
40. Schauer N, Semel Y, Balbo I, Steinfath M, Repsilber D, Selbig J, Pleban T, Zamir D, Fernie AR: **Mode of inheritance of primary metabolic traits in tomato**. *The Plant cell* 2008, **20**(3):509-523.
41. Fu J, Keurentjes JJB, Bouwmeester H, America T, Verstappen FWA, Ward

JL, Beale MH, de Vos RCH, Dijkstra M, Scheltema RA *et al*: **System-wide molecular evidence for phenotypic buffering in Arabidopsis**. *Nature genetics,* 2009, **41**, 166-167.

42. Coffman CJ, Wayne ML, Nuzhdin SV, Higgins LA, McIntyre LM: **Identification of co-regulated transcripts affecting male body size in Drosophila**. *Genome biology* 2005, **6**(6):R53.
43. Jansen RC, Nap JP, Mlynarova L: **Errors in genomics and proteomics**. *Nature biotechnology* 2002, **20**(1):19.
44. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes**. *American journal of human genetics* 2006, **78**(6):1011-1025.
45. Wu C, Delano DL, Mitro N, Su SV, Janes J, McClurg P, Batalov S, Welch GL, Zhang J, Orth AP *et al*: **Gene set enrichment in eQTL data identifies novel annotations and pathway regulators**. *PLoS genetics* 2008, **4**(5):e1000070.
46. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE: **Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks**. *Nature genetics* 2008, **40**(7):854-861.
47. Chesler EJ, Miller DR, Branstetter LR, Galloway LD, Jackson BL, Philip VM, Voy BH, Culiat CT, Threadgill DW, Williams RW *et al*: **The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics**. *Mamm Genome* 2008, **19**(6):382-389.
48. Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, Beavis WD, Belknap JK, Bennett B, Berrettini W *et al*: **The Collaborative Cross, a community resource for the genetic analysis of complex traits**. *Nature genetics* 2004, **36**(11):1133-1137.
49. Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang C *et al*: **Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes**. *PLoS genetics* 2005, **1**(5):e60.
50. Gonzalez-Martinez SC, Huber D, Ersoz E, Davis JM, Neale DB: **Association genetics in Pinus taeda L. II. Carbon isotope discrimination**. *Heredity* 2008, **101**(1):19-26.
51. Kim S, Zhao K, Jiang R, Molitor J, Borevitz JO, Nordborg M, Marjoram P: **Association mapping with single-feature polymorphisms**. *Genetics* 2006, **173**(2):1125-1133.
52. Rosenberg NA, Nordborg M: **A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations**.

Genetics 2006, **173**(3):1665-1678.

53. Zhao J, Paulo MJ, Jamar D, Lou P, van Eeuwijk F, Bonnema G, Vreugdenhil D, Koornneef M: **Association mapping of leaf traits, flowering time, and phytate content in Brassica rapa**. *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada* 2007, **50**(10):963-973.
54. Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P *et al*: **An Arabidopsis example of association mapping in structured samples**. *PLoS genetics* 2007, **3**(1):e4.
55. Wayne ML, McIntyre LM: **Combining mapping and arraying: An approach to candidate gene identification**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(23):14903-14906.

Chapter 7
Critical reasoning on causal inference in genome-wide linkage and association studies

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 that they are already valid in simple scenarios of two traits linked to or associated with the same genomic region. We argue that more comprehensive analysis and Bayesian reasoning are needed and that these can overcome some of these pitfalls, although not in every conceivable case. We conclude that causal inference methods can still be of use in the iterative process of mathematical modeling and biological validation.

Originally published as:
Critical reasoning on causal inference in genome-wide linkage and association studies.
Li Y*, **Tesson BM***, Churchill GA, Jansen RC, de Haan G.
Trends in Genetics 2010 Dec; 26(12):493-8.
*equal contributions

7.1 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 as yet unknown but potentially causal relationships.

Methods for causal inference have been introduced as early as the 1920s[3] and have been further developed and applied since then in genetic epidemiology and other fields [2, 5, 6]). Causal inference is a formal statistical procedure that aims to establish predictive models. For example, if a reduction in the level of a crucial 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 crucial to the process of target discovery in pharmaceutical research.

Recent genome-wide linkage studies (GWLS) on model organisms [8-10] and genome-wide association studies (GWAS) on humans [11] have successfully connected molecular and classical traits into networks with arrows indicating inferred causal relationships [4, 12-19]. 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/loci, QTL; Glossary) 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 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).

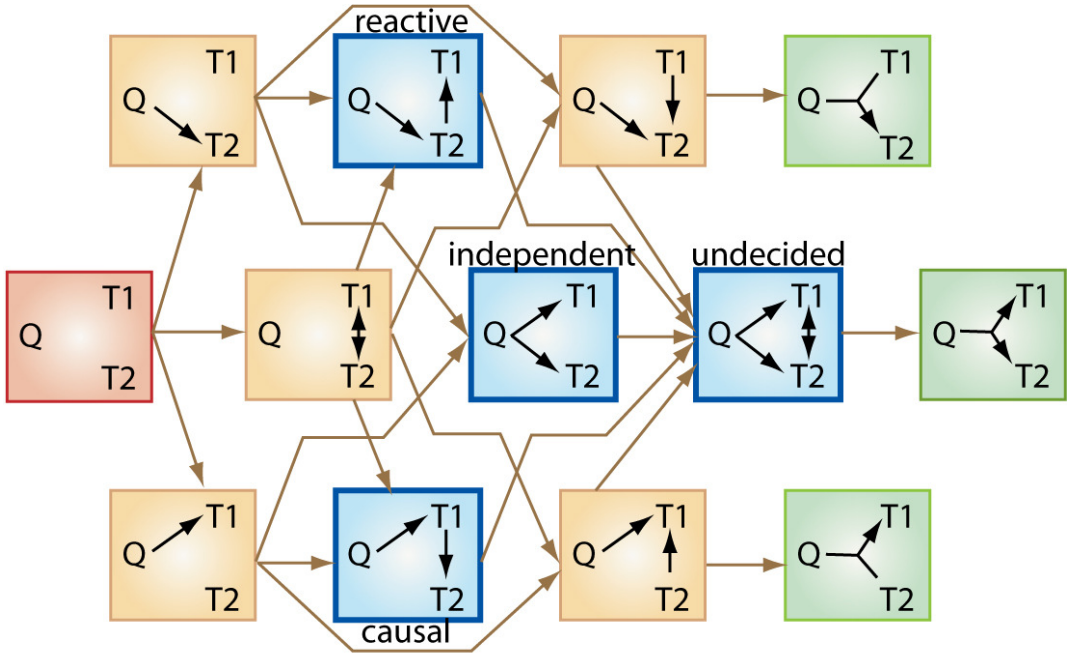


Figure 1 - Triad models. Many different causal relationships are possible within a triad of two traits (T1 and T2) and a QTL (Q). The simplest case (red box) to the left shows no causality, in which case the QTL and the two traits do not influence each other. In the next set of models (yellow), at least 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, for example the QTL could modulate the causal effect of T1 on T2. Equivalently, these models could 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 could form an important part of actual biological networks. The brown arrows indicate which of the models are nested and can thus be compared directly by statistical testing.

Biological variation in the two traits beyond that induced by the common QTL is the key for 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 1**), subtract the biological and QTL variation in T1 from T2 and we are left with the additional or 'residual' variation in T2 that is unrelated to the QTL. If we attempt the reciprocal analysis, the additional variation in T2 could 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 for distinguishing 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 1** and **Table 1** for the statistical details). Although 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.

Box 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 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 I). 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 1**).

Positive predictive value: the positive predictive value is probability of a declared causal connection being true. We incorporate prior knowledge (**Box 2** and **Glossary**): 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. The positive predictive value is then 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 2**).

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

Table 1 - Equations for regression parameters in the basic independent and causal model (first scenario in the main text). T1 and T2 have mean zero and equal QTL effect; this can always be achieved by subtracting the means and re-scaling. 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. For variance and covariance equations, 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 $n_H=0$.

7.2 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 [20]. The method will declare particular 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 can in some cases be incorrectly identified as causal by the method. As a consequence, the potential number of false causal links arising from, say, 80% of independent trait-trait pairs can overwhelm the number of true causal links arising from the 20% of 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 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 2**).

We present three different scenarios to illustrate the properties of the method. In 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 can originate from an independent perturbation such as another QTL affecting T2 but not T1, or from an environmental perturbation affecting T2 but not T1. The correlation between T1 and T2 results fully from the causal relationship between the two traits. Exact analytical equations can be used to compute the population size required to attain the desired levels of sensitivity and positive predictive value (**Box 1**). This requires specifying the size of the QTL effect, the frequency in the population of the major QTL allele, and the prior belief 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 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 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, in other words sensitivity is low.

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, such as 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 can appear to be 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 truly 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 2** is still approximately adequate if the hidden factor variance is small, in other words equals at most the QTL variance.

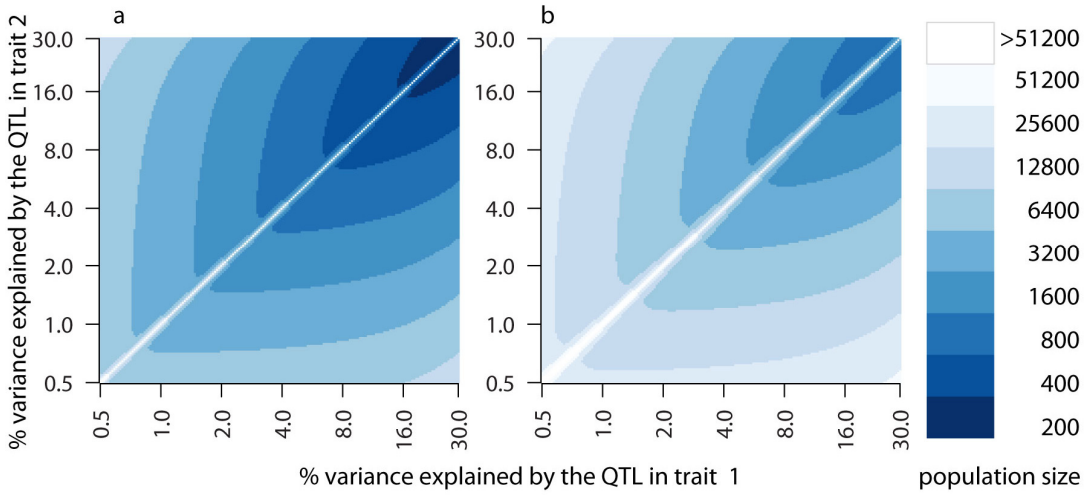


Figure 2 - Population sizes required for reliable causal inference. Here we show the required population size in **(a)** genome-wide linkage studies (GWLS) and **(b)** genome-wide association studies. Each color represents a different population size; the scale is shown in the right panel. These numbers have been calculated from the equations in **Box 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 at 10% and 90% in GWAS. Furthermore we use Bayesian reasoning (**Box 2**): we assume *a priori* that only 1% (20%) of the QTL-trait-trait connections is truly causal in GWLS (GWAS).

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, can also introduce a kind of measurement error. Measurement variation is never 'biologically' propagated from one trait to another trait, but 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 appear to be independent (yes, yes) if the 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 [16]. Again, increasing sample size will not be helpful to attain the desired levels of sensitivity and positive predictive value.

7.3 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 1**) to potentially 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 [8, 10, 12] which, although not identical, lead to comparable results [14]. Other computational methods for causal inference such as structural equation modeling [21, 22] or Bayesian network analysis [23] 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 sizes, 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 [24] or Kyoto Encyclopedia of Genes and Genomes (KEGG) [25] pathways should also be considered when weighing evidence for causal links. The use of more informative priors (**Box 2**) provides better prioritizing and filtering of the large numbers of possible triads, and could reduce the population sizes required 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, but 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 [26, 27]. Although it might not be necessary in inbred line cross studies, it is crucial 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 [28]. The effects of hidden factors affecting larger numbers of traits can be detected and corrected for by dimension-reduction methods ([28-32]). Causal inference can then be applied to the residual data. However, these multivariate analysis methods also have the potential to remove from the data signals that are 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 1**. For example, fitting a model such as the top-right yellow model in **Figure 1** could provide a powerful case for the causal signal in the data [19, 21, 22]. The green models in **Figure 1** with more complex correlation structure can also be informative and have been explored [19]. If two traits have multiple QTL in common, then this can be taken as additional evidence that the two traits are connected in the network [7]. 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 1**) to assess the combined effect [33].

7.4 Concluding remarks

Many in the scientific community share a healthy skepticism of causal inference and, as we have shown, for good reasons. Nevertheless we conclude that causal inference in linkage or association analysis could 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 as the development of techniques that address the effects of experimental factors, study design and population structure. Reasonable caution is still warranted and statistical methods of causal inference should be viewed as a necessary step in an era of high-throughput data generation and discovery.

Box 2. Bayesian Reasoning.

Bayes rule [1] 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[4].
- **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[7]. 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.

7.5 Acknowledgements

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.

Glossary

Allele frequencies
At a given polymorphic locus, the different alleles can differ in prevalence within the population studied. In GWLS using a cross originating from two inbred founders, the QTL has two alleles at equal frequencies in the population under study. By contrast, due to a combination of random segregation, drift and selection, allele frequencies in GWAS 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 can have hidden structure, or in case-control studies where sampling could 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 [2] 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 the original scale or after a transformation).

Genome-wide association studies (GWAS)
A genome wide association study is an experiment in which the genomes of unrelated individuals are screened for genetic markers (typically millions of single nucleotide polymorphisms, SNPs) 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 SNPs) at which allelic variation correlates with variation in studied traits. Examples of GWLS include experimental crosses such as recombinant inbred panels, intercrosses and backcrosses.

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.

Glossary (continued)

Quantitative Trait Locus (QTL)
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.

- **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.
- **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).
- **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.

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 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.

7.6 References

1. Stephens M, Balding DJ: **Bayesian statistical methods for genetic association studies**. *Nat Rev Genet* 2009, **10**(10):681-690.
2. Pearl J: **Causality: models, reasoning, and inference**, 2nd ed edn: Cambridge University Press; 2000.
3. Wright S: **Correlation and causation**. *J Agric Res* 1921, **20**:557-585.
4. Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, Thieringer R, Berger JP, Wu MS, Thompson J, Sachs AB *et al*: **An integrative genomics approach to the reconstruction of gene networks in segregating populations**. *Cytogenet Genome Res* 2004, **105**(2-4):363-374.
5. Duffy DL, Martin NG: **Inferring the direction of causation in cross-sectional twin data: theoretical and empirical considerations**. *Genet Epidemiol* 1994, **11**(6):483-502.
6. Spirtes P, Glymour C, Scheines R: **Causation, Prediction, and Search**. New York: Springer-Verlag; 1993.
7. Jansen RC, Nap JP: **Genetical genomics: the added value from segregation**. *Trends Genet* 2001, **17**(7):388-391.
8. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK *et al*: **Variations in DNA elucidate molecular networks that cause disease**. *Nature* 2008, **452**(7186):429-435.
9. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE: **Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks**. *Nat Genet* 2008, **40**(7):854-861.
10. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C *et al*: **An integrative genomics approach to infer causal associations between gene expression and disease**. *Nat Genet* 2005, **37**(7):710-717.
11. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S *et al*: **Genetics of gene expression and its effect on disease**. *Nature* 2008, **452**(7186):423-428.
12. Chen LS, Emmert-Streib F, Storey JD: **Harnessing naturally randomized transcription to infer regulatory relationships among genes**. *Genome Biol* 2007, **8**(10):R219.
13. Aten JE, Fuller TF, Lusis AJ, Horvath S: **Using genetic markers to orient the edges in quantitative trait networks: the NEO software**. *BMC Syst Biol* 2008, **2**:34.
14. Millstein J, Zhang B, Zhu J, Schadt EE: **Disentangling molecular relationships with a causal inference test**. *BMC Genet* 2009, **10**:23.
15. Chaibub Neto E, Ferrara CT, Attie AD, Yandell BS: **Inferring causal phenotype networks from segregating populations**. *Genetics* 2008,

179(2):1089-1100.
16. Rockman MV: **Reverse engineering the genotype-phenotype map with natural genetic variation**. *Nature* 2008, **456**(7223):738-744.
17. Bing N, Hoeschele I: **Genetical genomics analysis of a yeast segregant population for transcription network inference**. *Genetics* 2005, **170**(2):533-542.
18. Li H, Lu L, Manly KF, Chesler EJ, Bao L, Wang J, Zhou M, Williams RW, Cui Y: **Inferring gene transcriptional modulatory relations: a genetical genomics approach**. *Hum Mol Genet* 2005, **14**(9):1119-1125.
19. Kulp DC, Jagalur M: **Causal inference of regulator-target pairs by gene mapping of expression phenotypes**. *BMC Genomics* 2006, **7**:125.
20. Visscher PM, Hill WG, Wray NR: **Heritability in the genomics era--concepts and misconceptions**. *Nat Rev Genet* 2008, **9**(4):255-266.
21. Li R, Tsaih SW, Shockley K, Stylianou IM, Wergedal J, Paigen B, Churchill GA: **Structural model analysis of multiple quantitative traits**. *PLoS Genet* 2006, **2**(7):e114.
22. Liu B, de la Fuente A, Hoeschele I: **Gene network inference via structural equation modeling in genetical genomics experiments**. *Genetics* 2008, **178**(3):1763-1776.
23. Zhu J, Wiener MC, Zhang C, Fridman A, Minch E, Lum PY, Sachs JR, Schadt EE: **Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations**. *PLoS Comput Biol* 2007, **3**(4):e69.
24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
25. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.
26. Li Y, Breitling R, Jansen RC: **Generalizing genetical genomics: getting added value from environmental perturbation**. *Trends Genet* 2008, **24**(10):518-524.
27. Akey JM, Biswas S, Leek JT, Storey JD: **On the design and analysis of gene expression studies in human populations**. *Nat Genet* 2007, **39**(7):807-808; author reply 808-809.
28. Ghazalpour A, Doss S, Kang H, Farber C, Wen PZ, Brozell A, Castellanos R, Eskin E, Smith DJ, Drake TA *et al*: **High-resolution mapping of gene expression using association in an outbred mouse stock**. *PLoS Genet* 2008, **4**(8):e1000149.
29. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adany R, Aromaa A *et al*: **Multiple common variants for celiac disease influencing immune gene expression**. *Nat*

Genet 2010, **42**:295-302.

30. Fehrmann RS, de Jonge HJ, Ter Elst A, de Vries A, Crijns AG, Weidenaar AC, Gerbens F, de Jong S, van der Zee AG, de Vries EG *et al*: **A new perspective on transcriptional system regulation (TSR): towards TSR profiling**. *PLoS One* 2008, **3**(2):e1656.
31. Leek JT, Storey JD: **Capturing heterogeneity in gene expression studies by surrogate variable analysis**. *PLoS Genet* 2007, **3**(9):1724-1735.
32. Stegle O, Parts L, Durbin R, Winn J: **A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies**. *PLoS Comput Biol* 2010, **6**(5):e1000770.
33. Sargon JD: **The estimation of economic relationships using instrumental variables.** *Econometrica* 1958, **26**:393-415.

Chapter 8
Scaling up classical genetics to thousands of molecular traits: promises and challenges

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 review 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.

Originally published as:
Scaling up classical genetics to thousands of molecular traits: promises and challenges.
Tesson BM*, Li Y*, Breitling R, Jansen RC
Proceedings of the 9th World Congress on Genetics applied to Livestock. 2010 Aug 1-6, Leipsig, Germany

*equal contributions

8.1 Introduction

Genetical genomics [1, 2] 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 [3] for unraveling molecular regulatory networks: a PubMed search on relevant keywords currently yields 191 scientific publications [4], 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) in numerous model species (including yeasts, plants, worms, flies, mice, rats and humans) [5-11], and they have exposed the plasticity of the eQTL that control those traits with respect to environmental condition, tissue type or cellular context [12-16]. 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 [17] and have suggested molecular pathways through which some genetic variants can cause diseases [18-20].

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.

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.

8.2.1 Population

In genetical genomics studies, multiple testing caused by the mapping of large

numbers of phenotypes reduces the available statistical power. Linkage mapping on 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 experiments 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. Other types of crosses, such as the mouse collaborative cross [21], the *Arabidopsis thaliana* Multiparent Advanced Generation Inter-Cross [22], or Heterogeneous Stock (HS) for rat [23], may offer 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 [18, 24-27] and are emerging in model organisms studies using outbred populations [28].

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 [24, 25]. Meta-analyses use statistical methods for combining *p*-values [29], 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 [30]. 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.

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 [31, 32]. 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 [33, 34]. 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. 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 [34-36].

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) [37]. In practice, the approaches for calculating the significance thresholds accounting for multiple testing used in genetical genomics are mostly relying on permutations [38, 39], since standard approaches [40] 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 [39]. 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 [41]. 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 [42].

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 distant 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 [43].

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 (see e.g. [44, 45]. In the genetical genomics context, these samples come from genetically diverse individuals, possibly observed over multiple conditions. Under the principle of "guilt by association", coexpression can be used to predict similar gene functions and is indicative of possible co-regulation.

From the network, modules of coexpressed genes can be obtained, i.e. 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" [8]. Searching for cliques in a network containing thousands of nodes poses a serious computational burden and several algorithms have been designed to alleviate it [46]. 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 [47, 48], but without the strong constraint imposed by cliqueness.

Connectivity (also known as degree) represents the amount of edges reaching a gene in the coexpression network. Genes with high connectivity, termed "hubs", have been claimed to be enriched for essential genes [45]. Connectivity is therefore used to prioritize between genes belonging to modules of interest.

Similar correlation-based approaches can be used to study metabolites [49]. 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 [50]. 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 [17].

By using partial correlations, *i.e.* conditioning on selected other nodes in the network, it is possible to remove indirect edges from the network [51-53]. 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 [54, 55]. 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 [1]. 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 *distant* eQTL [2]. 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 [56, 57]. Chen et al. provided a method to quantify the likelihood of each causal link [58]. 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 [59]. Chaibub Neto et al. propose a likelihood-based method to compare graph configurations in which the non-propagated variation present in the downstream trait is explicitly modeled by non-shared QTL(s)[60]. 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 [61].

Some attempts have been made to combine co-expression networks with QTL-based causal inference: either by orienting undirected edges of coexpression networks [60, 62] or by inferring causal relationships between entire modules and clinical traits by studying the eigengenes representing those modules or selected genes from those modules [63, 64].

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, 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 [17]. 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 [65, 66] or iGA [67]. 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 [68], 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 [69]. 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 [67, 70].

8.5 Conclusion

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 [17, 71, 72] and broader (larger sample sizes, combining genetic perturbation with other factors such as environmental factors) [35, 73]. 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

[24, 74, 75] 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 [48, 56, 58, 60] promises to be accompanied by new breakthroughs by removing one of the major obstacles on the way towards reliable network inference [61].

8.6 References

1. Jansen RC, Nap JP: **Genetical genomics: the added value from segregation**. *Trends Genet* 2001, **17**(7):388-391.
2. Jansen RC: **Studying complex biological systems using multifactorial perturbation**. *Nat Rev Genet* 2003, **4**(2):145-151.
3. Sieberts SK, Schadt EE: **Moving toward a system genetics view of disease**. *Mamm Genome* 2007, **18**(6-7):389-401.
4. webCite: **PubMed search 17-05-2010 for "eQTL" OR "Genetical Genomics"**. In.; 2010.
5. Brem RB, Yvert G, Clinton R, Kruglyak L: **Genetic dissection of transcriptional regulation in budding yeast**. *Science* 2002, **296**(5568):752-755.
6. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G *et al*: **Genetics of gene expression surveyed in maize, mouse and man**. *Nature* 2003, **422**(6929):297-302.
7. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF *et al*: **Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'**. *Nat Genet* 2005, **37**(3):225-232.
8. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA *et al*: **Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function**. *Nat Genet* 2005, **37**(3):233-242.
9. DeCook R, Lall S, Nettleton D, Howell SH: **Genetic regulation of gene expression during shoot development in Arabidopsis**. *Genetics* 2006, **172**(2):1155-1164.
10. Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, Fischer J, Maatz H, Kren V, Pravenec M *et al*: **Heritability and tissue specificity of expression quantitative trait loci**. *PLoS Genet* 2006, **2**(10):e172.
11. Ruden DM, Chen L, Possidente D, Possidente B, Rasouli P, Wang L, Lu X, Garfinkel MD, Hirsch HV, Page GP: **Genetical toxicogenomics in Drosophila identifies master-modulatory loci that are regulated by developmental exposure to lead**. *Neurotoxicology* 2009, **30**(6):898-914.
12. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M *et al*: **Common regulatory variation impacts gene expression in a cell type-dependent manner**. *Science* 2009, **325**(5945):1246-1250.
13. Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, Le J, Koka V, Lam KC, Gagne V *et al*: **Global patterns of *cis* variation in human cells revealed by high-density allelic expression analysis**. *Nat*

Genet 2009, **41**(11):1216-1222.

14. Gerrits A, Li Y, Tesson BM, Bystrykh LV, Weersing E, Ausema A, Dontje B, Wang X, Breitling R, Jansen RC *et al*: **Expression quantitative trait loci are highly sensitive to cellular differentiation state**. *PLoS Genet* 2009, **5**(10):e1000692.
15. Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JA, Hazendonk E, Prins P, Plasterk RH, Jansen RC *et al*: **Mapping determinants of gene expression plasticity by genetical genomics in C. elegans**. *PLoS Genet* 2006, **2**(12):e222.
16. Smith EN, Kruglyak L: **Gene-environment interaction in yeast gene expression**. *PLoS Biol* 2008, **6**(4):e83.
17. Fu J, Keurentjes JJ, Bouwmeester H, America T, Verstappen FW, Ward JL, Beale MH, de Vos RC, Dijkstra M, Scheltema RA *et al*: **System-wide molecular evidence for phenotypic buffering in Arabidopsis**. *Nat Genet* 2009, **41**(2):166-167.
18. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S *et al*: **Genetics of gene expression and its effect on disease**. *Nature* 2008, **452**(7186):423-428.
19. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C *et al*: **Mapping the genetic architecture of gene expression in human liver**. *PLoS Biol* 2008, **6**(5):e107.
20. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E *et al*: **Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma**. *Nature* 2007, **448**(7152):470-473.
21. Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, Beavis WD, Belknap JK, Bennett B, Berrettini W *et al*: **The Collaborative Cross, a community resource for the genetic analysis of complex traits**. *Nat Genet* 2004, **36**(11):1133-1137.
22. Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R: **A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana**. *PLoS Genet* 2009, **5**(7):e1000551.
23. Hansen C, Spuhler K: **Development of the National Institutes of Health genetically heterogeneous rat stock**. *Alcohol Clin Exp Res* 1984, **8**(5):477-479.
24. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adany R, Aromaa A *et al*: **Multiple common variants for celiac disease influencing immune gene expression**. *Nat Genet* 2010, **42**:295-302.
25. Heap GA, Trynka G, Jansen RC, Bruinenberg M, Swertz MA, Dinesen LC, Hunt KA, Wijmenga C, Vanheel DA, Franke L: **Complex nature of SNP**

genotype effects on gene expression in primary human leucocytes. *BMC Med Genomics* 2009, **2**:1.

26. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S *et al*: **Genome-wide associations of gene expression variation in humans**. *PLoS Genet* 2005, **1**(6):e78.
27. Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG *et al*: **Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes**. *Nat Genet* 2007, **39**(10):1208-1216.
28. Ghazalpour A, Doss S, Kang H, Farber C, Wen PZ, Brozell A, Castellanos R, Eskin E, Smith DJ, Drake TA *et al*: **High-resolution mapping of gene expression using association in an outbred mouse stock**. *PLoS Genet* 2008, **4**(8):e1000149.
29. Whitlock MC: **Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach**. *J Evol Biol* 2005, **18**(5):1368-1373.
30. Gatti DM, Harrill AH, Wright FA, Threadgill DW, Rusyn I: **Replication and narrowing of gene expression quantitative trait loci using inbred mice**. *Mamm Genome* 2009, **20**(7):437-446.
31. Fisher R: **The design of experiments**, 6th edn. London: UK: Olivier and Boyd; 1951.
32. Wit E, McClure JD: **Statistics for Microarrays; Design, Analysis and Inference**. Chichester: John Wiley & Sons; 2004.
33. Fu J, Jansen RC: **Optimal design and analysis of genetic studies on gene expression**. *Genetics* 2006, **172**(3):1993-1999.
34. Lam AC, Fu J, Jansen RC, Haley CS, de Koning DJ: **Optimal design of genetic studies of gene expression with two-color microarrays in outbred crosses**. *Genetics* 2008, **180**(3):1691-1698.
35. Li Y, Breitling R, Jansen RC: **Generalizing genetical genomics: getting added value from environmental perturbation**. *Trends Genet* 2008.
36. Li Y, Swertz MA, Vera G, Fu J, Breitling R, Jansen RC: **designGG: an R-package and web tool for the optimal design of genetical genomics experiments**. *BMC Bioinformatics* 2009, **10**:188.
37. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Statist Soc* 1995, **57**:289-300.
38. Churchill GA, Doerge RW: **Empirical threshold values for quantitative trait mapping**. *Genetics* 1994, **138**(3):963-971.
39. Breitling R, Li Y, Tesson BM, Fu J, Wu C, Wiltshire T, Gerrits A, Bystrykh LV, de Haan G, Su AI *et al*: **Genetical genomics: spotlight on QTL hotspots**. *PLoS Genet* 2008, **4**(10):e1000232.
40. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**.

Proc Natl Acad Sci U S A 2003, **100**(16):9440-9445.

41. Knijnenburg TA, Wessels LF, Reinders MJ, Shmulevich I: **Fewer permutations, more accurate P-values**. *Bioinformatics* 2009, **25**(12):i161-168.
42. Anderson M, Ter Braak C: **Permutation tests for multi-factorial analysis of variance**. *Journal of Statistical Computation and Simulation* 2003, **73**:85-113.
43. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB: **Rare variants create synthetic genome-wide associations**. *PLoS Biol* 2010, **8**(1):e1000294.
44. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS: **Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks**. *Proc Natl Acad Sci U S A* 2000, **97**(22):12182-12186.
45. Carter SL, Brechbuhler CM, Griffin M, Bond AT: **Gene co-expression network topology provides a framework for molecular characterization of cellular state**. *Bioinformatics* 2004, **20**(14):2242-2250.
46. Baldwin NE, Chesler EJ, Kirov S, Langston MA, Snoddy JR, Williams RW, Zhang B: **Computational, integrative, and comparative methods for the elucidation of genetic coexpression networks**. *J Biomed Biotechnol* 2005, **2005**(2):172-180.
47. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks**. *Science* 2002, **297**(5586):1551-1555.
48. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis**. *Stat Appl Genet Mol Biol* 2005, **4**:Article17.
49. Kose F, Weckwerth W, Linke T, Fiehn O: **Visualizing plant metabolomic correlation networks using clique-metabolite matrices**. *Bioinformatics* 2001, **17**(12):1198-1208.
50. Steuer R: **Review: on the analysis and interpretation of correlations in metabolomic data**. *Brief Bioinform* 2006, **7**(2):151-158.
51. de la Fuente A, Bing N, Hoeschele I, Mendes P: **Discovery of meaningful associations in genomic data using partial correlation coefficients**. *Bioinformatics* 2004, **20**(18):3565-3574.
52. Bing N, Hoeschele I: **Genetical genomics analysis of a yeast segregant population for transcription network inference**. *Genetics* 2005, **170**(2):533-542.
53. Keurentjes JJ, Fu J, de Vos CH, Lommen A, Hall RD, Bino RJ, van der Plas LH, Jansen RC, Vreugdenhil D, Koornneef M: **The genetics of plant metabolism**. *Nat Genet* 2006, **38**(7):842-849.
54. Choi Y, Kendziorski C: **Statistical methods for gene set co-expression analysis**. *Bioinformatics* 2009, **25**(21):2780-2786.

55. Tesson BM, Breitling R, Jansen RC: **DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules**. *submitted* 2010.
56. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C *et al*: **An integrative genomics approach to infer causal associations between gene expression and disease**. *Nat Genet* 2005, **37**(7):710-717.
57. Li R, Tsaih SW, Shockley K, Stylianou IM, Wergedal J, Paigen B, Churchill GA: **Structural model analysis of multiple quantitative traits**. *PLoS Genet* 2006, **2**(7):e114.
58. Chen LS, Emmert-Streib F, Storey JD: **Harnessing naturally randomized transcription to infer regulatory relationships among genes**. *Genome biology* 2007, **8**(10):R219.
59. Millstein J, Zhang B, Zhu J, Schadt EE: **Disentangling molecular relationships with a causal inference test**. *BMC Genet* 2009, **10**:23.
60. Chaibub Neto E, Ferrara CT, Attie AD, Yandell BS: **Inferring causal phenotype networks from segregating populations**. *Genetics* 2008, **179**(2):1089-1100.
61. Li Y, Tesson BM, Churchill GA, Jansen RC: **Critical preconditions for causal inference in genome-wide association studies** *under review* 2010.
62. Aten JE, Fuller TF, Lusis AJ, Horvath S: **Using genetic markers to orient the edges in quantitative trait networks: the NEO software**. *BMC Syst Biol* 2008, **2**:34.
63. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK *et al*: **Variations in DNA elucidate molecular networks that cause disease**. *Nature* 2008, **452**(7186):429-435.
64. Plaisier CL, Horvath S, Huertas-Vazquez A, Cruz-Bautista I, Herrera MF, Tusie-Luna T, Aguilar-Salinas C, Pajukanta P: **A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia**. *PLoS Genet* 2009, **5**(9):e1000642.
65. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Muller R, Meese E, Lenhof HP: **GeneTrail--advanced gene set enrichment analysis**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W186-192.
66. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.
67. Breitling R, Amtmann A, Herzyk P: **Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments**. *BMC Bioinformatics* 2004, **5**:34.
68. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L: ***Trans*-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors**. *Nat Genet* 2003, **35**(1):57-64.

69. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes**. *Am J Hum Genet* 2006, **78**(6):1011-1025.
70. Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, Peeters AJ, Vreugdenhil D, Koornneef M, Jansen RC: **Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci**. *Proc Natl Acad Sci U S A* 2007, **104**(5):1708-1713.
71. Johannes F, Colot V, Jansen RC: **Epigenome dynamics: a quantitative genetics perspective**. *Nat Rev Genet* 2008, **9**(11):883-890.
72. Ferrara CT, Wang P, Neto EC, Stevens RD, Bain JR, Wenner BR, Ilkayeva OR, Keller MP, Blasiole DA, Kendziorski C *et al*: **Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling**. *PLoS Genet* 2008, **4**(3):e1000034.
73. Jansen RC, Tesson BM, Fu J, Yang Y, McIntyre LM: **Defining gene and QTL networks**. *Curr Opin Plant Biol* 2009, **12**(2):241-246.
74. Kang HM, Ye C, Eskin E: **Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots**. *Genetics* 2008, **180**(4):1909-1925.
75. Leek JT, Storey JD: **Capturing heterogeneity in gene expression studies by surrogate variable analysis**. *PLoS Genet* 2007, **3**(9):1724-1735.

Samenvatting

Variatie in DNA sequencies staat aan de oorsprong van phenotypische variatie. Het begrijpen van de mechanismes hoe deze variatie doorwerkt in weefsel, organen en organismes is het doel van systeem genetica. Genetical genomics gebruikt quantitatieve methoden om de in de natuur aanwezige DNA varianten te identificeren die ten grondslag liggen aan de variatie in mRNAs, eiwitten en metabolieten. Deze thesis onderzoekt het vermogen van Genetical Genomics om de moleculaire interacties te ontraffelen die zorgen voor complexe phenotypes (zoals ziekte) en hoopt bij te dragen aan de grotere ambities binnen de systeem genetica.

De verschillende stappen in een genetical genomics onderzoek worden onder de loep genomen (**Hoofdstuk 3**), maar eerst passen we deze methode toe in bloedcellen van muizen (**Hoofdstuk 2**). De resultaten tonen dat effecten van DNA variatie op gen expressie sterk beinvloed worden door de huidige cellulaire ontwikkelingsfase, wat de noodzaak aantoond om rekening te houden met cellulaire ontwikkelingfases. In **Hoofdstuk 4** wordt een nieuwe permutatie strategie gepresenteerd, die bescherming bied tegen het foutief interpreteren van eQTL hotspots.

Verder wordt methodiek geintroduceerd voor het analyseren van differentiele coexpressie, complementair aan klassieke differentiele expressie analyse. (**Hoofdstuk 5**)

Het ontraffelen van de rol van genen, eiwitten of metabolieten bij een specifiek phenotype kan niet zonder de causale verbindingen daartussen. De kracht en onmacht van huidige analyse methoden voor causale inferentie binnen Genetical Genomics wordt onderzocht (**Hoofdstukken 6 en 7**).

Tot slot, een discussie over de huidige en toekomstige uitdagingen binnen de systeem genetica zoals experimenteel ontwerp, statistische significantie en genetische netwerken.

Curriculum Vitae

Bruno Tesson was born in Bourges, France, in 1983. He completed his secondary education in 1999, and went on to study Mathematics and Physics in preparatory classes in Orléans. In 2002, he was admitted into the Ecole Nationale des Télécommunications de Bretagne (ENST Bretagne, Brest, France), an engineering school in the field of information technologies. In 2006, he graduated from ENST Bretagne as a Telecommunication Engineer. In parallel, he obtained a Masters in Bioinformatics from Chalmers University of Technology (Gothenburg, Sweden). In May 2006, he started his PhD project at the Groningen Bioinformatics Center of the University of Groningen under the supervision of Prof. dr. R.C. Jansen and Prof. dr. R. Breitling. The focus of the project was on systems genetics, and more precisely, he worked on bioinformatics approaches using natural genetic variation to gain insights into the biological mechanisms that determine complex phenotypes. In November 2010, he joined Institut Curie (Paris, France) where his research is about the identification of novel therapeutic targets for breast cancer through integrative analysis of genomic, transcriptomic and proteomic data.

Publications and conference presentations

Li Y*, **Tesson BM***, Churchill GA, Jansen RC. *Critical reasoning on causal inference in genome-wide linkage and association studies.* **Trends Genet.** 2010 Dec;26(12):493-8

Tesson BM, Breitling R, Jansen RC. *DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules.* **BMC Bioinformatics**. 2010 Oct 6;11:497.

Tesson BM*, Li Y*, Breitling R, and Jansen RC. *Scaling up classical genetics to thousands of molecular traits: promises and challenges.* **Proceedings of the 9th World Congress of Genetics Applied to Livestock Production**. August 2010.

Kooistra SM, van den Boom V, Thummer RP, Johannes F, Wardenaar R, **Tesson BM**, Veenhoff LM, Fusetti F, O'Neill LP, Turner BM, de Haan G, Eggen BJ. *Undifferentiated embryonic cell transcription factor 1 regulates ESC chromatin organization and gene expression.* **Stem Cells**. 2010 Oct;28(10):1703-14.

Swertz MA, Velde KJ, **Tesson BM**, Scheltema RA, Arends D, Vera G, Alberts R, Dijkstra M, Schofield P, Schughart K, Hancock JM, Smedley D, Wolstencroft K, Goble C, de Brock EO, Jones AR, Parkinson HE; Coordination of Mouse Informatics Resources (CASIMIR); Genotype-To-Phenotype (GEN2PHEN) Consortiums, Jansen RC. *XGAP: a uniform and extensible data model and software platform for genotype and phenotype experiments.* **Genome Biol.** 2010;11(3):R27.

Gerrits A*, Li Y*, **Tesson BM***, Bystrykh LV, Weersing E, Ausema A, Dontje B, Wang X, Breitling R, Jansen RC, de Haan G. *Expression quantitative trait loci are highly sensitive to cellular differentiation state.* **PLoS Genet.** 2009 Oct;5(10):e1000692.

Tesson BM, Jansen RC. *eQTL analysis in mice and rats.* Methods Mol Biol. 2009;573:285-309.

Jansen RC, **Tesson BM**, Fu J, Yang Y, McIntyre LM. *Defining gene and QTL networks.* **Curr Opin Plant Biol.** 2009 Apr;12(2):241-6.

Breitling R, Li Y, **Tesson BM**, Fu J, Wu C, Wiltshire T, Gerrits A, Bystrykh LV, de Haan G, Su AI, Jansen RC. *Genetical genomics: spotlight on QTL hotspots.* **PLoS Genet.** 2008 Oct;4(10):e1000232.

**: equal contributions*

Systems genetics of hematopoietic differentiation. Talk at **Systems Genetics: from man to microbe, from genotype to phenotype.** Oct 1st 2009, Groningen, Netherlands

A multi-cell type genetical genomics approach to study cell fate decisions during hematopoietic development. Talk at the **7th Annual Complex Trait Consortium Conference.** June 2nd 2008, Montreal, Canada

Acknowledgements

This thesis would not have been possible without the support of many people around me and I would like to thank them with a few words.

Dear Ritsert, your continuous support and trust have been invaluable to me. Your guidance has been critical to the success of this PhD and I hope we will have more fruitful collaborations in the future.

Dear Rainer, our many discussions and your always insightful advices have been tremendously stimulating. I have truly benefitted from your ideas and from your enthusiasm and it is comforting to know that your suggestions are always only one email away.

Dear Yang, I am very glad to have had the opportunity of working with you. In the process, I have not only gained a valuable scientific partner from whom I have learned a lot, but also a great friend.

Dear Klazien, apart from being always very helpful and incredibly efficient, your friendly presence has always been crucial in making our work environment welcoming.

Dear Elena, Richard, Tauqeer, Gonzalo, Frank, Nino, Andris, Morris, Danny, Marnix, Anna, Lying, Rudi, Martijn, Jingyuan, Joeri, René, George, Erik and all my other colleagues from GBiC which are too many to name here, thank you for all the good times we shared and the help that I have received from you. I also want to thank my collaborators from the Stem Cell Biology group at the University Medical Center of Groningen, Alice Gerrits, Dr. Leonid Bystrykh and Prof. dr. Gerald de Haan, as well as Prof. dr. Gary Churchill from the Jackson Laboratory.

Dear Chalmerists who have shared the same PhD plight: Christin, Darima and Tejas, thanks for being great friends that have been a constant source of comfort through good and bad times.

Dear friends that I met in Groningen, thanks for making my stay in this lovely city so full of good memories, you are too many to name, but you know who you are!

Finalement, je remercie du fond du cœur ma nombreuse famille dont l'affection et le soutien sans faille sont des richesses inestimables.