Chapter 1
Introduction: Generalized Genetical Genomics

1.1 Introduction

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

1.2 Concepts of generalized genetical genomics

1.2.1 Multifactorial experimentation

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

This has important consequences for genetic strategies for studying molecular networks, including genetical genomics or expression genetics (Jansen and Nap 2001, Jansen 2003, Jansen and Nap 2004, Rockman and Kruglyak 2006, Haley and
de Koning 2007). Genetical genomics measures molecular phenotypes, such as gene expression, protein abundance or metabolite levels, in many genetically diverse individuals and uses classic quantitative trait mapping to identify the underlying regulatory influences (see Box 1.1 for a brief outline of this concept). However, the resulting molecular network will be specific for a single experimental condition (e.g. one species, one tissue type or one physical condition).

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

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

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

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

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

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

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

Table 1.1: Molecular profiling technologies

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

1.2.2 Environments that matter

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

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

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

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

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

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

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

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

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

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

However, the cellular environment also changes in a more subtle way without external intervention. For instance, different tissues have widely different functions and molecular profiles. One would expect the underlying network structures to be distinct. Each tissue will have its own susceptibility to genetic polymorphisms. A mutation in an oncogene could, for example, lead to upregulation of cell proliferation genes in one tissue but not in another. A generalized genetical genomics approach can determine how variable such heritable differences are across several tissues (Figure 1.1d).
Furthermore, molecular networks will also change along differentiation trajectories, and different genotypes will show different dynamics of molecular traits during development (Figure 1.1e). One QTL might control stem cell genes during initial lineage commitment, whereas another QTL might control the same set of genes during terminal differentiation.

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

1.2.3 Controlled environmental perturbation

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

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

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

Figure 1.2: Designing a genetical genomics experiment with multiple environmental factors. (a) The optimal allocation of individuals with different genotypes to environmental conditions is a challenge in genetical genomic experiments. (i) We show 150 recombinant inbred lines (RILs) as balls, with the colors referring (in a simplified way) to genotypes: the more similar the colors the more similar the genetic fingerprint of two RILs. Each cell in the table represents a combination of different levels of two environmental factors (indicated by different levels of grey). (ii) The proposed strategy is to allocate 100 different RILs into four environments evenly, measuring 25 different RILs in each condition, keeping the microarray effort unchanged and leaving 50 RILs unused.

(b) Simulation results comparing QTL resolution for three different strategies: using 25 lines in a single condition (green); using the same 25 lines in each of the four conditions (black) and using a total of 100 lines, 25 different ones in each condition (red). A single QTL at the 10th marker was simulated. The support intervals (1.5 drop-off) are indicated by the bars, showing that using 100 different lines dramatically improves the mapping resolution.
special interest, such as a previously detected phenotypic QTL, or across the entire genome. We note that by minimizing the sum of the variances of the parameter estimates of interest (A-optimality (Kerr and Churchill 2001, Wit et al. 2005a)) using an optimization algorithm, such as simulated annealing (Wit and McClure 2004, Kirkpatrick et al. 1983), an optimal allocation can be found (Fu and Jansen 2006). In the optimization, the experimenter can, of course, give less weight to parameters of lesser interest, which will then be estimated with lower accuracy. For example, if the emphasis is on one or more genome regions of special interest, parameters for the markers in these regions can be given full weight in the optimization algorithm, whereas parameters for other markers can be given lesser or even zero weight. As a result, mapping resolution can improve (Figure 1.2b) and the power for finding QTLs can be increased (Figure 1.2c).

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

First, the resolution and power of QTL mapping depends on the number of genetically different samples in linkage and association studies. From this point of view, it would be wise to include as many genetically different samples as possible and not to replicate them across environments, because more recombination events will be observed and rare alleles are more likely to be present in the samples (Darvasi and Soller 1997). Figure 1.2b compares the resolution available from different design strategies: it is obvious that maximizing the genetic diversity leads to the sharpest QTL peak with tightest support interval and thus implicitly the most specific list of candidate regulators.
Let us assume there are 150 recombinant inbred lines (RILs) available and 100 single-color arrays can be used to measure the genome-wide expression level (Figure 1.2a), and that there are two different environmental factors, such as drug treatment (factor 1) and pathogen exposure (factor 2). Each factor has two different levels: different amounts of drug and low versus high pathogen exposure.

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

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

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

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

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

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

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

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

(c) The challenge of batching. In this hypothetical example, a group of crossed plant lines with different genotypes (as shown by grey versus white flower centers) are collected, and transcript abundance is measured (as indicated by the gradient of red flowers petals). It is clear that genotype correlates with gene expression level, and this can be mapped as an eQTL, indicating that a specific genetic locus plays an important role in deciding the gene expression...
maximize the number of informative genotype differences at one or more regions of particular interest, such as a previously detected phenotypic QTL, or to maximize the number of informative genotype contrasts across the entire genome. Figure 1.2c compares random and optimal selection. One can see that it is possible to achieve an almost balanced genotype ratio even in unfavorable circumstances.

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

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

1.2.4 Challenges: controlling the uncontrolled

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

Figure 1.3: (continued) level. However, if on closer examination, the samples were collected at different times of the day and, accidentally, samples with one genotype at a particular locus were more common at early and very late time points, this unintentional correlation of genotype with an uncontrolled external factor (such as daylight) would lead to a spurious mapping result for those genes that vary in expression in a diurnal fashion. (The observed expression difference between genotypes is confounded by the effect of daylight on expression). In many real life cases, the uncontrolled factor is much less obvious than in this example and will therefore be very hard to detect with sometimes dramatic consequences (see Ref. (Alberts et al. 2005) for an example of how spurious linkage between batch and certain genome regions can lead to ghost regulators being detected). Therefore, careful randomization and statistical consideration of batch effects is essential (Akey et al. 2007).
fluence the interpretation of generalized genetical genomics studies. For example, if samples are collected at the same developmental stage using a morphological or molecular selection marker, any unanticipated interaction between genotype and the selection marker can lead to spurious results (Figure 1.3a,b). To overcome this, one can include time or stage of sampling in the experimental design as an additional factor to check for potential interaction patterns. Other hidden factors can influence the molecular profiling itself. For example, if the samples were profiled environment-by-environment, any unanticipated interaction between the genotype and other uncontrolled factors changing during the profiling could also lead to spurious results (Figure 1.3c). Here one might want to include the order of profiling in the experimental design as an additional factor at multiple levels and use complete or incomplete block designs to eliminate any imbalance between factors of interest and the blocks (e.g. based on block-wise profiling of ten samples). In addition, methods such as surrogate variable analysis (Leek and Storey 2007) can be applied for detecting such uncontrolled environmental or genetic factors in an ANOVA model. As in the case for controlled environmental perturbations, including uncontrolled factors in the QTL model will not lead to a significant loss of QTL mapping power but will result in more robust interpretations.

### 1.2.5 Concluding remarks and future perspectives

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

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

We look forward to others enriching their genetical genomics experiments by
1. Introduction: Generalized Genetical Genomics

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

1.3 Thesis contribution and organization

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

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

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

Generalized genetical genomics has already been applied in a number of biological studies in C. elegans, yeast (Smith and Kruglyak 2008) and mouse. Chapter 5 describes the first genome-wide genetic study of gene expression plasticity by generalized genetical genomics and investigates whether environment-induced plastic responses of gene expression show heritable differences. We used recombinant inbred lines of the nematode worm C. elegans that were derived from parental lines
originally collected in Bristol (UK) and Hawaii (USA), and measured genome-wide gene expression at two different temperatures. Quantitative trait locus mapping uncovered genes with genetically determined differences in their plastic response to temperature changes, and a majority of them were found to be regulated by genes at a different genome position (regulated in \textit{trans}).

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

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

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