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.

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

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 a 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.
Figure 2 - Systems genetics: an integrative strategy.

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
1.5 References

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.


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.


