1 Introduction

1.1 An overview of gene expression

Gene expression is a fundamental cellular process by which a gene gives rise to a functional product and thus produces an observable phenotype. The phenotypic manifestation of genes is ensured by the synthesis of proteins and functional RNA molecules (e.g. rRNAs, tRNAs, microRNAs). These products are the consequence of a regulated flow of genetic information happening almost exclusively one way (Crick 1958): information is first transferred from DNA to RNA (transcription) followed by the transfer of information from RNA to protein in the case of protein-coding genes (translation) (Figure 1.1). A brief overview of these processes is presented in the following section.

Transcription

RNA polymerase II transcribes all eukaryotic protein-coding genes. To initiate transcription, the enzyme requires a set of additional proteins (transcription factors -TFs) which guide its positioning at the promoter and aid in pulling apart the two DNA strands, one of which acts as a template for RNA synthesis. The assembly of the transcription initiation machinery onto DNA is facilitated by the concomitant recruitment of chromatin-modifying enzymes, allowing access to the tightly chromatin-packaged DNA molecule. Following transcription initiation, other TFs guide the RNA polymerase into elongation mode. The single stranded pre-mRNA (primary transcript including both exons and introns) is synthesized in a 5' to 3' direction by adding ribonucleoside monophosphate residues to the free hydroxyl group at the 3′ end of the growing RNA chain (Strachan and Read 2004).

RNA processing

Eukaryotic transcription elongation is tightly coupled to RNA processing (McCracken, Fong et al. 1997). The first modification of the pre-mRNA is the addition of a 5' cap (a modified guanine) to the emerging transcript. This ensures that the cell distinguishes mRNAs from other types of RNA molecules and aids their transport to the cytosol. Following this processing step, introns are excised from the pre-mRNA by endonucleolytic cleavage and exons joined together through the process of RNA splicing.

Alternative splicing - mediated by a large RNA-protein complex called the spliceosome - can give rise to various polypeptide products (isoforms) resulting from different combinations of joined exons. This ability to produce multiple proteins from the same gene increases immensely the coding potential and complexity of eukaryotic genomes (Modrek, Resch et al. 2001; Modrek and Lee 2003). The 3' end of the RNA molecule is also processed, by the addition of a stretch of ~200 A nucleotides (poly-A tail), which helps direct the synthesis of the protein on the ribosome.

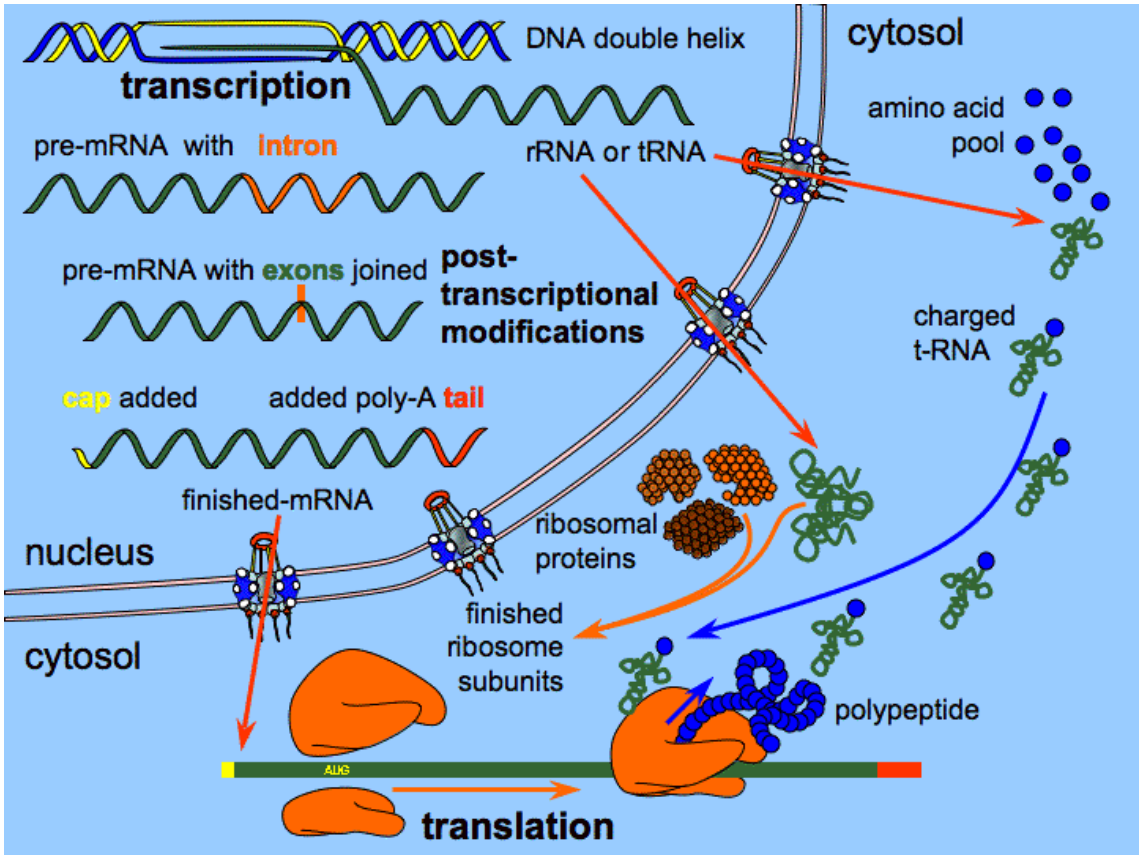

Figure 1.1 The process of eukaryotic gene expression. Inside the nucleus, RNA in transcribed from the DNA template into a primary RNA molecule (pre-mRNA). The pre-mRNA synthesis is followed by a series of processing steps including removal of intronic sequences (RNA splicing), 5' capping and 3' polyadenylation. The resulting processed mRNA molecule is exported into the cytoplasm where it engages with the translational machinery to give rise to the corresponding protein product. Modified from http://plantphys.info/plant_physiology/basiccytology1.shtml.

mRNA transport and translation

For some genes (e.g. rRNA genes, tRNA genes), RNA is the final gene product. Selected mature mRNA molecules however are transported through the nuclear pore into the cytoplasm, where they interact with the translational machinery and engage protein synthesis. Typically, the core of the mRNA is translated, while the flanking 5' and 3' sequences (UTRs – untranslated regions) are copied from the terminal exons to assist the stable binding of the mRNA to the ribosome and start polypeptide synthesis (Strachan and Read 2004). The assembly of the polypeptide is achieved by decoding the mRNA sequence as dictated by the triplet genetic code (three successive nucleotide sequences – codons – specify the corresponding amino acids). The decoding process is mediated by tRNAs bearing specific trinucleotide sequences (anticodons) and covalently bound amino acids which are subsequently inserted in the growing polypeptide chain. Once a stop codon is encountered, translation is terminated and the complete polypeptide released. Post-translational modifications involve attachment of functional groups (e.g. phosphoryl, carbohydrate), proteolytic cleavages or changing the chemical nature of selected amino acids (Mann and Jensen 2003).

1.2 Mechanisms of gene regulation

The brief overview of gene expression presented above highlights the complexity of the process and the multitude of steps involved in its completion. Any of these steps can be regulated to ensure the proper functioning of cells. Specifically, a cell can control the gene products it makes by (a) controlling when and how much of a given gene is transcribed (*transcriptional control*), (b) controlling how the RNA transcript is processed and spliced *(RNA processing control)*, (c) selecting which messenger RNAs are exported into the cytoplasm and where they should be localized *(RNA transport and localization control)*, (d) selecting which mRNAs should be translated *(translational control)*, (e) selectively destabilizing certain mRNA molecules *(mRNA degradation control)*, or (f) activating, inactivating or degrading specific protein products *(protein activity control)* (Alberts 2002). For most genes though, the most common regulatory control point is the initiation of transcription (Guenther, Levine et al. 2007). Given this, I emphasize below the most common points of transcriptional control and note that for the purpose of this thesis, transcript abundance (mRNA levels) was considered a proxy to gene expression.

1.2.1 Transcriptional control

Transcription of eukaryotic genes relies on two fundamental components: 1) stretches of defined DNA sequence in the gene's vicinity and 2) gene regulatory proteins (e.g. TFs) that recognize and bind to these sequences in order to recruit and activate the RNA polymerase. The sequence elements serving as recognition signals for the transcriptional apparatus are referred to as *cis*-acting, whereas gene regulatory proteins typically encoded elsewhere remotely in the genome (few megabases away from the gene or on another chromosome) are called *trans*-acting. Characterizing the full repertoire of regulatory elements is important and projects like the Encyclopedia of DNA elements (ENCODE - (Birney, Stamatoyannopoulos et al. 2007) have made important progress towards this goal. The pilot ENCODE project characterized in detail 1% of the human genome (~30 Mb) representing 44 carefully selected regions of variable gene content or containing functional elements revealed by comparative sequence analysis. The authors highlighted the pervasively transcribed nature of our genome by reporting that the majority of the human DNA sequence is represented in primary transcripts, many of which overlap considerably and include non-protein-coding regions. A multitude of new transcription start sites (TSS) was identified and the distribution of regulatory elements surrounding them was refined as being symmetrical, with no bias towards 5' regions as previously thought (Zhang, Paccanaro et al. 2007). The project offered also new insights into the relationship between chromatin structure and transcriptional control by showing that chromatin accessibility and patterns of histone modifications are predictive of transcriptional activity (Koch, Andrews et al. 2007).

1.2.1.1 *Cis* regulatory elements

The current standard view on transcription regulation involves the interplay of five major *cis*-regulatory elements (Maston, Evans et al. 2006):

1) Promoters, short stretches of DNA sequence immediately upstream of a gene, typically within 200 base pairs (bp) of the TSS. They are composed of different regulatory sequences (core promoter and nearby proximal regulatory elements) which function as a docking site for the basic transcriptional machinery (RNA polymerase and a set of general and promoter-specific TFs).

2) Enhancers, long-distance transcriptional control elements functioning as binding sites for activators, a class of TFs that increase the basal level of transcription initiated at the promoter. They control transcription in a spatial and temporal manner and are hence at

the basis of tissue-specific gene expression (Visel, Blow et al. 2009). A consensus DNA-looping model explains how the long physical distances between enhancers and the genes they regulate are overcome: the DNA between the enhancer and the core promoter loops out bringing the enhancer-bound proteins in proximity to the basal transcription complex.

3) Silencers, binding sites for TFs that reduce or repress transcription (repressors). They have been shown to act by blocking the binding of an activator (Harris, Mostecki et al. 2005), competing for an activator binding site (Li, He et al. 2004), or recruiting chromatin-modifying factors and thus blocking access to the promoter (Srinivasan and Atchison 2004).

4) Insulators, sequence boundary blocks (0.5-3 kb long) that prevent genes from being inappropriately regulated by neighbouring transcriptional elements. These DNA segments can preclude undesirable interactions between a distal enhancer and a promoter when situated in between the two (Geyer and Corces 1992; Kellum and Schedl 1992) or can act as barriers against spreading of repressive chromatin (heterochromatin), which might otherwise silence expression (Pikaart, Recillas-Targa et al. 1998).

5) Locus Control Regions (LCRs), groups of multiple *cis* regulatory elements acting upon an entire locus or cluster of genes (Li, Peterson et al. 2002). Each element in an LCR (enhancers, silencers, etc) affects expression differentially and only their cooperative activity determines its spatial/temporal expression properties. Moreover, LCRs seem to provide an open-chromatin domain for the gene cluster they regulate. DNase I hypersensitive sites - chromatin regions often preceding active promoters and having a high sensibility to cleavage by the DNase I nuclease - have been often observed in the proximity of LCRs (Lowrey, Bodine et al. 1992).

1.2.1.2 *Trans* regulatory elements

An abundance of distal protein regulators of transcription (*trans*-regulatory elements) further increases regulatory complexity. These proteins are mostly TFs (sequence specific DNA-binding proteins that mediate transcriptional activation or repression) or elements of chromatin modification complexes which assist the transcriptional apparatus to navigate through chromatin (Levine and Tjian 2003). TFs are broadly classified as general (factors such as the TATA box-binding transcription factor II D - TFIID - which assemble at the core promoter to form the preinitiation complex and are required for

transcription of most genes (Workman and Roeder 1987)) and tissue-specific (factors that ensure that certain genes are expressed only in certain tissues: e.g. the hepatic nuclear factor no 5 (HNF-5) modulating liver specific expression (Grange, Roux et al. 1991) or the epidermal-enriched factor KER1 (Leask, Rosenberg et al. 1990) controlling keratinocyte specific expression patterns).
Overall, the interplay of the multiple *cis* regulatory elements with the combinatorial activity of available TFs in specific chromatin-accessible genomic regions determines whether a transcript is being generated and if so, its level of steady-state expression (mRNA abundance). Mutations in any of these numerous components of the transcriptional machinery as well as any other post-transcriptional changes affecting mRNA stability (e.g. miRNA regulatory effects (Selbach, Schwanhausser et al. 2008)), splicing, cell signalling or protein-level modifications (Chen and Rajewsky 2007) influence gene activity.

1.3 Genetics of global gene expression

Gene expression underlies cellular and higher-order phenotypes by determining and maintaining proper transcript levels for each gene in a given cell-type. Understanding the genome-wide properties of regulatory control has been the focus of genetical genomics (Jansen and Nap 2001), a recent field of genetic analysis linking global gene expression with natural sequence variation. In this section, I present the main developments in the field and the methods employed to enhance the current knowledge on the genetics of gene expression.

1.3.1 Gene expression is a heritable quantitative trait

Despite their central role in shaping phenotypes, gene expression levels have been observed to differ significantly among individuals. These differences were first observed in model organisms such as yeast (Brem, Yvert et al. 2002) or mouse (Cowles, Hirschhorn et al. 2002), followed by similar observations in humans (Cheung, Conlin et al. 2003; Schadt, Monks et al. 2003; Morley, Molony et al. 2004).
Furthermore, evidence for familial aggregation of human expression profiles was found (Yan, Yuan et al. 2002; Cheung, Conlin et al. 2003; Monks, Leonardson et al. 2004), suggesting a heritable component of gene expression. Heritability estimates (h^2) capture the proportion of phenotypic variance among individuals in a population attributable to

genotypic differences. Therefore, evidence of heritability of a trait makes it amenable for genetic analysis. The lymphoblastoid transcriptome was the first to be estimated as variable among individuals, using pedigree analysis of samples from the Centre d' Etude du Polymorphisme Humain (CEPH) panel of lymphoblastoid cell lines (LCLs) (Cheung, Conlin et al. 2003; Monks, Leonardson et al. 2004; Morley, Molony et al. 2004). The initial heritability estimates differed between these studies, probably mostly due to small sample sizes and experimental artefacts that introduce additional expression variability. Nevertheless, they concurred that a large percentage of genes exhibit significant heritability levels (h^2). Monks etal. analyzed LCLs from 15 families and reported 762 genes of the 2430 (31%) differentially expressed as significantly heritable, with a median h^2 of 0.34 (Monks, Leonardson et al. 2004). Later on, Goring et al. examined expression in lymphocytes isolated from 1240 individuals from 30 large families and estimated that up to 85% of the 19,648 detected transcripts were significantly heritable (median h^2 of 0.23 among all expressed transcripts) (Goring, Curran et al. 2007). The authors also draw attention on the considerable influence on gene expression of environmental factors and the physiological state of an individual at the time of sample collection (e.g. time of blood draw). Overall, the studies above indicated that most transcript levels are influenced by an individual's genetic makeup and justified the upcoming efforts trying to identify the genetic determinants of gene expression variation.

1.3.2 Mapping expression quantitative trait loci (eQTLs)

The quantification of gene expression in numerous individuals from a population made it possible to treat the expression profile of each gene as a quantitative trait (Jansen and Nap 2001). This realization, together with the confirmation of a genetically determined component of gene expression, encouraged a series of efforts to map those regions of the genome that contribute to variation in transcript abundance (eQTLs) (Rockman and Kruglyak 2006).

Initially, small-scale experiments on the genetics of gene expression were performed. Allele-specific expression (ASE) assays confirmed that allelic differences in gene expression are common in autosomal non-imprinted genes (Yan, Yuan et al. 2002). Yan et al. compared relative expression levels of the two alleles for 13 genes in 96 individuals from the CEPH families. They observed allele-specific differences in six of the 13 tested genes, with a 1.3-4.3 fold difference between alleles. Reporter gene assays were also

informative with respect to the impact of genetic variation on gene expression (Hoogendoorn, Coleman et al. 2003). Hoogendoorn etal. screened for common polymorphisms the first 500 bp of the 5' flanking region of 170 genes and measured each promoter's ability to promote transcription in three human cell lines. The authors estimated that around a third of the promoter variants tested could significantly alter gene expression levels.

It was the development of microarray platforms however, that made it possible to shift from small-scale quantifications to genome-wide measurements, where transcript abundance of thousands of genes is determined simultaneously in a single experiment. These, combined with genetic variation information, allowed the identification of an abundance of loci with functional effects on gene expression. Two traditional approaches reviewed below have been used for eQTL mapping: linkage and association analysis (Hirschhorn and Daly 2005; Gilad, Rifkin et al. 2008).

1.3.2.1 Linkage mapping

Linkage mapping identifies genetic regions likely containing a causal variant by tracking the transmission pattern of chromosomes through families. The aim is to identify markers co-segregating with the trait of interest, as these are linked to the functional loci driving the phenotype. The main advantage of linkage studies is that they can be performed using a relatively low density of markers (<1000 microsatellites or slightly larger number of single nucleotide polymorphisms (SNPs) in humans) (Gilad, Rifkin et al. 2008). Some of the early genetical genomics studies identified regions controlling gene expression by using genome-wide linkage mapping in cell lines from individuals of the CEPH pedigrees (Monks, Leonardson et al. 2004; Morley, Molony et al. 2004). However, linkage mapping is only successful in detecting rare variants with high penetrance, such as those underlying monogenic 'Mendelian' disorders where the segregating causal allele is found in the same 10-20 cM region within each family (Hirschhorn and Daly 2005). In fact, in their study measuring expression of 23,499 genes in LCLs derived from 15 CEPH families, Monks et al. are only powered to detect eQTLs for 33 genes at a pointwise significance level of .000005. These are, as expected, large effects accounting for > 50% of the expression variance and having very high heritability (75% of the 33 eQTLs have a heritability > 0.76) (Monks, Leonardson et al. 2004). Typically, the regulatory regions uncovered by linkage mapping are also quite large. Morley etal. detect linkage peaks

>5Mb for approximately 1,000 expression phenotypes of the 3,554 expressed genes in LCLs from 14 CEPH families (Morley, Molony et al. 2004). Fine mapping of these large regions is very challenging and depends on the occurrence of recombination events within families. For detecting common variants (minor allele frequency (MAF) ≥ 5%) with smaller effect size on expression, association studies are much better powered and more suitable.

1.3.2.2 Association mapping

Association studies use phenotypes measured in collections of unrelated individuals and dense marker information from the same samples (typically > 500,000 genotyped SNPs in humans) in order to detect statistically significant correlations between marker genotypes and the analyzed trait (transcript abundance in the case of eQTL studies). Again, the assumption is that the causal locus is linked, or correlated with the markers showing statistical associations with the phenotype. The extent of this correlation is determined by linkage disequilibrium (LD), a property of the genome describing the non-random association of alleles at different loci in a population (Rockman and Kruglyak 2006). The preferential association of allelic combinations is reflected in the haplotype structure of the genome (Figure 1.2.), whereby a set of highly correlated genetic markers (LD blocks) undisrupted by recombination mechanisms are inherited together through generations (Paigen and Petkov 2010). The size of the LD blocks across the human genome is variable but nevertheless, they provide a much better resolution than linkage maps, with causal variants typically to be found within windows of few tens or hundreds of kilobases (kb) (Dawson, Abecasis et al. 2002).

Taking advantage of the correlation structure in the genome, the International HapMap project was launched in 2002 as a large-scale collaborative effort with the goal to identify and catalogue most of the common human genetic variation (Consortium 2003). The purpose of a detailed haplotype map (HapMap) of the human genome was to serve as a public resource of genetic markers and facilitate subsequent association studies with various phenotypes. The project was a success and its scale has been growing considerably throughout the years, reaching now its third phase.

Figure 1.2. The haplotype structure of the human genome. a) A short DNA sequence from the same chromosome is shown in four different individuals. The nucleotides are identical for most genomic positions except at three variable loci (SNPs) b) A particular combination of alleles observed in a population is called a haplotype. In this example, four haplotypes capture the sequence variation in the population of the DNA region in panel **a**. Only the 20 variable loci of the total 6,000 DNA bases represented are shown. c) To uniquely identify these four haplotypes, it is sufficient to genotype three tag SNPs out of the 20 variants. For example, Haplotype 1 can be recognized in any individual having the A-T-C pattern at the three tag SNPs. Typically, many chromosomes would carry the common haplotypes in the population. Figure adapted from the International HapMap Project, *Nature* 426, 789-796 (2003).

Phase 1 of the HapMap project aimed at genotyping at least one common SNP every 5 kb across the genome in each of the 269 samples belonging to four geographically distinct populations. Additionally, ten ENCODE regions, each 500 kb long, were sequenced in 48 individuals and all SNPs in these regions were genotyped in the full sample set The 269 samples consisted of: 90 individuals (30 parents-offspring trios) of northern and western European ancestry living in Utah from the Centre d'Etude du Polymorphisme Humain collection (abbreviated CEU), 90 individuals (30 trios) from the Yoruba in Ibadan, Nigeria (YRI), 45 Han Chinese from Beijing, China (CHB) and 44 Japanese from Tokyo, Japan (JPT). At this stage, a total of approximately 1.3 million SNPs were genotyped in each population (Consortium 2005).

In its second phase (HapMap 2), 270 individuals (the 269 Phase 1 individuals plus an additional JPT sample) were genotyped for a further 2.1 million SNPs. The resulting denser SNP map (approximately one common SNP per kb) contains an estimated 25-35% of the total 10 million SNPs expected across the human genome (Frazer, Ballinger et al. 2007).

The current and largest phase of the project (HapMap 3) involves an extension of the genotyped sample set, both by supplementing the initial four-population collection with more individuals and also by adding samples from seven other populations (http://hapmap.ncbi.nlm.nih.gov/). As such, over 4 million SNPs were genotyped from individuals of the Phase 1 and 2 populations (180 CEU, 90 CHB, 91 JPT, 180 YRI) and approximately 1.5 million SNPs were genotyped in 760 individuals of seven new populations (90 ASW: African ancestry in Southwest USA; 100 CHD: Chinese in Metropolitan Denver, Colorado, USA; 100 GIH: Gujarati Indians in Houston, Texas, USA; 100 LWK: Luhya in Webuye, Kenya; 90 MEX: Mexican ancestry in Los Angeles, California, USA; 180 MKK: Maasai in Kinyawa, Kenya; 100 TSI: Toscans in Italy).

In combination with large-scale expression data, these well-documented common genetic variation maps enabled the success of detecting eQTLs using population association studies (Cheung, Spielman et al. 2005; Stranger, Forrest et al. 2005). Building on their previous whole-genome linkage work, Cheung et al. used > 770,000 HapMap 1 SNP markers and mRNA abundance measurements from 57 CEU individuals to map eQTLs for previously identified expression phenotypes. Among the chosen 27 phenotypes with significant linkage evidence ($P < 3.7 \times 10^{-5}$), the authors confirmed 70% as having significant evidence of association ($P < 0.0001$). For all the concordant signals between the two methods, they were also able to narrow down the candidate functional regions, making thus use of the better resolution conferred by LD (Cheung, Spielman et al. 2005). Stranger etal. further explored the power of association studies by performing a genome-wide analysis of 630 genes in LCLs from 60 unrelated CEU individuals genotyped in HapMap 1. For the subset of 374 expressed genes, the authors detected eQTLs for up to 40 genes, with the majority of the eQTLs identified mapping in the proximity of the genes they associate with. Laying the ground for future genome-wide expression studies, Stranger etal. paid special attention to the multiple-testing problem and evaluated three statistical correction methods to reduce false positives, namely Bonferroni (Miller 1981), false discovery rate (FDR) (Storey and Tibshirani 2003) and permutations (Churchill and Doerge 1994). The Bonferroni method is prone to conservative estimates of significance since it does not account for the dependence of SNPs due to LD and treats each SNP – gene test as independent. The authors nevertheless report a generally good concordance among the different multiple-testing correction methods. Based on the highest enrichment of significant discoveries, the

results favoured permuting the expression values on the genotypes as a very suitable statistical correction strategy (Stranger, Forrest et al. 2005).

Besides multiple-testing, another caveat of genome-wide association mapping is the occurrence of false positives because of population substructure. In this case, allele frequency differences due to systematic ancestry differences between individuals having a dissimilar profile of interest (gene expression pattern or disease status) can cause spurious associations. Careful consideration of this issue lead to the development of appropriate statistical correction methods such as principal component analysis, modelling explicitly the inter-individual differences in ancestry prior to association testing (Price, Patterson et al. 2006). Altogether, these statistical advancements contributed unequivocally to the success of genome-wide association mapping.

In addition to single base variations (SNPs), structural DNA variations greater than 1 kb and present at variable copy number compared to the reference genome (copy number variants - CNVs) have also been successfully mapped. Initial observations suggested that CNVs are commonly present across the human genome and alluded to their substantial contribution to genetic variation in the population (Iafrate, Feuk et al. 2004; Sebat, Lakshmi et al. 2004). Consequently, their likely substantial effect on phenotypic variation resulting from gene dosage alteration, disruption of coding sequences or perturbation of long-range interactions (Kleinjan and van Heyningen 2005) elicited considerable attention. The contribution of CNVs to gene expression variation was assessed first in LCLs derived from the HapMap 1 samples (Stranger, Forrest et al. 2007). The association analysis between expression levels of 14,925 transcripts and correspondingly typed SNP and CNV variants in the 210 unrelated HapMap individuals revealed a series of *cis* effects: a total of 888 nonredundant genes associated with at least one SNP and 238 nonredundant genes associated with at least one autosomal CNV. The authors estimated that 83.6% of the expression variation is attributable to SNPs while CNVs capture 17.7% of the total expression variation in the current samples, with little overlap between them. Recently however, the higher CNV resolution enabled by tiling-array comparative genome hybridization (CGH) approaches indicates that CNV effects are much less dramatic than initially suspected. In conjunction with the denser SNP maps available, it was concluded that the contribution of common CNVs to

phenotypic variance (including thus mRNA levels) is already captured to a great extent by neighbouring SNPs (McCarroll, Kuruvilla et al. 2008; Conrad, Pinto et al. 2010).

1.3.2.3 *Cis* and *trans* eQTLs

One of the major advantages of eQTL mapping using the genome-wide association approach is that it permits the identification of new functional loci without requiring any previous knowledge about specific *cis* or *trans* regulatory regions. Typically in the eQTL mapping literature, regulatory variants have been characterized as either *cis* or *trans* acting, depending on the physical distance from the gene they regulate (Figure 1.3). In this thesis variants within one megabase (Mb) on either side of a gene's TSS were called *cis*, while those at least five Mb downstream or upstream of the TSS or on a different chromosome were considered *trans*-acting.

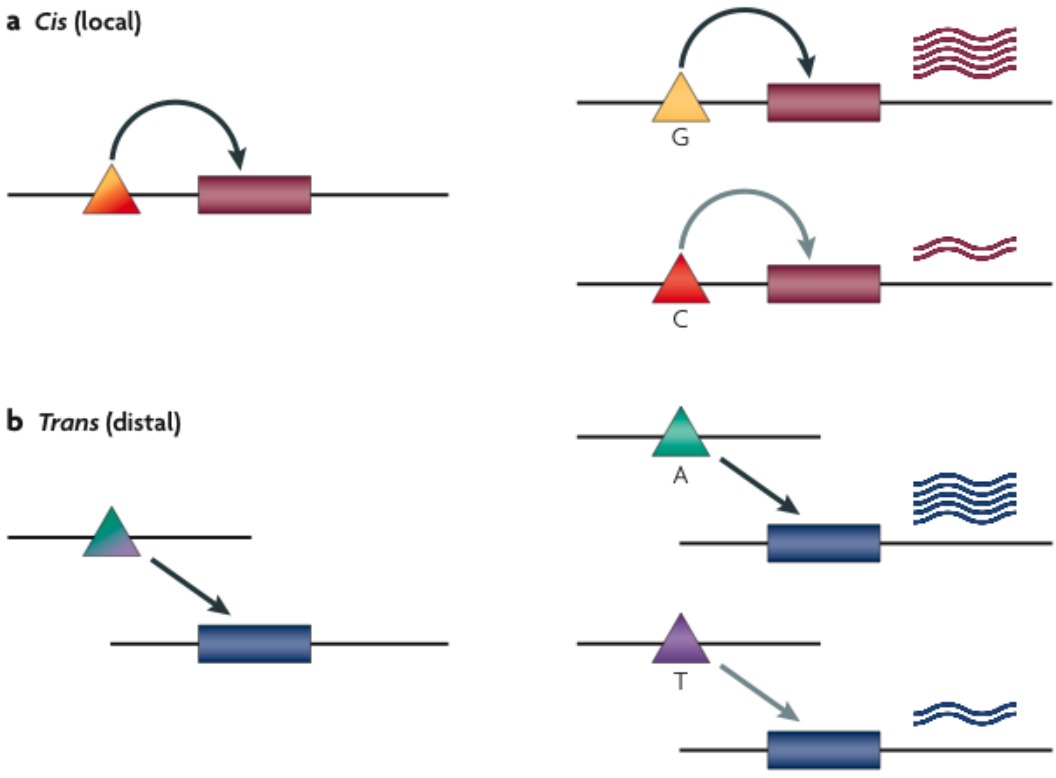

Figure 1.3. *Cis* and *trans* effects on transcript levels. Polymorphic regulatory variants (SNPs) affecting variation in a gene's transcript levels in *cis* (a) or in *trans* (b). The *cis* variant is located close to the gene it regulates. Individuals with the G allele at the *cis* eQTL have higher expression levels than individuals with the C allele. *Trans* variants are located at a much further genomic distance from the gene they regulate. There too, a particular allele (A in this case) drives high expression levels as opposed to the T allele determining lower mRNA levels. Modified from (Cheung and Spielman 2009).

Studies so far explain most of the variance in gene expression locally, by sequence variants in the vicinity of the associated genes. In a large-scale expression study where lymphocytes in 1,240 individuals were profiled, the authors identified 1,345 *cis*-regulated transcripts at an FDR rate of 5% of the total 19,658 tested (Goring, Curran et al. 2007). A study in our lab detected numerous *cis* effects in transformed B-cells. The analysis of 270 lymphoblastoid cell lines derived from the HapMap 2 individuals and genotyped for 2.2 million common SNPs revealed 831 genes of the 13,643 tested as having a significant *cis* eQTL (Stranger, Nica et al. 2007). Since power increases with the availability of larger sample sizes, the number of genes detected to have eQTLs is also expected to increase. Finding *trans* eQTLs has been less successful so far, mainly because interrogating the whole-genome for potential regulatory effects is a daunting statistical and computational task, requiring the correction for millions of tests. Whether the current enrichment of *cis* versus *trans* eQTLs reflects biological reality and is not just attributable to low power in *trans* is still under debate (Wray 2007; Wittkopp, Haerum et al. 2008).

1.3.3 Population differentiation of gene expression

Several studies have analyzed expression data in populations of different ancestry and revealed substantial differences at many loci. A study on 16 individuals of European and African descent estimated that 17% of genes were differentially expressed between populations (Storey, Madeoy et al. 2007). Differences were found also between European and Asian–derived populations for 1,097 of 4,197 genes tested (Spielman, Bastone et al. 2007). Larger scale studies confirmed the initial estimates. The eQTL study on 270 individuals of the four HapMap 2 populations of European (CEU), Asian (CHB, JPT) and African (YRI) descent reported that 17-29% of loci have significant differences in mean expression levels between population pairs (Stranger, Nica et al. 2007). While some of these observations are due to environmental factors (Idaghdour, Storey et al. 2008), genetics plays an important role in shaping the observed differences. Price etal. provide evidence for population differentiation due to genetic effects using cell lines derived from an admixed African American population (Price, Patterson et al. 2008). They estimated a mean value of 0.2 and a median of 0.12 in the proportion of gene expression variation attributable to population differences. A large proportion of the genetically determined variation in gene expression across populations has been

explained by different allele frequencies (Spielman, Bastone et al. 2007), suggesting that regulatory mechanisms are probably not fundamentally different between populations.

1.3.4 Multiple-tissue studies

So far, the majority of human eQTL studies have been performed exclusively on blood-derived cells or cell lines. This relatively easily accessible cell-type has been very useful in understanding the genetics of gene expression and continues to be a great resource. However, as gene expression signatures are cell-type specific (Alberts 2002), the question arises whether regulatory control of steady-state expression is also cell-type dependent. Estimates vary depending on the tissues being compared and the eQTL methods used, but generally, a significant tissue-specific component of *cis* regulation has been systematically reported.

Myers et al. analyzed for the first time the genetics of gene expression variability in the human brain. After expression profiling and genotyping 193 neuropathologically normal human brain samples, the authors estimated that 58% of the transcriptome is cortically expressed and identified significant eQTLs for 21% of the expressed transcripts (2,975 of the total 14,078 tested). A comparison of the cortical results with eQTLs previously identified in LCLs from CEPH individuals resulted in barely any overlap. While some degree of brain-specific control of gene expression is expected, the marked lack of overlap observed here is exacerbated by the different microarray platforms used and the distinct samples profiled in the two experiments (Myers, Gibbs et al. 2007). In a study comparing adipose and blood expression patterns between two Icelandic cohorts of considerable sample size (673 and 1,002 individuals respectively), 50% of the *cis* eQTLs detected were shared (Emilsson, Thorleifsson et al. 2008). Another study overlapping eQTLs identified in 93 autopsy-derived cortical tissue samples and 80 peripheral blood mononucleated cell samples outlined the distinct genetic control of expression in the two tissues, reporting <50% sharing (Heinzen, Ge et al. 2008). Finally, a study in our lab compared the regulatory landscape in three tissues (fibroblasts, LCLs and primary T cells) derived from the same set of 75 European individuals. Unlike previous studies, this unique dataset properly accounts for confounding factors such as differences in population samples, array platforms or statistical methods. The authors reported that 69-80% of *cis* eQTLs are cell-type specific, augmenting thus the need to study multiple tissues to determine the full spectrum of regulatory variants (Dimas, Deutsch et al. 2009).

1.3.5 Environmental and epistatic effects on expression

Gene expression is a complex trait shaped, in addition to genetic factors, by environmental conditions. Lifestyle (e.g. diet, smoking), geographic conditions or age have been shown to have a considerable impact on expression, sometimes even larger than that attributed to genetic effects (Idaghdour, Storey et al. 2008). Moreover, experimental treatment of cells can markedly change their expression patterns (Choy, Yelensky et al. 2008). Exposing cells to different perturbations also revealed that individuals differ in their response to external stresses (e.g. ionizing radiation) and lead to the identification of DNA variants that influence this differential response (Smirnov, Morley et al. 2009). Further such studies will be very useful for understanding the genetics of differential toxin response in view of improving drug administration. This has been the focus of pharmacogenomics, a field aiming to identify genetic determinants of drug response, i.e. polymorphisms influencing the activity of drug metabolizing genes (Evans and Relling 1999). Differential tissue and organ response to other disease relevant stimuli (e.g. insulin) is also likely to influence disease status and will need appropriate consideration in future association studies.

Interactions between genetic factors (Brem, Storey et al. 2005; Dimas, Stranger et al. 2008), but also between genetic and environmental factors (Gibson 2008) have an effect on gene expression as well. However, in the absence of good hypotheses for which particular combinations of DNA variants to test and under which model, detecting epistatic effects is statistically still very challenging.

1.4 Gene expression shapes cellular and high-order phenotypes

The role of gene regulation in shaping phenotypes has been pointed out early on, starting with the landmark paper of King and Wilson (King and Wilson 1975). Through comparative protein analysis, they were able to demonstrate that humans and chimpanzees are 99% genetically identical and thus suggested that significant species differences are probably due to gene regulatory variation. Since then, gene expression has been implicated in associations with a wide range of cellular and high-order phenotypes. A succinct review of a number of key examples certifying the role of gene expression variation in shaping phenotypes is presented below.

1.4.1 The role of expression in defining and maintaining cell-specificity

The control of gene expression is fundamental for the formation of specialized differentiated eukaryotic cells. Precise spatial and temporal gene regulation during development determines cell fate and helps maintain differentiated cell-specific signatures through subsequent cell generations (Alberts 2002). One example of a highly coordinated regulatory genetic switch mechanism is that involved in the formation of muscle cells during embryonic development. The development of muscle cells depends on the expression of myogenic proteins (MyoD, Myf5, myogenin, and Mrf4), a family of helix-loop-helix regulatory proteins. These bind to specific regulatory sequences surrounding muscle-specific genes and activate their transcription (Weintraub, Davis et al. 1991). Through a series of positive feedback loops, myogenic proteins further stimulate transcription of other gene regulatory proteins involved in muscle cell development. The deterministic role of gene expression in cell-type formation was additionally confirmed by observing the ability of myogenic proteins to trigger muscle differentiation in other cell-types (e.g. the human myogenic factor Myf5 induces myogenic phenotypes such as formation of multinuclei and synthesis of sarcomeric myosin heavy chains when transiently expressed in embryonic mouse fibroblasts (Braun, Buschhausen-Denker et al. 1989)).

Once cells differentiate into a certain cell-type, they remain specialized and transmit their specific expression signatures to daughter cells. This is attained by feedback loops wherein key gene regulatory proteins activate their own transcription or that of cell-type specific genes they interact with (Alberts 2002). Chromatin signatures also ensure the faithful propagation of cell-type specific expression, as unexpressed genes are packaged into compact chromatin forms, inaccessible to the transcriptional apparatus (Boyle, Davis et al. 2008). Specificity of gene expression in different cell-types has been extensively observed and Adams et al. were among the first to report this at a genome-wide scale (Adams, Kerlavage et al. 1995). In their study sampling 30 tissues with more than 1000 ESTs (expression sequence tags) each, they detected only eight genes matched by ESTs in all 30 tissues and 227 genes represented in at least 20 tissues. A following large-scale study on 46 human and 45 mouse tissues, organs and cell lines reported only 6% ubiquitously expressed genes and ascertained that tissue-specific gene clusters can be found in nearly all tissues examined (Su, Cooke et al. 2002).

1.4.2 Gene expression shapes complex phenotypes in the natural and disease range

Despite the essential role of regulatory control in ensuring normal functioning of cells, most biological systems are remarkably robust, showing abundant gene expression variation (Oleksiak, Churchill et al. 2002; Gilad, Oshlack et al. 2006). Much of the natural phenotypic variation including numerous adaptive features of various organisms has been associated with changes in gene expression. One of the early examples included the differential expression of the Hox gene *Ultrabithorax*, which has been shown to pattern fine hair outgrowths (trichomes) on the posterior femur of the second leg in *Drosophila* (Figure 1.4). The evolution of *cis*-regulatory elements of this gene, rather than the protein itself was indicated as responsible for these adaptive morphological changes (Stern 1998). Similarly, evolution of a *cis*-regulatory element in the *yellow* gene has been shown to contribute to the gain of a male-specific pigmentation spot in *Drosophila biarmipes* (Gompel, Prud'homme et al. 2005) while comparative analysis of expression patterns of growth factors in Darwin's Finches revealed that differential expression of *Bmp4* has a major role in determining beak morphology (Abzhanov, Protas et al. 2004).

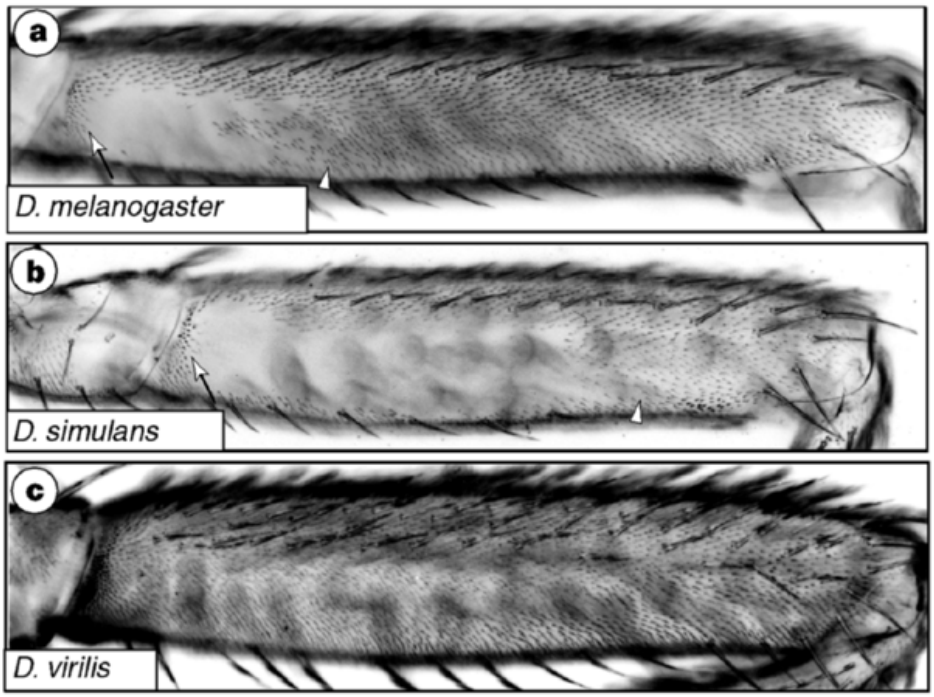

Figure 1.4. Different trichome patterns among *Drosophila* species. The posterior second femur of three *Drosophila* species is shown. The morphological differences observed are due to differential regulation of the Hox gene *Ultrabithorax.* Modified from (Stern 1998).

In humans, regulatory variation has been also associated with a series of phenotypic changes. A common well-studied example is that of lactase persistence (Ingram, Mulcare et al. 2009). Digestion of lactose, the sugar essential for nourishment of newborn mammals, is facilitated by lactase, a small intestinal enzyme encoded by the *LCT* gene. After weaning, the production of lactase decreases significantly in humans resulting in the inability to digest milk (lactose intolerance). Some humans however, are able to express *LCT* in adulthood (lactase persistence) and this has been especially observed in regions with traditional practice of milking (Figure 1.5), suggesting that the locus has been subject to strong positive selection (Holden and Mace 1997). The worldwide differential lactase expression has a genetic determinant and that was shown to be a *cis*-regulatory element (Wang, Harvey et al. 1995).

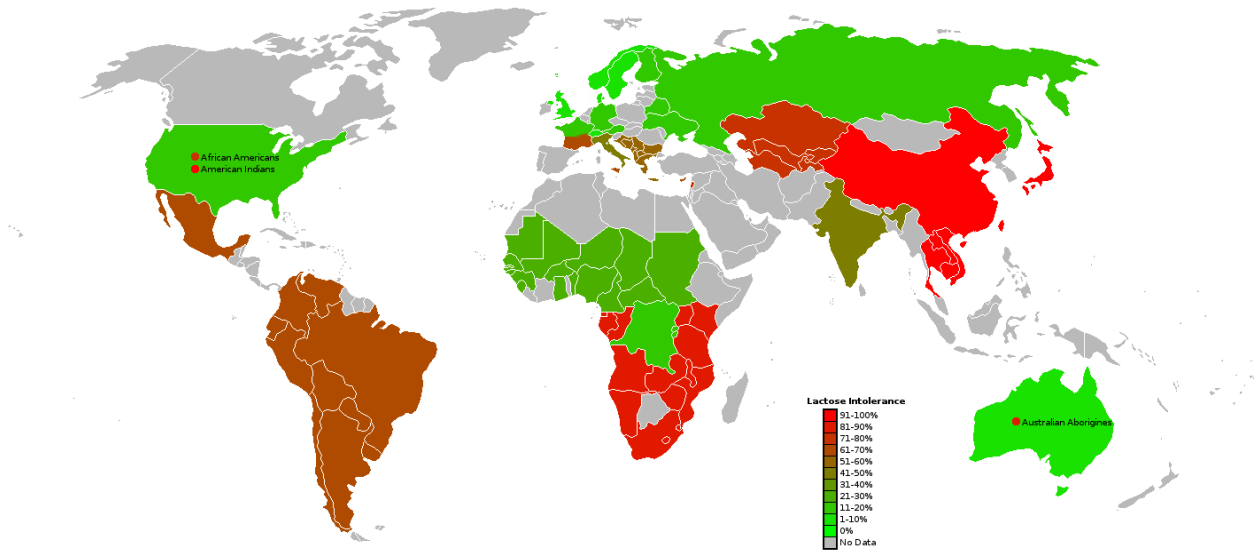

Figure 1.5. Geographic distribution of lactose intolerance. World map showing the distribution of lactose intolerance by region. Red indicates a high intolerance percentage and green a low percentage of lactose intolerance. The regions of low intolerance (or lactase persistence) coincide with areas of known cattle farming tradition. Image from http://en.wikipedia.org/wiki/Lactose_intolerance.

Gene expression changes beyond a tolerance limit can have more serious phenotypic consequences and prove detrimental. Decreased or complete loss of α-globin expression has been associated with α-thalassaemia (Weatherall 1998) and a regulatory SNP mapping in between the α-globin gene cluster and its upstream regulatory elements has been linked to the disease by causing significant down-regulation of the α^{D}, α2 and α1

genes (De Gobbi, Viprakasit et al. 2006). Low levels of adenomatous polyposis coli (*APC*) expression predispose to hereditary colorectal cancers (Yan, Dobbie et al. 2002) and over-expression of *C-MYC* (v-myc myelocytomatosis viral oncogene homolog) can lead to Burkitt's lymphoma (Boxer and Dang 2001). Specific regulatory polymorphisms inducing differential gene expression associated with complex traits have also been identified. Progression of coronary atherosclerosis has been associated with reduced expression of human stromelysin-1, regulated by a common *cis*-regulatory variant at the promoter (Ye, Eriksson et al. 1996). Susceptibility to autoimmune disorders has also been attributed to changes in expression: variants in the noncoding 3' region of the cytotoxic T lymphocyte antigen 4 gene (*CTLA4*) correlating with lower mRNA levels of a *CTLA4* splice variant were identified as disease determinant candidates for Graves' disease, autoimmune hypothyroidism and type 1 diabetes (Ueda, Howson et al. 2003). Taken together, these examples highlight the considerable range of phenotypic changes attributable to gene expression variation.

1.5 Genetics of complex diseases

Concomitant with the progress in understanding the genetics of gene expression, new insights into the genetic causes of common diseases have also been obtained. In the current section I briefly outline the developments that lead to this progress and the challenges that still exist in this area.

1.5.1 The road to genome-wide association studies (GWAS)

During the last twenty years, genetic mapping techniques allowed the elucidation of numerous rare monogenic disorders (Jimenez-Sanchez, Childs et al. 2001). Identifying genetic determinants of common diseases on the other hand, was lagging behind. The hitherto most common methods for uncovering disease genes, candidate gene approaches and family studies using linkage analysis, have been both failing to identify causative loci for complex traits.

Candidate gene studies are impractical since they are conditioned by often-unavailable information about disease biology. Even when a proposed relationship with a complex phenotype exists, finding causative variants within candidate genes has been unsuccessful, mainly because of the small sample sizes and the lenient statistical criteria by which associations were deemed causal. As such, many of the genotype-phenotype

associations reported based on candidate-gene approaches failed to replicate in independent studies (Ioannidis, Ntzani et al. 2001).

Linkage approaches were not far more successful, unsurprisingly given the properties of complex traits. Common complex disorders do not follow simple Mendelian inheritance patterns and are in addition characterized by multiple gene-gene and gene-environment interactions (Lander and Schork 1994). Each of these factors has a relatively small individual contribution to the determination of the ultimate disease phenotype. A number of additional aspects explain why it is improbable to find variants that impact common diseases and also co-segregate in families: (1) incomplete penetrance, whereby not all individuals inheriting the predisposing allele manifest the phenotype (2) locus and allelic heterogeneity, when mutations in any of several genes and different mutations within a gene respectively may give rise to the same phenotype or (3) pleiotropy, occurring when a single gene has multiple parallel phenotypic effects (Lander and Schork 1994). Under these circumstances, only few attempts to uncover complex disease loci using linkage analysis were successful (e.g. *NOD2* associated with Crohn's disease susceptibility (Hugot, Chamaillard et al. 2001; Ogura, Bonen et al. 2001)). Numerous other putative candidates were not further replicated, limiting in this way our understanding of complex diseases. In a classical paper in the field, Risch and Merikangas explained that the failure of replication was due to the limited power of linkage analysis to detect small genetic effects (Risch and Merikangas 1996). Thus, many of the proposed candidates were false positives and in fact, an unachievable sample size of more than 2500 families would be required to detect loci having low genotypic relative risk (typically ≤ 2 for complex disorders) with a minimum 80% power. Despite technical limitations at the time, the authors proposed genome-wide association studies (GWAS) as an alternative powerful approach.

1.5.2 The GWAS revolution

In the past few years, the ability of GWAS to help understand the genetic basis of complex disorders has become apparent (WTCCC 2007). The outburst of successful GWAS is owed to the availability of well-documented common human genetic variation maps (e.g. HapMap project (Frazer, Ballinger et al. 2007)), large patient samples with accurately recorded phenotypic information as well as appropriate statistical methods to assess significance (Rice, Schork et al. 2008). This approach has revealed a multitude of disease-susceptibility loci, now stored in an updated catalogue at the National Human

Genome Research Institute (NHGRI) (Hindorff, Sethupathy et al. 2009). For several common disorders such as type 1 (Hakonarson, Grant et al. 2007; Todd, Walker et al. 2007) and type 2 diabetes (Scott, Mohlke et al. 2007; Sladek, Rocheleau et al. 2007; Zeggini, Weedon et al. 2007; Zeggini, Scott et al. 2008), prostate cancer (Eeles, Kote-Jarai et al. 2008; Thomas, Jacobs et al. 2008) or inflammatory bowel disease (Parkes, Barrett et al. 2007; Rioux, Xavier et al. 2007), a multitude of predisposing loci has been reported. Most of these studies featured case-control designs, whereby a group of selected individuals diagnosed with a disorder of interest (cases) is compared to a group of people not ascertained for that phenotype (controls). The goal is to detect susceptibility alleles having marked frequency differences between the two groups. This design requires population stratification corrections and careful case/control selection in order to avoid misclassification biases (classification of cases as non-diseased), which can all decrease the power to detect associations (McCarthy, Abecasis et al. 2008).

Most recently, GWAS have been performed on population-based cohorts, offering insights into the genetics of continuous traits, be they anthropomorphic like height (Weedon, Lettre et al. 2007; Lettre, Jackson et al. 2008) or disease relevant (e.g. fat mass (Frayling, Timpson et al. 2007; Loos, Lindgren et al. 2008), lipids (Saxena, Voight et al. 2007; Willer, Sanna et al. 2008; Teslovich, Musunuru et al. 2010)). The discovery of multiple susceptibility variants per complex trait, each of small effect size, as well as their considerable pleiotropic overlap (Stratton and Rahman 2008) is gradually shifting the focus from viewing disease as a dichotomous trait towards a quantitative view, where common disorders are the extremes of a spectrum of quantitative traits (Figure 1.6) (Dermitzakis 2008; Plomin, Haworth et al. 2009).

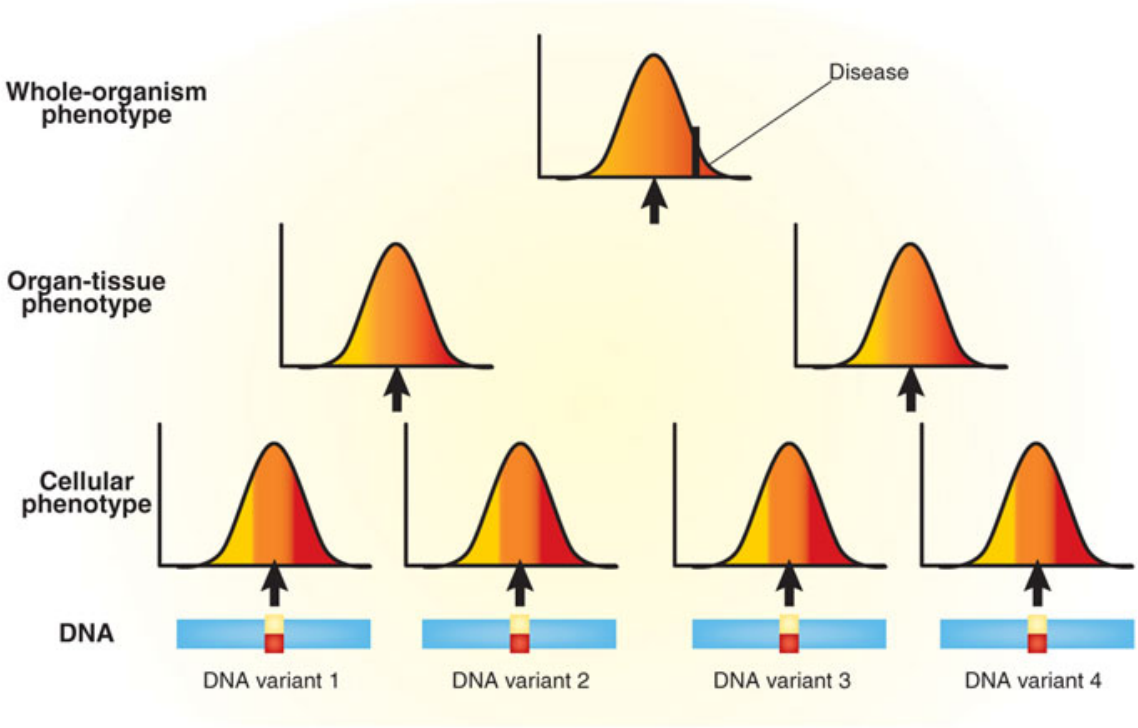

Figure 1.6. Common disorders as quantitative traits. Disease state can be viewed as the tail of a spectrum of continuous phenotypes**. I**n this schematic example**,** four unlinked DNA variants determine the whole-organism phenotype through changes at the cellular level, which in turn affect intermediate organ-tissue phenotypes. Yellow to red gradients represent the effect of each of the four red and yellow DNA variants at the different ends of the phenotypic spectra. At the cellular level, these effects are easy to interpret and to detect, whereas at organismal level, the power to detect them is reduced owing to the large number of direct and indirect intermediate interactions. Adapted from (Dermitzakis 2008).

A wide range of quantitative trait data of potential disease relevance is nowadays being collected and combined into large-scale meta-analyses. With greater power, such efforts identify disease affecting loci and the intermediate quantitative traits underlying them (e.g. Prokopenko etal. and Dupuis et al. look at fasting glucose levels as a continuous trait, find variants that associate with glucose concentrations and subsequently identify type 2 diabetes susceptibility loci (Prokopenko, Langenberg et al. 2009; Dupuis, Langenberg et al. 2010)).

1.6 Promise of eQTL studies for disease genetics

Despite the impressive success of GWAS, there is a substantial gap between the susceptibility variants discovered and understanding how those respective loci contribute to disease. Frequently such loci map to genomic regions of no apparent function (non-coding) or the genome's tight correlation structure (LD) does not permit firm conclusions about functional effects (i.e. which is the causal variant and which gene function does it

affect). Under these circumstances, the need to incorporate additional information for interpreting GWAS results became evident. The direct link between DNA polymorphisms (usually SNPs) and variable transcript levels along with the increasing role attributed to regulatory variation in shaping phenotypic differences, nominated gene expression as an important mechanism underlying complex traits. Subsequently, I describe the main results obtained so far in support of this hypothesis.

1.6.1 GWAS SNPs can be strong eQTLs

Comparing expression levels of individual genes between cases and controls may not be sufficiently powered to detect significant differences (Cookson, Liang et al. 2009) and discriminating between causal and reactive expression changes would be a tough challenge. However, genetic markers simultaneously associated with disease status and eQTLs are very interesting: if one allele is more frequent in cases than controls and at the same time it is causal for gene expression effects of a nearby gene, which is itself important for the disease, then it is likely that causality can be established. Several recent studies have shown the value of this principle by incorporating eQTL analyses with GWAS results and thus proposing candidate disease genes. Moffatt etal. identified a series of strongly correlated SNPs in a 200 kb region of chromosome 17q23 associated with childhood asthma (Moffatt, Kabesch et al. 2007). The association region contained 19 genes, none of which had an evident disease role. Expression analysis on lymphoblastoid cell lines derived from the same families showed that the most significant GWAS SNPs also explained ~29.5% of the variance in transcript levels of one of those 19 genes, *ORMDL3* (ORM1-like 3), now the best candidate for further functional studies. Expression data has helped interpret some of the association signals for Crohn's disease as well. Initial findings of a recent GWAS included multiple susceptibility loci mapping to a 1.25 Mb gene desert region on chromosome 5 (Barrett, Hansoul et al. 2008). eQTL data showed that one or more of these loci act as long-range *cis* regulators of *PTGER4* (prostaglandin E receptor 4), a gene 270 kb away from the associated region whose homologue has been implicated in phenotypes similar to Crohn's disease in the mouse (Libioulle, Louis et al. 2007). Other similar examples for height (Gudbjartsson, Walters et al. 2008), systemic lupus erythematosus (Hom, Graham et al. 2008), type 1 diabetes (Hakonarson, Grant et al. 2007) or bipolar disorder (WTCCC 2007) support the use of eQTL data in aiding the interpretation of GWAS results.

However, not all cases are so straightforward, as shown by the association of the *SH2B1* (SH2B adaptor protein 1) locus to body mass index (BMI) (Willer, Speliotes et al. 2009). In this case, a non-synonymous genome-wide significant SNP in *SH2B1* was associated also with differential expression of two other genes (*EIF3C* – eukaryotic translation initiation factor 3, subunit C and *TUFM* – Tu translation elongation factor, mitochondrial). Functional evidence from mice, where mutating a *SH2B1* homologue leads to extreme obesity (Ren, Li et al. 2005) and from humans, where a chromosomal deletion encompassing *SH2B1* associates with severe early-onset obesity (Bochukova, Huang et al. 2010) strengthen the hypothesis that the missense SNP is the actual functional variant. This SNP is then most probably in high LD with a different causal regulatory variant, which affects *EIF3C* and *TUFM* expression. This is a typical example of a coincidental overlap of GWAS and eQTL results, which must be carefully distinguished from causal cases where both the GWAS SNP and the eQTL tag the same functional variant (Figure 1.7).

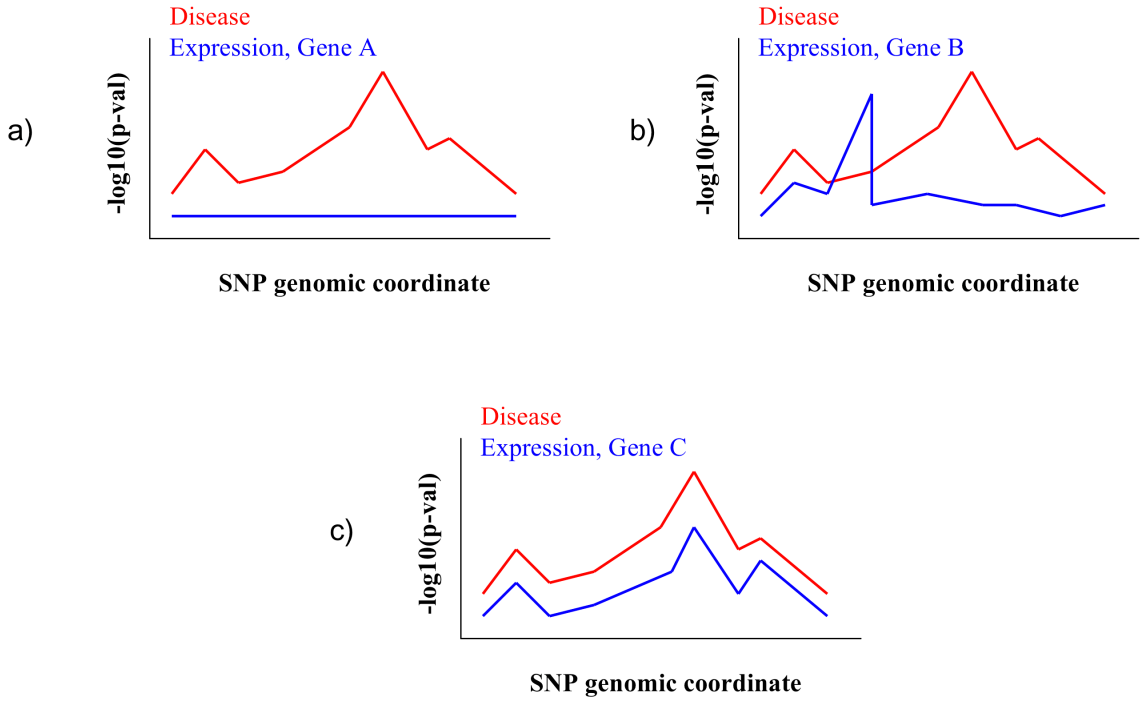

Figure 1.7. Interpreting GWAS results with eQTL data. Schematic representation of a genomic interval where same SNPs have been tested independently for associations with a disease (red) and transcript levels of a set of genes (blue). Three nearby genes are investigated for potential causal regulatory effects: (a) Gene A can be ruled out as it has no significant eQTL in the interval (b) The disease - associated interval harbours an eQTL for gene B, but the eQTL and the disease marker tag different functional variants (c) The GWAS SNP is a strong eQTL for Gene C and they likely tag the same functional effect as reflected by the similar association patterns at other tagging SNPs in the interval (Nica and Dermitzakis 2008).

Given the ubiquitous nature of regulatory variants (Stranger, Nica et al. 2007) and hence the high probability of such coincidental overlaps, integrative methods pinpointing true causal regulatory effects are desirable (Nica and Dermitzakis 2008). Nevertheless, since many traits manifest themselves only in certain tissues, such methods are only informative if expression measurements from disease-relevant cell-types are compared. Defining disease relevance of a cell-type is however yet another challenging task. The pathology of diseases is largely tissue-specific, but it remains mostly unknown how tissue-wide germline mutations lead to tissue-restricted disease effects (Lage, Hansen et al. 2008). Moreover, substantial overlap has been observed between pathways involved in progression of different diseases (Bentires-Alj, Kontaridis et al. 2006) and sometimes this overlap is not intuitive (Swanberg, Lidman et al. 2005; Torkamani, Topol et al. 2008). Therefore, confidently defining tissue relevance to a complex trait is yet unrealistic and expression datasets from seemingly irrelevant tissues should not be discarded at this stage, as they could be informative of disease biology.

1.6.2 Gene regulatory networks

The large-scale disease studies performed so far have uncovered multiple variants of small effect sizes affecting multiple genes. This suggests that common forms of disease are most probably not the result of single gene changes with a single outcome, but rather the outcome of perturbations of gene networks which are affected by complex genetic and environmental interactions (Schadt 2009). The numerous genetic factors involved in disease predisposition appear randomly distributed across the genome, but the expectation is that they are functionally linked and that these functional interactions are useful in prioritizing disease genes (Franke, van Bakel et al. 2006). DNA sequencing of tumour samples from pancreatic and brain cancer respectively, provided supporting evidence for this principle by identifying candidate genes belonging largely to core pathways involved in tumorigenisis or tumour progression (Jones, Zhang et al. 2008; Parsons, Jones et al. 2008). Recently, analysis of gene regulatory networks has offered important insight into complex disease mechanisms. In a study integrating co-expression networks and genotypic data from an F2 intercross population, Chen etal. identified a liver and adipose macrophage-enriched sub-network (MEMN) associated with metabolic syndrome relevant traits (Chen, Zhu et al. 2008). Three genes in this network, lipoprotein lipase (*Lpl*), lactamase β (*Lactb*) and protein phosphatase 1-like (*Ppm1l*) were validated by gene knockouts as causal obesity genes, strengthening the association of MEMN to

phenotypes characteristic to metabolic syndrome. A parallel study in humans identified a homologous transcriptional network constructed from adipose data, having substantial overlap with MEMN sub-modules and being enriched for genes involved in inflammatory and immune response (Emilsson, Thorleifsson et al. 2008). Subsequent eQTL mapping identified *cis*-regulatory variants affecting specific genes in this network and the joint analysis of the strongest *cis* eQTLs revealed substantial enrichment for variants associated to obesity related clinical traits. Classical genetic approaches would not be able to detect such variants with small individual effects. Identifying them as a group affecting gene networks which - when perturbed - result in a disease state, is in this case much better powered.

1.6.3 Candidate gene approach via transcriptome profiling

The major challenge when using transcriptome data for interpreting disease effects is distinguishing between causal and reactive changes in gene expression. An interesting approach to address this issue has been taken recently by Naukkarinen and colleagues, who use it to detect potential obesity candidate genes (Naukkarinen, Surakka et al. 2010). The authors made use of genome-wide expression data from adipose tissue and a unique collection of samples (a set of 13 monozygotic - MZ - twin pairs discordant for BMI) in order to devise an original candidate gene prioritization strategy. An additional cohort of 77 non-related individuals having a wide and representative BMI range had been profiled for adipose expression. The authors compared significant expression differences between lean and obese individuals in the MZ twins and the separate cohort, the rationale being that expression differences in the genetically identical twins represent likely reactive effects to obesity, whereas expression differences in the non-related individuals are a combination of causal and reactive determinants. The difference of the two sets would constitute a plausible collection of genes causally implicated in obesity risk. Variants in these genes (197 SNPs in 27 genes) showed a significant excess of low P-values when tested for association with BMI in a large cohort. Among the top associated SNPs, seven mapped to the same gene, F13A1 (coagulation factor XIII, A1 polypeptide), a newly proposed obesity susceptibility candidate. Variants in this gene were replicated in another independent cohort of ~2,000 samples, yet further validation of the associations with BMI is still required in larger independent cohorts. The choice of the candidate cell-type for expression quantification is an obvious issue. The authors acknowledge that for example, MC4R, a known obesity gene has been excluded, as it is

not expressed in fat. Given the poor candidate tissue knowledge, choosing a relevant tissue for such experiments is currently very challenging for many traits. Furthermore, identical twins discordant for a phenotype of interest, while very informative, are a rare sample collection. Nevertheless, this informed candidate gene strategy is an interesting example of how to identify further disease variants when limited by the sample sizes available. For many complex traits, identifying additional genetic associations beyond the initial 'low-hanging fruits' requires sampling of a vast set of individuals. For example, the six new loci associated with body mass index (BMI) recently reported by the GIANT consortium involved the analysis of more than 91,000 samples (Willer, Speliotes et al. 2009). Similarly large or larger sample sizes are unrealistic for other traits and novel strategies like the example presented here could be helpful for uncovering smaller genetic effects.

1.7 Thesis aims

The genetics of global gene expression has been extensively studied in recent years and it is now unquestionable that regulatory variants affecting transcript levels are ubiquitously distributed throughout the human genome. Concurrently, large-scale GWAS have shed light into the genetics of human complex traits and identified a multitude of susceptibility loci of modest effect-size each. Despite the statistical success in revealing DNA variant - trait associations, by themselves, these results alone don't necessarily lead to the identification of causal disease mechanisms. The goal of my thesis is to further the understanding of regulatory variation particularly with a focus on its role in complex diseases. Specifically, I address this by: (a) developing an empirical methodology (RTC) that directly combines eQTL and GWAS results in order to detect causal regulatory effects and prioritize candidate disease genes (Chapter 3) (b) exploring the complexity of regulatory tissue-specificity in multiple primary tissues derived from a set of twins (Chapter 4) (c) analyzing the implications of tissue-dependency in detecting causal regulatory effects for complex traits (Chapter 5). Taken together, the results presented here underline the informative value of expression phenotypes for explaining the biological properties behind genetic associations with complex traits and highlight the need to explore regulatory complexity in a variety of relevant cell-types.