1 INTRODUCTION

In this chapter I will:

- Define gene expression as the transfer of information from DNA to mRNA and then into protein.
- Explain that gene expression is a complex, quantitative trait with naturally occurring variation in its patterns and levels. These patterns have a key role in defining and maintaining cell types, and in shaping higher level phenotypes in the normal and disease ranges.
- Give a brief overview of the process of gene expression.
- Outline that this process can be regulated at many levels, the most important of which is transcription initiation which involves the action of cis and trans regulatory elements.
- Argue that a component of variation in expression levels is heritable and arises as a consequence of genetic variation in cis and trans-acting regulatory elements.
- Outline strategies employed to uncover genetic variants responsible for expression variation.
- State the aims of this thesis.

1.1 WHAT IS GENE EXPRESSION?

The process by which a gene gives rise to a functional product is called gene expression (Lewin 2008). Gene expression enables the phenotypic manifestation of genes, results in the production of a protein or a functional RNA molecule (e.g. rRNA, tRNA, microRNA) and is necessary for cells to operate.

Gene expression is a complex trait shaped by genetic (Monks, Leonardson et al. 2004; Morley, Molony et al. 2004; Cheung, Spielman et al. 2005; Stranger, Forrest et al. 2007; Stranger, Nica et al. 2007), epigenetic (Eckhardt, Lewin et al. 2006; Petronis 2006), and environmental (Gibson 2008; Idaghdour, Storey et al. 2008) factors. Interactions between genetic factors (Brem, Storey et al. 2005; Dimas, Stranger et al. 2008), as well as those between genetic factors and the environment (Gibson 2008) also affect the expression levels of genes. As a result, this phenotype exhibits continuous variation among individuals and has the properties of a quantitative trait (Dermitzakis 2008).

In this thesis I address protein-encoding genes whose expression can be divided in two stages: transfer of information from DNA to RNA (transcription) and transfer of information from RNA to protein (translation). A number of mechanisms control these processes, including chromatin condensation, alternative splicing, DNA methylation, transcription initiation, mRNA stability, translational control, post-translational control and protein degradation. In eukaryotic cells the most common point of control is transcription initiation (Stranger and Dermitzakis 2005). In the following sections I discuss the role of gene expression in shaping phenotypes, I give a brief overview of this biological process and outline how it is controlled.

1.2 Gene expression defines phenotypes

The idea that gene expression levels play a role in shaping phenotypes is not new. In her doctoral thesis, Marie-Claire King revolutionized evolutionary biology by proving, through the comparative study of proteins, that human and chimpanzee genomes are 99% identical. In their landmark paper, King and Wilson (1975, pg 107) stated that human and chimpanzee macromolecules “are so alike that regulatory mutations may account for the biological differences between these species”. In the following sections I discuss that naturally occurring variation in expression levels is widespread, that expression patterns define and maintain cell types and sculpt higher level phenotypes in the natural and disease ranges.

1.2.1 Naturally occurring variation in gene expression levels

Developmental biology was one of the first fields that recognised the importance of expression patterns in shaping phenotypes. Spatially and temporally regulated expression is critical for the complex developmental programmes that result in the highly specialised cell types of higher eukaryotes. Following development and differentiation, the control and maintenance of appropriate levels and patterns of gene expression are vital cellular processes. Although expression of genes in the right quantity range, at the right time and in the right place is largely responsible for normal functioning of cells, biological systems can also display remarkable robustness. In most species studied (including yeast, fruit flies, mice, and humans) ample tolerance of naturally occurring variation in expression levels has been detected (Hartman, Garvik et al. 2001; Jin, Riley et al. 2001; Brem, Yvert et al. 2002; Schadt, Monks et al. 2003; Stranger and Dermitzakis 2005; Stranger, Forrest et al. 2005; Boone, Bussey et al. 2007; Gibson 2008). In a cross between two strains of yeast for example, profound differences in gene expression were found, with nearly half (2,698 out of 6,215) of all the genes in the genome being differentially expressed (DE) (Brem, Yvert et al. 2002). A study exploring human natural gene expression variation in 16 individuals of European and African descent found that 83% and 17% of genes were DE among individuals and populations respectively (Storey, Madeoy et al. 2007). In another study, three populations of apparently healthy Moroccan Amazigh (Berbers) were found to differ for over a third of their leukocyte transcriptome (Idaghdour, Storey et al. 2008). Taken together these results demonstrate that naturally occurring variation in expression levels is widespread within species.

1.2.2 Gene expression patterns define cell type specificity

Although there is space for expression variation, temporal and spatial regulation of gene expression patterns is critical for defining cell types during development in higher eukaryotes. Furthermore gene expression patterns have a role in the maintenance of cellular and tissue function following differentiation. Some examples are discussed below.

1.2.2.1 Gene expression defines cell type during development

Mammalian skeletal muscles are highly distinctive cells whose differentiation is triggered by expression of specific myogenic proteins including MyoD, Myf5, myogenin and Mrf4. The potency of these proteins was highlighted in a study where their expression in skin fibroblasts triggered muscle differentiation (Alberts 2008). Eye development also illustrates the developmental role of gene expression. In *Drosophila*, mice and humans this process involves highly regulated expression patterns of the gene *Ey* (*Drosophila*) and *Pax-6* (vertebrates). *Ey* expression triggers the formation of a specific cell type, but also of an entire organ, composed of different cell types and arranged in three dimensional space (Alberts 2008).

The critical role of expression patterns even for very closely related genes was demonstrated in an experiment where the transcription factor (TF) gene *SOX10* was deleted in mouse embryos and replaced with *SOX8*, a closely related gene that has overlapping expression patterns (Kellerer, Schreiner et al. 2006). SOX10 has a role in neural crest development and is defective in the human Shah-Waardenburg syndrome. It is essentially expressed in neural crest derivatives that form the peripheral nervous system and in the adult central nervous system (Bondurand, Kobetz et al. 1998). Both genes perform very similar functions and regulate processes such as enteric nervous system development and oligodendrocyte differentiation. Despite their similarities, SOX8 phenotypic rescue of SOX10 deficiency was variable for different tissues: development of glial cells and neurons in the sensory and sympathetic parts of the peripheral nervous system was almost normal, but melanocyte development was as defective as in SOX10-deficient mice. Furthermore rescuing of defects in enteric nervous system development and oligodendrocyte differentiation was limited. These results highlight the importance of tissue-specific gene expression and demonstrate that the extent of functional equivalence depends on cell type (Kellerer, Schreiner et al. 2006).

1.2.2.2 Cell type-specific patterns of gene expression in differentiated cells

Once cells have undergone differentiation, expression profiles remain critical for maintenance of cellular function. In one of the first studies to explore genome-wide expression signatures, over 1,000 expressed sequence tags (ESTs) were sampled in 30 tissues (Adams, Kerlavage et al. 1995). Substantial tissue specificity of gene expression was detected with only eight genes sharing ESTs across all tissues, and 227 genes being represented in at least 20 tissues. A subsequent study interrogated transcription levels in 46 human and 45 mouse tissues, organs, and cell lines spanning a broad range of biological conditions (Su, Cooke et al. 2002). Only 6% of the genes interrogated were found to be ubiquitously expressed and hierarchical clustering identified groups of genes with specific expression patterns in nearly all tissues examined. Another study explored expression patterns of human orthologue genes from chromosome 21 in mice using RNA *in situ* hybridization and reverse transcriptase polymerase chain reaction (RT-PCR) (Reymond, Marigo et al. 2002). Patterned expression was observed in several tissues including those affected in trisomy 21 phenotypes (central nervous system, heart, gastrointestinal tract, and limbs). Taken together these examples underline that gene expression is a phenotype displaying extensive cell type and tissue specificity.

1.2.3 Gene expression shapes normal range phenotypes

Variation in expression levels is to a large extent responsible for shaping higher level phenotypes in the normal range. In *Drosophila* expression variation in the developmental gene *Svb* underlies trichome pattern differences between species (McGregor, Orgogozo et al. 2007). Expression patterns of the Hox gene *Ubx* outline trichomes on the posterior femur of the second leg (Stern 1998), and male-specific wing pigmentation spots in *D. biarmipes* are a consequence of varying expression levels of the yellow pigmentation gene *y* (Figure 1 a) (Gompel, Prud'homme et al. 2005). In *Geospiza* (Darwin's finches) diverse beak shape and morphology is in part due to expression differences of the gene *Bmp4* (Abzhanov, Protas et al. 2004). Expression patterns in the mesenchyme of upper beaks correlates with beak morphology (Figure 1 b) and when misexpressed in chicken embryos, Bmp4 causes morphological transformations that parallel the beak morphology of the large ground finch *G. magnirostris.*

The predominant differences in branching patterns in domesticated maize (*Zea mays mays*) and its wild ancestors, the teosintes (*Z. mays parviglumis* and *mexicana*) arise in part from expression differences of the gene *tb1* (Clark, Wagler et al. 2006). In *Gasterosteus aculeatus* (sticklebacks), expression variation of the developmental gene *Pitx1* in pelvic and caudal fin precursors results in pelvic reduction and major skeletal changes (Shapiro, Marks et al. 2004). Modified gene expression levels of the prairie vole gene *V1aR* give rise to differences in receptor distribution patterns in the brain. This is thought to affect a range of socio-behavioural traits, including social recognition and investigation, social odour tasks and parental care (Hammock and Young 2005).

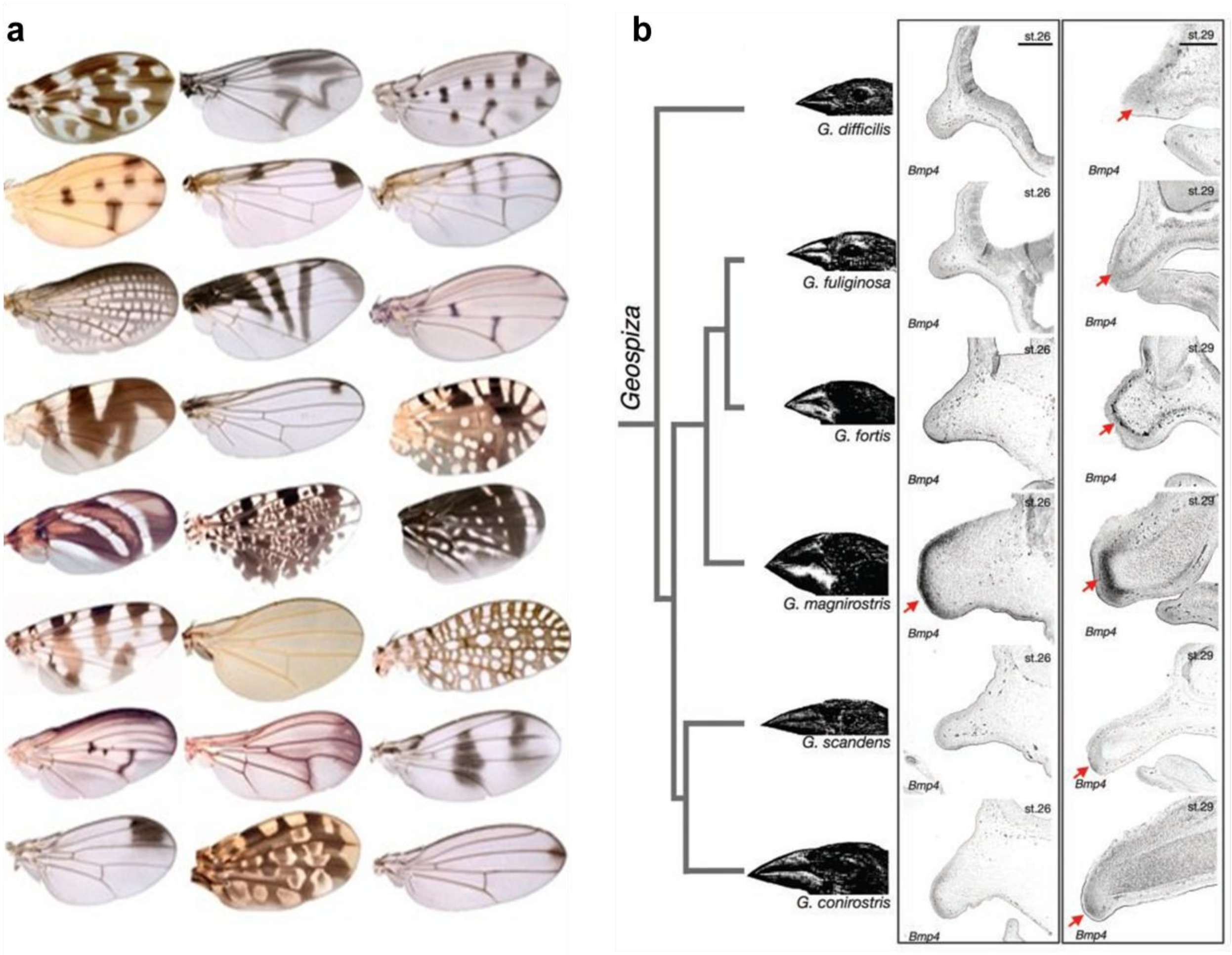
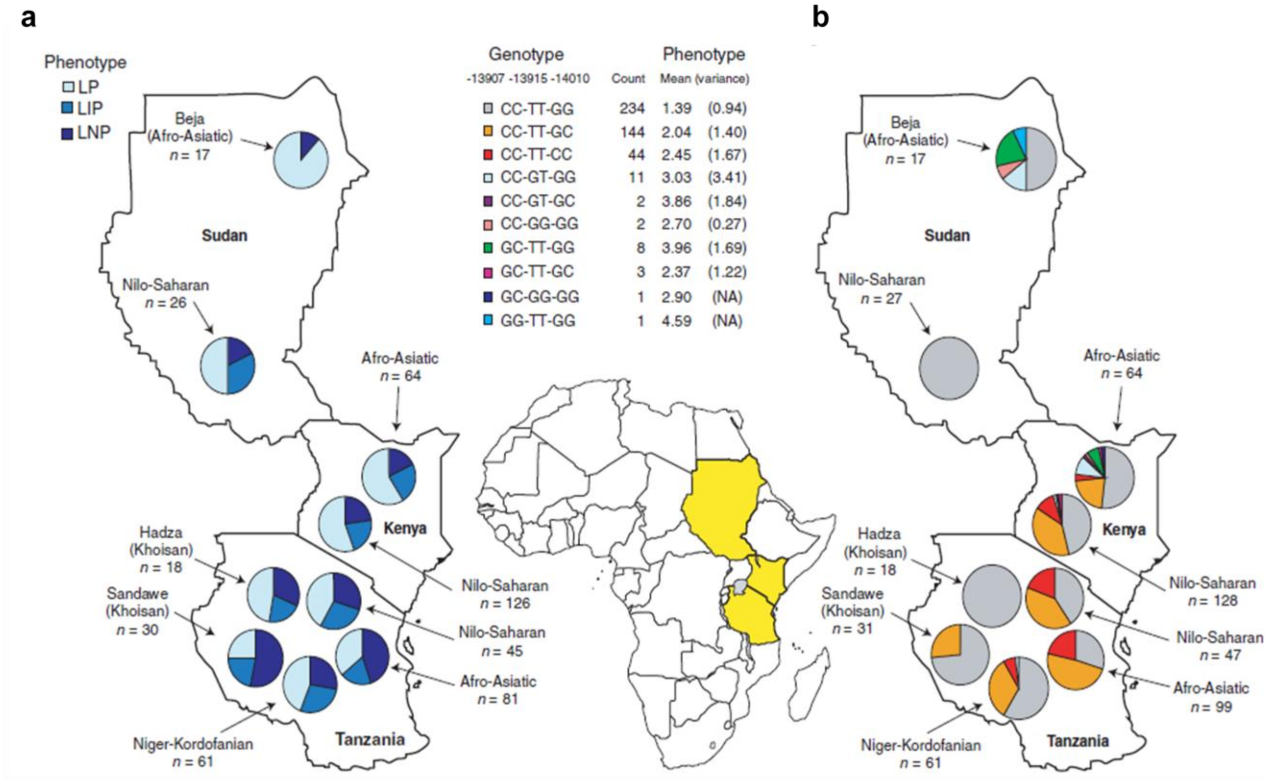


Figure 1. Variation in gene expression levels and patterns underlies phenotypic differences. **a)** Like butterflies, different species of *Drosophila* decorate their wings with a great diversity of spots and patterns. Expression of a single gene produces pigmentation patterns and acts as a molecular switch that controls where pigmentation is deployed. This finding explains how expression can be controlled to produce the seemingly endless array of patterns, decoration and body architecture found in animals. Photo by N Gompel and B Prud'homme (from http://www.news.wisc.edu/newsphotos/fruitfly.html). **b)** Expression differences of Bmp4 give rise to beak morphology differences in *Geospiza*. *G. difficilis* is the most basal species of this genus, and the rest of the species form two groups: ground and cactus finches, with distinct beak morphologies. At stage 26 (middle panel) Bmp4 is strongly expressed in the mesenchyme of the upper beak of *G. magnirostris* and at lower levels in *G. fortis* and *G. conirostris*. No Bmp4 was detected in the mesenchyme of *G. difficilis*, *G. fuliginosa*, and *G. scandens*. At stage 29 (right panel) Bmp4 continues to be expressed at high levels in the distal beak mesenchyme of *G. magnirostris*. Broad domains of Bmp4 expression are detectable in *G. fuliginosa* and *G. fortis*. A small domain of Bmp4 expression is also found in the distal mesenchyme of *G. conirostris*, and weaker expression is seen in *G. scandens* and *G. fortis* (red arrows). Adapted from (Abzhanov, Protas et al. 2004).

In humans the ability to digest lactose, the main carbohydrate in milk, declines rapidly after weaning. This is due to decreasing levels of lactase-phlorizin hydrolase, which metabolises lactose and is encoded by the gene *LCT*. Adult expression of LCT results in the ability to digest milk and other dairy products in adulthood (lactase persistence or lactose tolerance) and differences in LCT levels lead to the differences in lactase persistence observed in a number of populations across the world (Tishkoff, Reed et al. 2007) (Figure 2). Overall these examples highlight the role of gene expression in shaping natural range phenotypes.



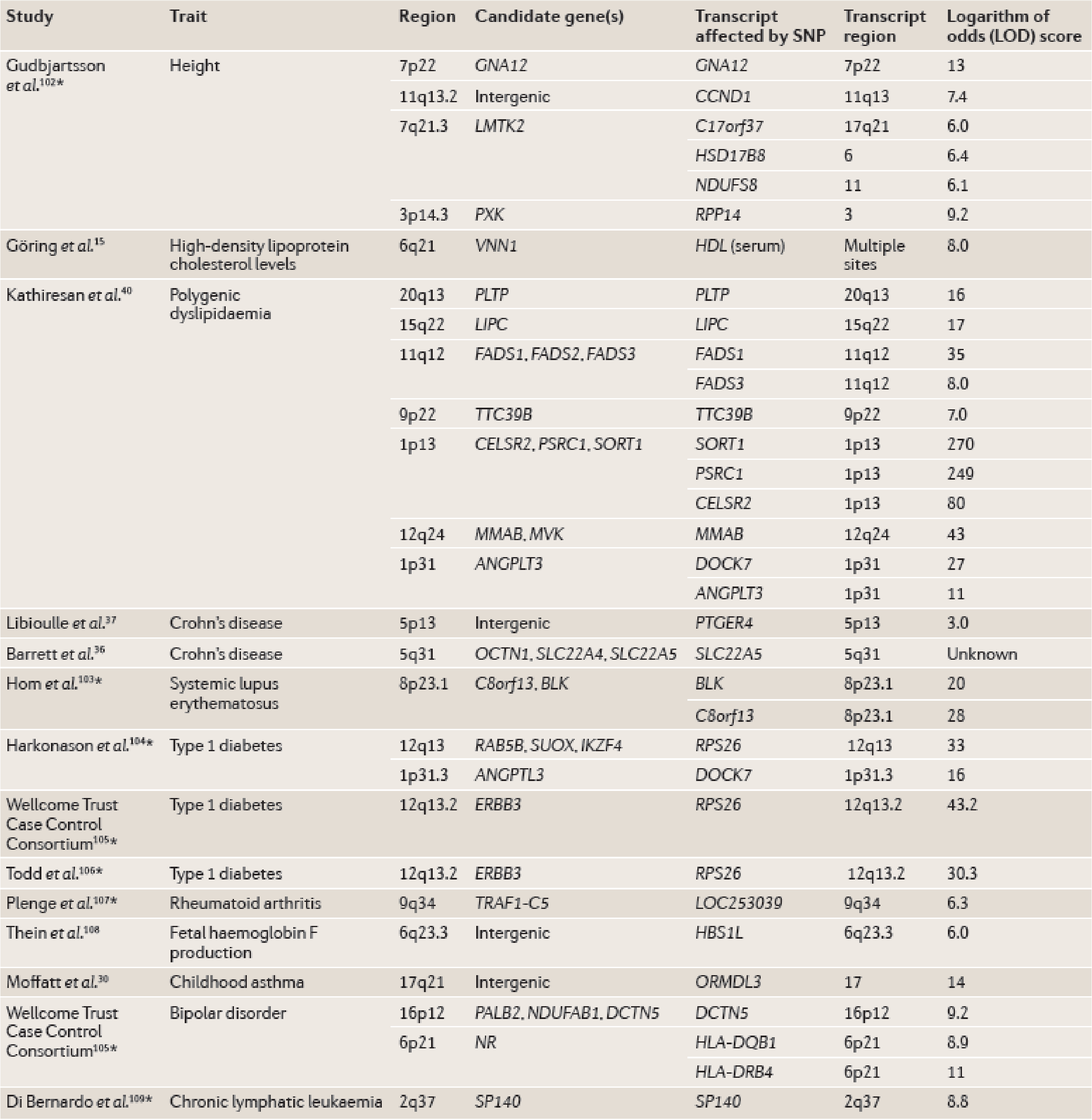
Genotype -13907 -13915 -14010	Count	Phenotype Mean (variance)	
■ CC-TT-GG	234	1.39	(0.94)
■ CC-TT-GC	144	2.04	(1.40)
■ CC-TT-CC	44	2.45	(1.67)
□ CC-GT-GG	11	3.03	(3.41)
■ CC-GT-GC	2	3.86	(1.84)
■ CC-GG-GG	2	2.70	(0.27)
■ GC-TT-GG	8	3.96	(1.69)
■ GC-TT-GC	3	2.37	(1.22)
■ GC-GG-GG	1	2.90	(NA)
■ GG-TT-GG	1	4.59	(NA)

Figure 2. Lactase persistence differences across human populations are due to differences in adult expression of lactase. a) The degree of lactase persistence is represented by a pie chart for each geographic region (LP: lactase persistence, LIP: lactase intermediate persistence, LNP: lactase non-persistence). **b)** Proportion of compound genotypes of variants that influence levels of LCT (G/C-13907, T/G-13915 and C/G-14010). The pie charts are in the approximate geographic location of the sampled individuals. Adapted from (Tishkoff, Reed et al. 2007).

1.2.4 Gene expression can shape disease phenotypes

Variation in gene expression can have a detrimental impact on cells and tissues if expression profiles are perturbed beyond the range of tolerance (Hartman, Garvik et al. 2001; Stranger and Dermitzakis 2006; Nica and Dermitzakis 2008; Cookson, Liang et al. 2009). Many-fold over-expression of C-MYC can lead to Burkitt's lymphoma (Boxer and Dang 2001), a reduction of APC expression is associated with a pronounced predisposition to hereditary colorectal cancers (Yan, Dobbie et al. 2002) and partial or complete loss of α-globin expression can lead to α-thalassaemia (Weatherall 1998). Subtle changes in gene expression can also contribute to disease phenotypes, as is the case for Type 1 diabetes, whose manifestation depends on the genetic background of individuals (Eaves, Wicker et al. 2002). Type 1 diabetes was one of the first instances where genetic variation driving gene expression was shown to be associated with disease risk (Bennett, Lucassen et al. 1995; Kennedy, German et al. 1995). The insulin-linked polymorphic region (ILPR), mapping 5′ of the *INS* gene is composed of a series of tandemly-repeated sequences that contain high affinity binding sites for the TF Pur-1. Allelic variation in these sequences was shown to influence *INS* transcription levels and risk for diabetes. Table 1 (Cookson, Liang et al. 2009) summarises cases from the literature and public databases in which trait and disease phenotypes arise in part due to variation in expression levels.

The link between human tissue-specific gene expression and pathological manifestations has been demonstrated in multiple studies. Lage et al. (2008) mapped 2,000 disease genes to the tissues they affect and identified 1,500 disease-associated complexes. The expression patterns of complex components were analysed and disease genes were found to be over-expressed in the normal tissues where defects eventually cause pathology. For example a complex involved in XY sex reversal was found to be testis-specific and was down-regulated in the ovaries. Tissue specificity of expression was identified for complexes with a role in Parkingson disease, cardiomyopathies and muscular dystrophies.



Study	Trait	Region	Candidate gene(s)	Transcript affected by SNP	Transcript region	Logarithm of odds (LOD) score
Gudbjartsson *et al.*[102]*	Height	7p22	*GNA12*	*GNA12*	7p22	13
		11q13.2	Intergenic	*CCND1*	11q13	7.4
		7q21.3	*LMTK2*	*C17orf37*	17q21	6.0
				HSD17B8	6	6.4
				NDUFS8	11	6.1
		3p14.3	*PXK*	*RPP14*	3	9.2
Göring *et al.*[15]	High-density lipoprotein cholesterol levels	6q21	*VNN1*	*HDL* (serum)	Multiple sites	8.0
Kathiresan *et al.*[40]	Polygenic dyslipidaemia	20q13	*PLTP*	*PLTP*	20q13	16
		15q22	*LIPC*	*LIPC*	15q22	17
		11q12	*FADS1, FADS2, FADS3*	*FADS1*	11q12	35
				FADS3	11q12	8.0
		9p22	*TTC39B*	*TTC39B*	9p22	7.0
		1p13	*CELSR2, PSRC1, SORT1*	*SORT1*	1p13	270
				PSRC1	1p13	249
				CELSR2	1p13	80
		12q24	*MMAB, MVK*	*MMAB*	12q24	43
		1p31	*ANGPLT3*	*DOCK7*	1p31	27
				ANGPLT3	1p31	11
Libioulle *et al.*[37]	Crohn's disease	5p13	Intergenic	*PTGER4*	5p13	3.0
Barrett *et al.*[36]	Crohn's disease	5q31	*OCTN1, SLC22A4, SLC22A5*	*SLC22A5*	5q31	Unknown
Hom *et al.*[103]*	Systemic lupus erythematosus	8p23.1	*C8orf13, BLK*	*BLK*	8p23.1	20
				C8orf13	8p23.1	28
Harkonason *et al.*[104]*	Type 1 diabetes	12q13	*RAB5B, SUOX, IKZF4*	*RPS26*	12q13	33
		1p31.3	*ANGPTL3*	*DOCK7*	1p31.3	16
Wellcome Trust Case Control Consortium[105]*	Type 1 diabetes	12q13.2	*ERBB3*	*RPS26*	12q13.2	43.2
Todd *et al.*[106]*	Type 1 diabetes	12q13.2	*ERBB3*	*RPS26*	12q13.2	30.3
Plenge *et al.*[107]*	Rheumatoid arthritis	9q34	*TRAF1-C5*	*LOC253039*	9q34	6.3
Thein *et al.*[108]	Fetal haemoglobin F production	6q23.3	Intergenic	*HBS1L*	6q23.3	6.0
Moffatt *et al.*[30]	Childhood asthma	17q21	Intergenic	*ORMDL3*	17	14
Wellcome Trust Case Control Consortium[105]*	Bipolar disorder	16p12	*PALB2, NDUFAB1, DCTN5*	*DCTN5*	16p12	9.2
		6p21	*NR*	*HLA-DQB1*	6p21	8.9
				HLA-DRB4	6p21	11
Di Bernardo *et al.*[109]*	Chronic lymphatic leukaemia	2q37	*SP140*	*SP140*	2q37	8.8

Table 1. Trait and disease phenotypes with an identified gene expression component. Disease-linked associations with significant expression quantitative trait loci (QTLs) from the literature and public databases From (Cookson, Liang et al. 2009).

Tissue specificity in disease pathology has also been addressed from an evolutionary standpoint in a study exploring the relationship between disease genes, tissue specificity, and evolutionary rates (Winter, Goodstadt et al. 2004). Cell type specificity is known to correlate positively with gene evolution rates and ubiquitously expressed, slowly evolving housekeeping genes were found to be under-represented in human disease. Genes with a role in disease on the other hand, had secreted protein products and were highly expressed in tissues such as liver, kidney and lung. This observation is likely due to the effects of purifying selection and may assist in prioritization of candidate genes.

A recent study highlighted the role of a single TF in regulating markedly different cell type-specific programmes (Servitja, Pignatelli et al. 2009). HNF1A controls tissue-specific genetic programmes in pancreatic islets and the liver, and its deficiency causes a severe β-cell phenotype (HNF1A-deficient diabetes), but only subtle abnormalities in other tissues. The final phenotypic outcome of Hnf1a deficiency in mice was highly cell type-specific and resulted from an integrated failure of multiple direct and indirect functions of this gene in pancreatic islets and liver. Due to the breadth of Hnf1a-dependent transcriptional programmes, the authors suggest that correction of defects causing β-cell dysfunction should not focus on restoring individual target gene activity, but should aim at manipulating proteins or pathways acting on the β-cell HNF1A-dependent programme. Taken together these examples outline the multiple effects of gene expression patterns and the role of gene regulation in determining disease risk.

1.3 The mechanism of gene expression

In eukaryotic cells, protein-encoding genes are transcribed in the nucleus by RNA polymerase II. The RNA transcript produced is the messenger RNA (mRNA) which acts as an intermediary between the gene and the protein product. A further step, known as translation, is necessary to convert the information carried by the mRNA into a protein (Clark 2005). In the following section I outline the process of gene expression for protein-encoding genes (summarised in Figure 3).

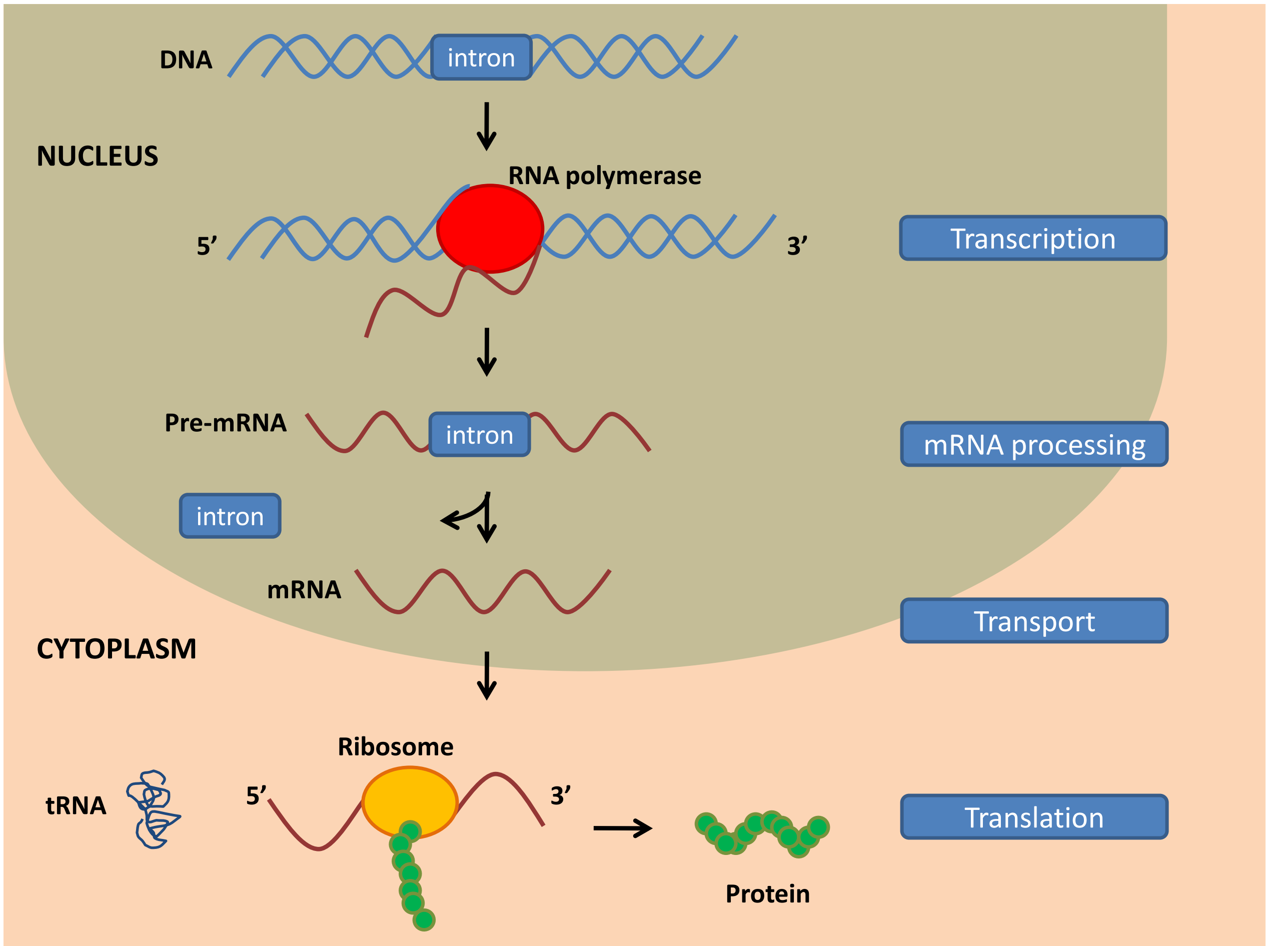


Figure 3. The process of gene expression for protein-encoding genes. RNA is transcribed from a DNA template by RNA polymerase II in the nucleus to produce pre-mRNA (transcription). The pre-mRNA undergoes a series of processing steps including splicing, 5′ capping and 3′ polyadenylation (mRNA processing). Processed mRNA molecules are transferred from the nucleus into the cytoplasm (transport) where they engage with ribosomes and other components of the translational machinery that direct polypeptide synthesis (translation). Adapted from (Clark 2005).

1.3.1 Transcription

RNA polymerase II uses nuclear DNA as a template and ribonucleoside triphosphates to produce pre-mRNA molecules in a 5′ to 3′direction. Chain elongation is achieved by the addition of ribonucleoside monophospate residues to the free hydroxyl group at the 3′ end of the growing pre-mRNA chain (Strachan and Read 2004). This process gives rise to the primary transcript (or pre-mRNA), an RNA molecule complementary to the full sequence of the gene (exons and introns).

1.3.2 mRNA processing

The pre-mRNA undergoes a series of processing steps in the nucleus including splicing, 5′ capping and 3′ polyadenylation. Splicing is mediated by the spliceosome, a large RNA-protein complex, which recognises sequences at exon/intron boundaries (splice junctions). Intronic RNA segments are removed by endonucleolytic cleavage and exonic segments are joined end-to-end (spliced). The end product is a shorter RNA product (mRNA) that contains the information encoding a protein (exons). Alternative splicing can bring together different combinations of exons to produce versions of the polypeptide product (isoforms).

Further mRNA processing involves addition of a methylated nucleoside, 7-methylguanosine (5′ cap) to the first 5′ nucleotide of the RNA transcript, as well as addition of a 3′ polyA tail. 5′ caps and 3′ polyA tails facilitate transfer of mRNA molecules to the cytoplasm, ensure RNA stability, and assist recognition by the translational machinery (Strachan and Read 2004). In some instances in somatic cells, mRNA molecules undergo RNA editing, which results in a coding sequence difference between mRNA and DNA sequence. mRNA editing of *APOB* gene transcripts in the liver for example introduces a stop codon in the mRNA transcript and gives a much shorter product from the one generated by the unedited mRNA molecule in the intestine (Navaratnam, Bhattacharya et al. 1995; Lewin 2008).

1.3.3 mRNA transport and translation

Following post-transcriptional processing, mRNA molecules migrate from the nucleus to the cytoplasm where they engage with ribosomes and other components of the translational machinery that direct polypeptide synthesis (Strachan and Read 2004). The central part of mRNA molecules encodes the amino acid sequence whereas 5′ and 3′ mRNA ends are untranslated regions (UTRs) (transcribed from the first and terminal exons respectively) with a role in binding and stabilizing mRNA on ribosomes.

Assembly of polypeptides from their constituent amino acids is governed by the triplet genetic code with successive groups of three nucleotides (codons) in the linear mRNA sequence encoding an individual amino acid. Decoding of mRNA is mediated through tRNA molecules that bear specific trinucleotide sequences (anticodons) and covalently bound amino acids. Recognition of the complementary codon on the mRNA ensures that the appropriate amino acid is inserted in the growing polypeptide chain. Translation products are frequently modified, usually through covalent attachment of hydroxyl, phosphoryl, carbohydrate and lipid groups to amino acid side chains. Upon modification, polypeptides may undergo cleavage to generate smaller, mature proteins (e.g. β-globin, plasma proteins, neuropeptides) (Strachan and Read 2004).

1.4 Regulation of gene expression

As discussed, gene expression is influenced by genetic, epigenetic, and environmental factors that give rise to expression differences between species, populations and cell types. Furthermore, interactions between genetic factors (Brem, Storey et al. 2005; Boone, Bussey et al. 2007; Dimas, Stranger et al. 2008) as well as those between genetic factors and the environment (Gibson 2008) have a key role in shaping expression levels. Although expression regulation involves multiple levels, in eukaryotes the most common point of control is initiation of transcription. For the purposes of this thesis the products of transcription (mRNA levels) were regarded as a proxy to gene expression (see section 2.3).

1.4.1 Transcriptional regulation of gene expression

The simplest model of transcription of protein-coding genes in eukaryotes involves recruitment of RNA polymerase II, which recognises and binds to combinations of short DNA sequences in the proximity of a gene. These sequence elements, are referred to as cis-acting and serve as recognition signals for TFs that engage in gene expression regulation by guiding and activating the polymerase (Strachan and Read 2004). Transcription initiation is influenced by DNA sequences located further away (or on another chromosome) from the gene whose activity is being regulated. These sequences are known as trans-acting and encode proteins that influence transcription levels (e.g. TFs). In this thesis cis regulatory elements are defined as those mapping within a 2 Mb window centred on the probe midpoint or the transcription start site (TSS) of a gene (see section 2.4) and trans-acting regulatory elements are those mapping outside this 2 Mb window or on another chromosome.

Multiple cis and trans elements act in conjunction with each other to control transcription initiation and mRNA levels for a given gene (Stranger, Forrest et al. 2005; Stranger, Nica et al. 2007; Dimas, Deutsch et al. 2009). The identity of regulatory sequences, the TFs present, and their binding affinities all play an important role in transcription initiation. Mutations altering the nucleotide sequence of any of these elements, or the nucleotide sequence of the transcript (affecting its stability), may have substantial effects on mRNA transcript levels (Stranger and Dermitzakis 2005). The genomic distribution and complexity of cis and trans sequence elements, as well as the architecture of the regulatory landscape is an area of active research and a substantial effort is underway to annotate regulatory elements in the genome. A pilot project in which 1% of the genome was studied (Birney, Stamatoyannopoulos et al. 2007) revealed that the distribution of regulatory sequences is variable, with elements being scattered across the genome. In the following paragraphs I describe well-studied regulatory sequence elements, proteins and RNA molecules.

1.4.1.1 Promoters

Promoters are short sequence cis-acting elements that cluster in the immediate upstream region of a gene's coding sequence, often within 200 base pairs (bp) of the TSS, and control transcription initiation. RNA polymerase and a number of general TFs bind to the promoter region of a gene, which is typically made up of different components, to form the basal transcription complex. Upon binding, the polymerase is activated and RNA synthesis is initiated. Well-studied promoter elements are discussed in Box 1.

The **initiator box (Inr)** is a sequence bound by general TF TFIID at the site of transcription initiation. The first transcribed base of the mRNA is usually an A with a pyrimidine on each side.

The **core promoter sequence** directs the basal transcription complex (RNA polymerase and general TFs) to initiate transcription of a gene. In the absence of additional regulatory elements, it permits constitutive gene expression, but at very low (basal) levels. Core promoters are typically located very close to the TSS, at nucleotide positions -45 to +40, and include the **TATA box**, the **BRE sequence**, and the **DPE**. The TATA box (TATAAA sequence or a variation of it), found approximately at -25 of the TSS, is usually surrounded by GC rich sequences and is recognised by the TATA-binding protein subunit of TFIID. Immediately upstream of the TATA element is the BRE sequence which is recognised by TFIIB. The DPE (downstream promoter element) is located at approximately +30 and is recognised by TFIID.

Non-core promoter elements map immediately upstream of the core promoter, typically spanning the region of -50 to -200 bp and include **GC boxes (Sp1 boxes)** and **CCAAT boxes**. GC boxes are found in multiple copies within 100 bp of the TSS and are bound by the general TF Sp1. CCAAT boxes are strong determinants of promoter efficiency, locate approximately 80 bp upstream of the TSS, and are recognised by TFs CTF and CBF. Both GC and CCAAT boxes function to modulate basal transcription levels of the core promoters, operating as essentially as enhancer sequences.

Box 1. Promoter elements. Adapted from (Strachan and Read 2004).

1.4.1.2 Enhancers

Enhancers are positive control sequence elements, located at variable and often considerable distances from a gene, that increase the basal level of transcription initiated through promoter elements. They are short DNA sequences and may contain several elements recognised by TFs in a ubiquitous or tissue-specific manner (Heintzman, Hon et al. 2009; Visel, Blow et al. 2009). Upon TF binding, the DNA between the enhancer element and the promoter loops out and allows the proteins bound to the enhancer to interact with the basal transcription complex (Strachan and Read 2004). A well-studied enhancer is the locus control region (LCR) located 50-60 kilobases (kb) upstream of the *β-globin* gene whose expression it activates.

1.4.1.3 Silencers

Silencers have similar properties to enhancers, but act to reduce expression levels by inhibiting the transcriptional activity of genes. They have been reported in various positions relative to human genes: close to the promoter, upstream of the TSS, and within introns. Classical silencer elements are position-independent sequences that direct an active transcriptional repression mechanism (Strachan and Read 2004). Negative regulatory elements are position-dependent sequences that exert passive repression of transcription and often act by interfering with activators rather than by obstructing the movement of RNA polymerase.

1.4.1.4 Insulators

Insulators (boundary elements) are regions of DNA spanning a few hundred to a few thousand bases (typically 0.5-3 kb) which block the spreading of agents that affect transcription in a positive or negative manner, and divide chromosomes into regulatory neighbourhoods (Clark 2005). They contain clusters of GC rich sequences that bind multiple copies of zinc-finger proteins known as insulator binding proteins (IBPs). In many cases their action can be countered by methylation of GC sequences.

1.4.1.5 Response elements

Response elements are usually located a short distance upstream of promoter elements (1 kb upstream of the TSS) and are responsible for modification of transcription in response to environmental stimuli. Response elements can respond to specific hormones (e.g. retinoic acid or steroid hormones such as glucocorticoids) or to intracellular second messengers such as cyclic AMP (Strachan and Read 2004).

1.4.1.6 Transcription factors

RNA polymerase II transcribes genes following binding of TFs to specific regulatory DNA within the gene and its vicinity. TFs are typically regarded as trans-acting elements and may bind to the promoter region around genes or to distant enhancer sequences. Activators are TFs that stimulate transcription and repressors are those with antagonistic effects. TFs can be general (e.g. components of the basal transcription complex such as TFIIB or TFIID) or tissue-specific (e.g. HNF1A which controls tissue-specific expression in pancreatic islets and the liver). General TFs are required for transcription from all promoters occupied by RNA polymerase II and their binding results in basal levels of transcription. Specialised TFs modulate basal transcription levels and influence the activity of specific gene sets, usually in a tissue-specific manner.

1.4.1.7 MicroRNA

MicroRNAs (miRNAs) are single stranded, 21–24 nucleotide, regulatory RNA molecules abundant in animals, plants and viruses (Flynt and Lai 2008). They are encoded by genes from whose DNA they are transcribed, but are not translated into protein. Instead each transcript is processed into a short stem-loop structure called a pre-miRNA and finally into a functional miRNA. miRNA molecules are fully or partially complementary to mRNA molecules and their main function is to down-regulate gene expression through partial base pairing with their target mRNAs. Base pairing either inhibits translation of target mRNA molecules or speeds up deadenylation causing mRNA degradation (Williams 2008).

1.4.2 Other mechanisms of gene expression regulation

Although transcription is the primary means of expression regulation, gene activity can be modulated post-transcriptionally, through mechanisms involving mRNA processing, transport and stability at the mRNA level, as well as translation, processing, targeting and stability at the protein level. Furthermore, expression regulation can be achieved epigenetically through DNA methylation, histone modification and the action of non-coding RNA molecules. A detailed overview of these mechanisms can be found in Genes IX (Lewin 2008).

1.5 GENETIC VARIATION IN GENE EXPRESSION

As described, gene expression is a complex, quantitative trait controlled at many levels and sculpted by numerous factors. In this thesis I address the genetic component of expression variation, or the fraction of transcript level differences that arises as a consequence of genetic variation in DNA sequences. Broadly speaking, genetic variation influencing gene expression can manifest itself in four major ways: gene expression differences among populations, among individuals in a population, among tissues, and in response to environmental factors. In this section I outline a number of landmark studies that have contributed to our understanding of the genetic component of gene expression.

The first series of large-scale studies aiming to uncover regulatory DNA variation focused on model organisms. A genetic component for naturally occurring variation in gene expression was documented in yeast (Brem, Yvert et al. 2002; Steinmetz, Sinha et al. 2002), maize (Schadt, Monks et al. 2003), fruit flies (Jin, Riley et al. 2001; Wittkopp, Haerum et al. 2004), and mice (Sandberg, Yasuda et al. 2000; Cowles, Hirschhorn et al. 2002; Lo, Wang et al. 2003; Schadt, Monks et al. 2003). In humans, familial aggregation of expression profiles was demonstrated by Cheung et al. (2003) who showed that variability in transcript abundance was lower in more closely related individuals. Gene expression heritability estimates for the same individuals showed that approximately 25% of the genes studied had significant heritable variation (Schadt, Monks et al. 2003). This implied a heritable component of gene expression variation among humans and laid the groundwork for subsequent studies in primates (Enard, Khaitovich et al. 2002) and humans (Cheung, Conlin et al. 2003; Monks, Leonardson et al. 2004; Morley, Molony et al. 2004; Pastinen, Sladek et al. 2004; Stranger, Forrest et al. 2005; Dixon, Liang et al. 2007; Goring, Curran et al. 2007; Stranger, Forrest et al. 2007; Stranger, Nica et al. 2007; Dimas, Deutsch et al. 2009).

To date, most studies interrogating the genetic basis of regulatory variation have explored the effects of single variants on gene expression. Experiments in yeast however have revealed that the inheritance of over half of all transcripts is influenced by interacting locus pairs (Brem, Storey et al. 2005). Interactions between genetic factors have also been shown to occur in humans, for example in studies where the functional impact of coding variants is modified by regulatory variants nearby (Dimas, Stranger et al. 2008; Wang, Cruchaga et al. 2009). However, systematic measures of the extent of genetic interactions are lacking (Flint and Mackay 2009). Furthermore, although numerous large-scale studies have identified loci with a role in expression regulation, in most cases the candidate regions defined are broad and identification of true functional variants is pending. Finally, the bulk majority of studies to date have explored gene expression in a single cell type, usually in Epstein-Barr virus (EBV)-transformed B-cells (lymphoblastoid cell lines or LCLs), as these can be easily obtained from B-cells in blood samples and maintained in the laboratory. The extent of cell type specificity of gene expression (described in section 1.2.2) underscores the need to explore expression systematically in other cell types and catalogue cell type-specific regulatory variation. With the increasing realisation of the role of regulatory variation in shaping phenotypes in health and disease, detection and precise identification of single and interacting variants in multiple populations and cell types is a priority.

1.6 Detecting regulatory variation

Technological advances in the last decade, especially the development of microarray platforms, have made it possible to move from low and medium-throughput quantification of gene expression (e.g. reporter, or allele-specific expression assays (ASE)) to genome-wide quantification of mRNA levels. Transcript abundance for each of thousands of genes can be determined in a single experiment with mRNA intensity values reflecting mRNA levels. mRNA intensity exhibits continuous variation among individuals and mapping gene expression variation is a typical quantitative trait exercise (Stranger and Dermitzakis 2005; Dermitzakis and Stranger 2006). The rationale used to map quantitative trait loci (QTLs) for continuous phenotypes such as weight and height is also employed to detect expression QTLs (eQTLs) (Mackay, Stone et al. 2009). In human populations two approaches have been employed for eQTL mapping: linkage and association mapping (Dermitzakis and Stranger 2006; Gilad, Rifkin et al. 2008; Mackay, Stone et al. 2009).

1.6.1 Linkage mapping

Linkage mapping tracks the transmission of chromosomes through families using pedigrees and requires data on phenotypes and markers for each family member. The aim is to identify markers whose transmission patterns correlate with the phenotype, the implication being that these markers are linked to causal variants driving the phenotype (Gilad, Rifkin et al. 2008). The advantage of linkage mapping is that it requires a relatively low density of markers (<1,000 for microsatellites and slightly higher numbers for single nucleotide polymorphisms (SNPs)). However, it provides coarse-grained (low resolution) localisation, as it depends on the occurrence of recombination events within families for finer mapping (Gilad, Rifkin et al. 2008). Some of the first genome-wide studies on gene expression in humans employed a linkage approach, using cell lines from individuals of Centre d'Étude du Polymorphisme Humain (CEPH) pedigrees (Monks, Leonardson et al. 2004; Morley, Molony et al. 2004). This approach is powerful when functional variants are rare and there is allelic heterogeneity (different mutations at the same locus that give rise to the same phenotype), as is the case for β-thalassaemia which can be caused by several different mutations in the *β-globin* gene (Dermitzakis and Stranger 2006). If the variants affecting gene activity are of smalle effect size (minor allele frequency (MAF) > 5%), linkage is relatively underpowered and association mapping performs better.

1.6.2 Association mapping

Association mapping identifies markers whose genotypes show a statistical association to the phenotype of interest (in this case mRNA abundance). A statistically significant association for a given marker implies that it is linked to a functional regulatory variant. In its simplest form, association mapping uses samples of unrelated individuals and dense genotyping data (e.g. 500,000 SNPs for a genome-wide study in humans). It is the most powerful method to date for the detection of common variants, provided that the causal sequences are in strong linkage disequilibrium (LD) with the genotyped SNPs (Dermitzakis and Stranger 2006). Additionally, with sufficiently dense genotyping, association mapping is more likely to detect variants with small or medium effect sizes (Gilad, Rifkin et al. 2008). Although this approach rarely detects true functional variants, the resolution provided is much higher compared to linkage, with functional variants mapping within hundreds of kb of associated markers depending on the extent of LD. One potential caveat of association mapping is the occurrence of false positives arising as a consequence of population structure, but this can be resolved using methods that correct for structure (Price, Patterson et al. 2006).

1.7 GENETIC VARIANTS TESTED IN ASSOCIATION STUDIES

An estimated 99.9% of the 6 billion nucleotides making up the human genome is identical across individuals (Sachidanandam, Weissman et al. 2001). The remaining 0.01% that varies between any two randomly chosen individuals consists of variation occurring on different scales and ranges from single base changes to alterations in copy number of larger segments. Genetic variants in the human genome include SNPs, insertion/deletion polymorphisms (indels), retroposon insertions, variation in the number of copies of a tandem repeat (mini and microsatellites), copy number variants (CNVs), inversions and variants that are a combination of some or all the above.

In this thesis genetic variation in the form of SNPs was associated with transcript levels to detect eQTLs. SNPs are the simplest and most common type of genetic variant, constituting roughly 75% of the total variation observed in humans (Levy, Sutton et al. 2007). They are the smallest unit of polymorphism and arise from the exchange of a single base in the DNA sequence (Hartl and Clark 2007). Traditionally, a DNA position is said to be polymorphic when alleles are found at a frequency between 1% and 99% in the population. The human genome is estimated to contain over ten million SNPs, seven million of which are designated as common (MAF ≥ 5% across the entire population) (Kruglyak and Nickerson 2001; Crawford, Akey et al. 2005). The International HapMap Consortium, launched in 2002 aimed to identify and catalogue these variants to quantify the extent of genetic similarities and differences between humans (International HapMap Consortium 2003). Currently, over four million SNPs in 1,301 individuals from eleven geographically distinct populations have been assayed (see section 2.1.1). Depending on their position in the genome, SNPs can be non-coding or coding. For the 1.5% of the genome that encodes proteins, the redundancy of the genetic code means that in some cases specific amino acids can be encoded by multiple codons. Synonymous SNPs are those base substitutions that do not alter the amino acid sequence, while coding or non-synonymous SNPs (nsSNPs) are those that lead to a change of a single amino acid.

1.8 Thesis aims

Studies addressing the genetic component of gene expression have uncovered an abundance of common genetic variation influencing gene expression and have defined a field of intense study over the past few years. It is now well-established that regulatory polymorphisms are widespread in the human genome, with cis and trans-acting loci regulating transcript levels of genes. Most studies to date however have explored the effects of single genetic variants and have interrogated expression in a single cell type. Furthermore, although these studies have made a very important first step in detecting regions harbouring regulatory variants, few have identified precise functional variants. In this thesis I aim to further our understanding of regulatory variation by: a) exploring the effect of interactions between genetic variants on transcript levels (Chapter 3), b) dissecting the fine-scale architecture of the cis regulatory landscape (Chapter 4) and by c) exploring the extent of cell type specificity of regulatory variation (Chapter 5). Uncovering regulatory variation and understanding its function will help elucidate developmental programmes and patterns of cell type specificity and will also shed light on processes determining natural range and disease phenotypes.