1 Introduction

The ultimate phenotypic effect of a gene product depends on two different components; the identity and structure of the product itself, and the spatial and temporal regulation of its expression. The former is defined largely by the coding sequence of the gene, although post-translational modifications on the protein also play a part. The precise relationship between coding sequence and primary protein product has been thoroughly elucidated since the discovery of the structure of deoxyribose nucleic acid (DNA) in 1953, and is now a firm fixture at the base of molecular biology. The latter component, however, remains far less well understood despite increasing attention and resources being focused on it. Detailed studies of particular gene loci in both model organisms and humans have helped elucidate some of the mechanisms that control gene expression (Wright et al. 1984; Whitehead and Sackstein 1985; Bulger et al. 2002; Ting and Trowsdale 2002), as well as some of the sequence elements that are involved in these processes. However, it has proved difficult to generalise these to the whole genome. The variety of possible regulatory mechanisms and elements has meant that, despite longstanding interest in the regulatory aspect of phenotype (King and Wilson 1975), nothing remotely close to the genetic code for protein-coding genes exists in the regulatory sense.

In the post-genomic era, it has become clear that the number of genes in a genome is not necessarily correlated with the perceived complexity of an organism. The fact that fewer than 25,000 transcriptional units are present in humans suggests that a large component of the myriad of known phenotypes and diseases must be accounted for by regulatory rather than coding variation. A compelling sign of this is that the proportion of highly conserved bases outside of protein-coding genes increases with overall biological complexity, suggesting that a significant component of this complexity is underlain by non-coding, and presumably regulatory sequences (Siepel et al. 2005). In recent years, renewed efforts have been made to study the non-coding genome in search of the identity and mechanism of action of sequence elements that regulate gene expression. This has been easier in model organisms than humans, with yeast being a particularly productive system for inferring gene regulatory networks and elements (Ren et al. 2000; Lee et al. 2002). In humans, the most notable of these is the ENCODE project (ENCyclopaedia Of DNA Elements), which aims to

functionally annotate regulatory elements in 1% of the human genome (Consortium 2004b).

This thesis has sought to explore the impact of putative *cis*-regulatory sequence on gene expression by discovering variation in promoter sequences, and testing them to identify mutations that have an effect on promoter activity. Promoters are currently the only regulatory element that can be readily predicted on the basis of a positional relationship with known genes, and is therefore the most reliable place to start when exploring the mechanistic basis of gene expression regulation

1.1 Transcriptional regulation

The information contained within genes is converted to a useful product by first transcribing the DNA into mRNA, which is then in turn translated into a protein sequence. This in turn undergoes post-translational processing before becoming an active finished protein. While the mechanics of this process that underpins all of life are, not surprisingly, conserved to the point of ubiquity, the regulatory events that control them have undergone fundamental change over evolutionary time. In prokaryotes, transcriptional regulation is relatively simple, with a general scheme consisting of co-regulated genes being transcribed together in polycistronic operons, and with the transcription initiation being regulated almost completely by the binding of transcription factors (TFs) in 5' flanking sequence of the first gene in the operon. In eukaryotes, genes are transcribed as individual units, and concordance of regulation across multiple genes is achieved by having common regulatory signals affecting each. In addition, regulatory DNA elements are often spread over larger distances relative to the genes they regulate, and there is more heterogeneity in the type of regulatory mechanisms in use. In humans and other mammals, transcriptional regulatory mechanisms can be divided into two classes; TFs and epigenetic mechanisms.

The large number of TFs in the human genome gives rise to the potential for an extremely large combination of possible regulatory signals. They are usually the terminal components of signalling cascades relaying signals from a variety of sources, thus ensuring the correct spatio-temporal expression of the genes they control. They

are regulated both at the level of transcription (and hence by other TFs) and post-translational modification. Of course, TF genes are subject to the same transcriptional regulatory mechanisms as other protein-coding genes. Cascades of linked TFs can be set up, where one factor regulates the expression of a further TF gene, whose product in turn regulates one or more downstream TF genes. A good example is the regulation of gene expression in liver cells, where an array of TFs including c/EBP, HNF-1α, HNF-4α and HNF-3β are involved in a regulatory cascade resulting from growth hormone stimulation (Rastegar, Lemaigre, and Rousseau 2000). It is also common for TFs to regulate their own expression. Examples include Pit-1 (Rhodes et al. 1993) and c/EBP (Legraverend et al. 1993; Timchenko et al. 1995). Post-translational modification of TFs that are already present allows dynamic and hence rapid regulation of their activity. There are several different levels at which they can be regulated. These include the phosphorylation (e.g. the MAP kinase pathway), ligand-binding (e.g. steroid hormone receptors) and dimerisation (e.g. Fos and Jun) (Lewin 2003). In most cases, the reactions that generate these modifications are the result of equilibrium between two enzymes, each of which carries out the forward or reverse reaction (e.g. a kinase and a phosphatase with the same substrate). Modifications are often brought about by changes in the balance of the equilibrium, usually by one of the two enzymes being post-translationally modified itself. In this way, these modifications are rapidly reversible on the withdrawal of a signal.

Epigenetic mechanisms of gene control are those that do not directly rely on the DNA sequence itself, but rather on its higher order modifications and chromatin structure. They can be divided into two components; chromatin modulation and DNA methylation. The expression level of a gene is directly related to the accessibility of the gene promoter to the basal transcription machinery, and this is heavily influenced by the state of the chromatin in which that promoter resides. Chromatin that is densely packed with tightly-spaced nucleosomes is associated with transcriptional silencing, whereas open chromatin with more widely-spaced nucleosomes allows Pol II and its associated factors to reach the genes and is thus associated with transcriptional activation. Chromatin conformation is largely controlled by post-translational modifications to amino acid residues on the tails of the histone proteins that make up the nucleosome. These modifications can take a variety of forms, including acetylation, methylation, phosphorylation, ubiquitination and sumoylation

(Nightingale, O'Neill, and Turner 2006). Each type of modification contributes a distinct effect to the chromatin environment. Acetylation is the best studied of these modifications, and takes place on lysine residues in histone tails. Hyperacetylated histones are associated with more open chromatin and transcriptional activation (Schubeler et al. 2004). Hypoacetylated histones are associated with transcriptionally repressed regions (especially heterochromatin). The specific effects of a modification can depend not only on the modifying group, but also on the residue being modified and the extent of the modification. For example, methylation at lysine 4 of histone H3 is associated with transcriptionally active chromatin, with tri-methylation at this position having a higher association than mono- or di-methylation (Schubeler et al. 2004). In contrast, methylation of lysine 9 of the same histone is associated with repressed gene expression. Again, the degree of methylation is correlated with the functional implications of the modification, with mono- and di-methylation acting as euchromatic silencing markers and tri-methylation being enriched in pericentromeric heterochromatin (Rice et al. 2003). All these modifications are regulated by pairs of enzymes that either attach or remove the modifying group. These proteins are often co-regulator proteins recruited to the genome by TFs via protein-protein interactions. Many known co-activator proteins such as p300/CBP, Gcn5, and PCAF have histone acetylase activity (Sterner and Berger 2000; Roth, Denu, and Allis 2001), whereas transcriptional repressors including NCoR/SMRT and Sin3 recruit histone de-acetylase enzymes (Pazin and Kadonaga 1997; Kuzmichev and Reinberg 2001).

The other arm of the epigenetic regulatory machinery is DNA methylation. While the extent of methylation and the type of nucleotide motifs methylated varies greatly, in mammals it takes place almost exclusively on cytosines in CpG dinucleotides. Heavily methylated DNA is greatly inhibited in its ability to bind proteins. This means that genes whose flanking regions are methylated are transcriptionally silenced, as neither the basal transcription machinery nor TFs can bind. Methylated DNA can also act as a binding site for transcriptional repressor proteins that form part of repressor complexes including histone deacetylase activity, such as the Sin3 and NuRD complexes. This in turn leads to repressive chromatin states. Methylation is central to the processes of X-inactivation and imprinting (Strathdee, Sim, and Brown 2004), both of which involve the long-term silencing of particular sets of genes. The extent to which it is involved in dynamic gene regulation in normal human cells is

less clear. Examples are known of promoters being differentially methylated in different tissues in a manner that correlates with differential gene expression. These include 14-3-3σ (Umbricht et al. 2001) and HoxA5 (Strathdee et al. 2006). The RT6 gene in rats was also found to be differentially expressed in different populations of T-cells, and alterations of the methylation status of the promoter could induce or silence expression (Rothenburg et al. 2001b). However, the majority of promoters seem to be unmethylated in most tissues, including those in which the genes are not expressed.

1.1.1 A bestiary of genomic non-coding regulatory elements

Essentially all regulatory events that affect transcriptional regulation are mediated by proteins that bind to the DNA, whether these are TFs or histones, as well as any co-activator proteins that mediate indirect contact between DNA binding proteins. It is through these proteins that signals are passed from upstream in the regulatory pathway to result in the recruitment of the transcription machinery at the transcription start site (TSS). Most DNA binding proteins have some degree of specificity for the DNA sequence they bind. This allows the regulatory inputs that mediate the transcription of each gene to be controlled by the positioning of binding sites at appropriate sites in the genome such that their interactions would lead to the recruitment of Pol II at any given locus. There are several known classes of DNA elements, each of which fulfil a distinct purpose. Within each class there is a high degree of sequence heterogeneity, and very few can be predicted solely on the bases of sequence or relative positioning to other elements. Here, the major classes of regulatory DNA elements are described, and their known mechanisms of action will be briefly explained.

1.1.1.1 Promoters

Promoters were the first non-coding control elements to be discovered and studied, and are the sequences immediately flanking genes where the transcription machinery assembles before initiating the synthesis of mRNA. They usually contain a number of binding elements for various components of the basal transcription machinery, as well as for TFs that relay regulatory signals to the promoter from other sources either intra- or extra-cellular. While the individual binding sites may or may not be orientation-

dependent, the promoter itself is generally dependent on the relative order of the binding sites. Thus, most promoters are directional, although a significant proportion of them are bidirectional, and can control the transcription of genes on both strands from the same stretch of sequence (Trinklein et al. 2004). Promoters are described in more detail in section 1.2.

1.1.1.2 Enhancers/Silencers

Enhancers were among the earliest regulatory elements other than promoters to be discovered (Khoury and Gruss 1983), and are DNA elements typically no longer than a few hundred base pairs in total that cause an increase in the expression of their target genes. Unlike promoters, they have no predictable spatial relationship with the TSS, typically being found many tens of kb away from the TSS. Their effects can be exerted regardless of distance and whether they are 5' or 3' of the start of the gene (many enhancers are found within introns (Kleinjan et al. 2001; Lettice et al. 2002)). Their effects are also independent of internal orientation, and can enhance transcription of a gene even if they are reversed (Kong et al. 1997; Blackwood and Kadonaga 1998). Compositionally, enhancers have much in common with promoters in that they contain multiple binding sites for a variety of transcriptional activator proteins, which then interact with the basal transcription machinery to modulate expression. While there has been some debate about the precise mechanism of this interaction, it is now becoming increasingly clear that some form of DNA looping and interaction between proteins bound to the enhancer and promoter takes place (Carter et al. 2002; Dekker et al. 2002; Tolhuis et al. 2002) This interaction can be either direct or via intermediary proteins (Lemon and Tjian 2000). Enhancers can change the expression level of a gene significantly, sometimes by several orders of magnitude (Li et al. 2001). They can also confer tissue-specificity to the expression of the genes they regulate. For example, the enhancer for the creatine kinase gene includes binding sites for myocyte enhancer binding factor 2, a muscle-specific TF, thus restricting the expression of the gene to muscle cells. Some enhancers also allow the induction of a gene in response to an external stimulus, thus forming a distinct functional component of the regulatory machinery for a given gene or genes (e.g. the glucocorticoid response element (Yamamoto 1985; Evans 1988).

Silencer elements are functionally similar to enhancers, but act to suppress gene expression rather than promote it. As enhancers were discovered first and have been much more extensively studied, far more is known about them than about silencer elements.

1.1.1.3 Insulators

Many enhancers and silencers are gene-specific, regulating the expression of some nearby genes and not others (Butler and Kadonaga 2001) While some of this specificity may be due to the nature of the protein complexes that bind to particular enhancers and promoters, is also thought that the organisation of the genome into functional compartments, where regulatory elements only interact with other elements and genes within that compartment, plays an important role in expression regulation (Bell, West, and Felsenfeld 2001). Such compartmentalisation is partly mediated by particular DNA elements called insulators, boundary elements or enhancer blockers. These function to block interactions between enhancers on one side and promoters on the other. As such, they are position-dependent elements that only work if they are between an enhancer and a promoter and not if they are to one side of both. They are also generally orientation-independent, although some do function more efficiently in one orientation than the other (Bell and Felsenfeld 2000; Hark et al. 2000). Insulators have been most extensively studied in drosophila, but the number of known vertebrate insulator elements is rapidly increasing (West, Gaszner, and Felsenfeld 2002).

Insulators, like other DNA regulatory elements function through the binding of proteins. While a number of proteins involved in insulator function have been discovered in *Drosophila*, CTCF is currently the only protein known to fulfil this function in vertebrates (West and Fraser 2005). Several mechanisms have been proposed for insulator function. These include insulators and their associated proteins competitively inhibiting enhancer action at promoters by interacting with the enhancer proteins or sterically inhibiting enhancer-promoter interactions by sequestering them in separate chromatin loops (West and Fraser 2005). Insulators are not simply fixed and irreversible boundaries, with some having been shown to be regulated by DNA methylation (Bell and Felsenfeld 2000; Hark et al. 2000; Filippova et al. 2001). Methylated DNA blocks the binding of CTCF (and any other proteins that may bind

to that site) and thus can turn the effect of insulators on and off. Such a mechanism has been shown to be involved in the control of gene expression in at least some cases of imprinting (Kanduri et al. 2000).

1.1.1.4 Locus control regions

Locus control regions (LCRs) are DNA elements that modulate the transcriptional potential of a region of the genome, without necessarily having direct enhancer activity themselves. Like enhancer elements, their effects are position-independent, although they have also been found to depend on copy number (Carson and Wiles 1993; Li, Harju, and Peterson 1999). They are thought to exert a "priming" effect on the genes they control, rather than directly inducing transcription at particular promoters. These genes are not necessarily functionally related (Spitz, Gonzalez, and Duboule 2003), with LCRs controlling certain stretches of the genome rather than individual genes. A gene regulated by an LCR in a tissue-specific manner can sometimes be accompanied by aberrant transcription of a neighbouring "bystander" gene, even if that gene is not functionally relevant to the tissue (Cajiao et al. 2004).

LCRs seem to have different mechanisms of action depending on the particular locus. Initially, they were thought to modulate the chromatin state of the surrounding genome, thus opening up the promoters of the genes for transcription subject to further regulatory signals. This seems to be clearly the case in the growth hormone (GH) locus, where deletion of parts of the LCR results in dramatic changes to histone acetylation and chromatin conformation, and hence the expression of a transgene integrated into the site (Ho et al. 2002; Ho, Liebhaber, and Cooke 2004). However, while deletion of the LCR in the β-globin locus also abrogates gene expression, it does not alter histone modification markers at promoters or DNaseI hypersensitivity across the locus (Schubeler, Groudine, and Bender 2001; Sawado et al. 2003). A number of mechanisms have been proposed for individual well-studied LCRs that involve the induction of complex chromatin loops by proteins binding to individual sites within the LCR. There is also a proposal that some LCRs function by controlling the localisation of the DNA containing the genes themselves into transcriptional factories within the nucleus (Ragoczy et al. 2003). There seems to be no single model

that universally applies to LCR function, and it is an interesting area for further research.

1.1.2 Transcription Initiation in Eukaryotes

Human cells contain three functionally distinct RNA polymerase enzymes, each of which is responsible for the transcription of different kinds of RNA molecules. RNA Pol I transcribes ribosomal RNA (rRNA), and accounts for the majority of RNA polymerase activity in the cell by quantity. RNA Pol III transcribes tRNAs and other small non-coding RNAs. RNA Pol II is responsible for transcribing mRNA from protein-coding genes, and as such is at the apex of regulatory processes that regulate the production of proteins and the phenotypic destiny of the cell. The basic mechanism of transcription initiation at Pol II promoters has been well-characterised for a certain class of promoter containing a TATA-box (see later), though the mechanism in other promoter classes is less clear. The assembly of the transcription machinery and escape of Pol II have been the subject of many detailed reviews and textbook chapters (Dvir, Conaway, and Conaway 2001; Lewin 2003), and as such will be covered only briefly here. The RNA Pol II holoenzyme itself is not capable of sequence-specific binding to DNA on its own, and requires the presence of numerous other proteins in order to recognise the promoter accurately and carry out high levels of transcription. These additional components are called basal transcription factors, to distinguish them from other families of TFs.

The first step in the initiation mechanism is the binding of the basal TF $TF_{II}D$ to the promoter a few bases upstream of the TSS. $TF_{II}D$ is itself made up of multiple protein subunits, consisting of TATA-binding protein (TBP) and a set of TATA-associated factors (TAFs) in varying proportions. The TBP component recognises the TATA box, and is the key element in correctly positioning the initiation complex in TATA-containing promoters. A series of factors subsequently binds in the following order; $TF_{II}A$, $TF_{II}B$ and $TF_{II}F$. With each additional factor bound, the DNA footprint of the pre-initiation complex increases. Only after $TF_{II}F$ binds does the Pol II holoenzyme join the complex. Transcription begins on binding of $TF_{II}E$, and the phosphorylation of the pol II carboxy-terminal domain by another basal factor, $TF_{II}H$.

In TATA-less promoters, $TF_{II}D$ retains its role in positioning the complex, but is able to recognise other promoter motifs, particularly the initator sequence described in section 1.2.1.2 (Smale 1997). The TAFs making up a given $TF_{II}D$ are also important in promoter sequence recognition. The place of TBP in the $TF_{II}D$ complex is sometimes taken by a similar protein, TBP-like factor (TLF). This protein, which is 60% similar to TBP, is expressed in all multicellular organisms. It does not bind the TATA box, and its mechanism is not known, but it likely plays a role in initiation from some TATA-less promoters.

1.1.3 The position of the promoter in the regulatory framework of the cell

The process of gene expression, from DNA to finished protein, can be regulated at multiple points. These include the rate or timing of transcription initiation, the stability of the primary and processed mRNA transcript, the rate of translation and the regulation of post-translational modifications on the protein. While examples exist of regulatory influences at many of these stages *in vivo*, it is a widely-held view that the most crucial point of control is the initiation of transcription (Lewin 2003; Wray et al. 2003; Buckland 2006). This is a difficult fact to quantify definitively, as it would theoretically require complete knowledge of the regulatory pathways of every gene. However, all post-transcription control mechanisms require the presence of at least a primary transcript, and thus require transcription to be taking place before they can function. In addition, they are mostly inhibitory or destructive mechanisms, for example involving the degradation of transcript or protein. They are therefore not able to cause induction of genes in response to intra- or extra-cellular stimuli, and can only modulate the amount of gene product being produced from a gene that is already being transcribed.

This places the promoter in a crucial position in the regulatory hierarchy of a gene (Figure 1). The majority of signalling mechanisms known terminate with a change in the activity of a TF or some other method of changing the degree of transcription initiation described above. The promoter is essentially that of a logical signal integrator, where a wide range of regulatory inputs come together and are processed to produce a single scalar output; the rate of transcription initiation. This, plus the fact that promoters are the only regulatory elements with a predictable spatial relationship

to genes (Trinklein et al. 2003) make them prime candidates for the study of regulatory variation.

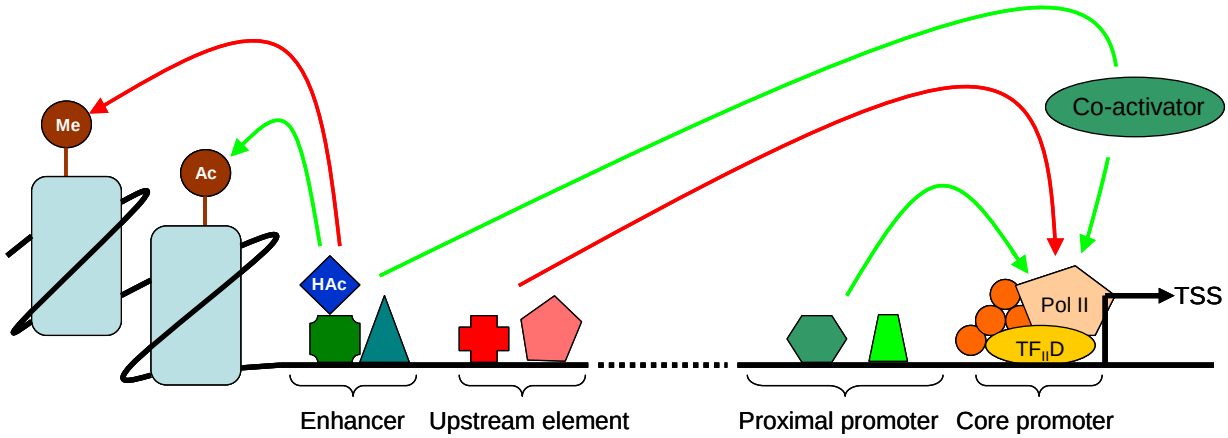

Figure 1. Diagrammatic representation of regulatory inputs into promoter function. The core promoter is the binding site for the basal transcription machinery, including RNA Pol II and $TF_{II}D$ and other general TFs. Both in the proximal promoter and in upstream enhancers and silencers are binding sites for a wide range of TFs which are in turn influenced by a myriad of cellular signalling pathways that relay information from intra- and extra-cellular sources. These TFs can have both stimulatory (green) and inhibitory (red) effects on the stability of the Pol II complex. These effects can be mediated by both direct contact between the factors and the Pol II complex, or by contact through intermediary co-activator proteins. Upstream elements can also affect transcription initiation by recruiting chromatin modification proteins (blue) such as histone acetylases. These then modify the tails of nearby histones to modify the chromatin into either more permissive (shown) or less permissive conformations depending on the enzymes recruited.

1.2 The Eukaryotic Promoter

The function of the eukaryotic promoter sequence itself can be split into two components; the definition of the correct TSS and orientation of the transcript, and the capacity to receive regulatory signals that govern that timing of transcription initiation. The former involves direct interaction with the basal transcription machinery in order to orient it with respect to the TSS, whereas the latter is regulated by the binding of TFs. This requirement for the binding of distinct entities gives rise to a functional partitioning of the promoter. However, the boundaries of these two arbitrary functional units are difficult to define for any specific promoter, as there is considerable heterogeneity in the functional motifs present in each promoter and their *in vivo* functionality is dependent on chromatin state and TF complement.

In vertebrates, the sequence feature most characteristic of promoters is their correlation with CpG islands. Vertebrate genomes in general contain only 20% of the

CG dinucleotides that would be expected from the base composition (Antequera 2003). This is because CpG's are targets of methylation on the cytosine residue, and the majority of such sites (around 80%) are methylated at any one time. Methylated CpG's are highly susceptible to mutation by deamination of the methyl-cytosine, converting it to a thymine (Figure 2). However, DNA methylation is also associated with transcriptional silencing when in the vicinity of genes, so methylation is generally reduced or absent in areas where gene expression is occurring. Unmethylated CpG's do not mutate any faster than other dinucleotides, and therefore consistently unmethylated genomic regions have CpG frequencies close to the expected level, and are called CpG islands (Gardiner-Garden and Frommer 1987; Wasserman and Sandelin 2004). An interesting question is whether hypomethylation of CpG islands flanking genes is a cause or a consequence of their status as promoters. That is to say, are promoters hypomethylated so that they can have promoter activity, or are they hypomethylated due to their interactions with DNA binding proteins or chromatin as a result of promoter activity? The fact that the majority of intergenic DNA is methylated implies the existence of a mechanism to either prevent methylation of promoters or to demethylate them after global DNA methylation, and in turn suggests that promoters are hypomethylated in preparation for their role as promoters. However, while it seems unlikely that hypomethylation is simply the passive result of a protective effect of the binding of TFs (as even untranscribed genes are often hypomethylated (Strathdee, Sim, and Brown 2004)), no human DNA demethylases have ever been discovered. The precise position of methylation in the evolution of gene regulation remains unknown. Even though they are the most common sequence characteristic of promoters, only around 60% of human promoters are found in CpG islands (Antequera and Bird 1993; Antequera 2003).

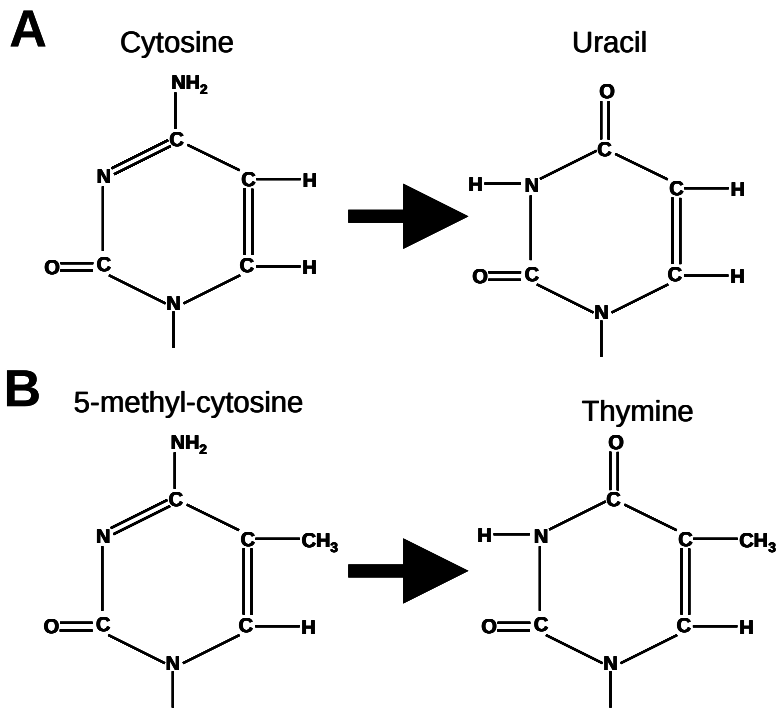

Figure 2. Deamination of cytosine and methylcytosine produce different bases. A) Cytosine bases are prone to spontaneous deamination, producing uracil as the resulting base. This is efficiently detected and repaired by the cellular repair machinery. B) Methyl-cytosine bases are prone to the same process, but due to the extra methyl group produce thymine on deamination. This makes it much less likely to be detected and repaired, leading to a higher probability that the mutation would become fixed into a daughter cell following the next round of DNA replication.

1.2.1 The Core Promoter

The "core promoter" is the sequence up to 40 base pairs upstream from the TSS, and contains the sequence elements that are bound by the Pol II complex. The "proximal promoter" is further upstream from the core promoter, and its extent is not currently definable from sequence information alone, as it is made up largely of transcription factor binding sites (TFBSs) that are themselves difficult to rigorously define (see later). The core promoter is the better understood of the two functional units, and the few promoter motifs that are well-characterised belong in this region (Figure 3).

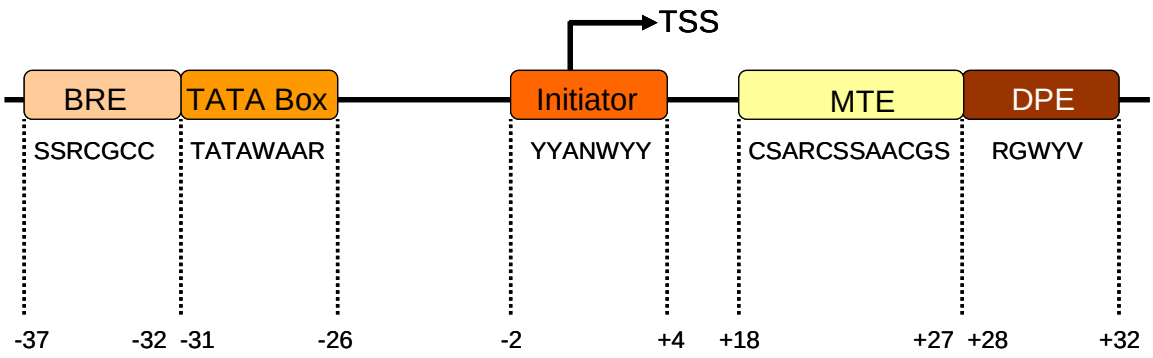

Figure 3. Known core promoter motifs in human promoters. The positions of the motifs are shown relative to the transcription start site (TSS), designated as base +1. Each of the elements is described in detail below. The consensus sequences are also shown in IUPAC ambiguity code notation. Figure adapted from (Jin et al. 2006).

1.2.1.1 TATA Box

The TATA box is an AT-rich element found at -25 to -30 bases from the TSS, with a consensus sequence of TATAWAAR . It was the first promoter element ever found in eukaryotes, and was identified by aligning viral, mammalian and *Drosophila* promoter sequences (Breathnach and Chambon 1981). Following its discovery, it was believed to be a near-ubiquitous and essential motif for transcription from Pol II promoters, particularly as it was repeatedly shown that introducing mutations in the TATA box sequence severely reduced if not eliminated transcription in *in vitro* systems, as well as displacing the TSS (Grosschedl and Birnstiel 1980; Wasylyk et al. 1980; Hu and Manley 1981). However, it is now known that TATA-containing promoters form a minority in most eukaryotic genomes. A survey of 1941 *Drosophila* promoters found a TATA sequence within one mismatch of the consensus in only 33% of promoters. In humans, a similar survey found 32% of 1031 Pol II promoters contained TATA boxes (Suzuki et al. 2001). More recent computational surveys have suggested that this figure is in fact only 20% (Jin et al. 2006).

The TATA box acts as a recognition site for the TATA binding protein (TBP), a key component in the assembly of the Pol II complex. X-ray crystallography of TBP bound to oligonucleotides containing strong TATA boxes suggested that the binding was unidirectional, and implied a role for the TATA box in determining transcript orientation. However, TBP only shows a moderate preference for binding in the forward orientation in solution, and artificially reversing the orientation of the TATA

box in the context of a complete promoter sequence failed to produce a reversal of the transcript (Xu, Thali, and Schaffner 1991; O'Shea-Greenfield and Smale 1992). Instead, the major role of the TATA box *in vivo* seems to be to regulate the location of the TSS at a certain distance downstream of it, with RNA Pol II itself and $TF_{II}B$ playing a crucial role. This was elegantly demonstrated in a study where basal TFs and Pol II holoenzyme from *S. pombe* were transferred to a *S. cerevisiae* system. *S. pombe* Pol II and $TF_{II}B$ were able to shift the TSS from 40-120 base pairs downstream of the TATA box (in native *S. cerevisiae*) to 30 bases downstream (as in native *S. pombe*) (Li et al. 1994).

1.2.1.2 Initiator

The initiator element encompasses the TSS, and has the consensus sequence YYANWYY in mammals (Smale and Baltimore 1989; Jin et al. 2006) with the adenosine residue in the sequence being the TSS (base +1). Though earlier work had suggested that the sequence immediately around the TSS was important in the maintaining the efficiency and precision of transcription initiation both in TATA-containing and TATA-less promoters (Talkington and Leder 1982; Dierks et al. 1983; Concino et al. 1984), it was first rigorously characterised in the TATA-less promoter of the terminal transferase (TdT) gene (Smale and Baltimore 1989). In this study, analysis of mutations across the TdT promoter showed that the -3 to +5 sequence was essential to accurate transcription for this gene (Smale and Baltimore 1989; Javahery et al. 1994).

Functionally, the initiator performs a similar role to the TATA box, providing a binding site for the basal transcription machinery and regulating the location of the TSS. When an initiator and a TATA box are found together in the same promoter, their behaviour is determined by their relative positions. If the TATA box is present in the -25 to -30 range relative to the initiator (hence the TSS), the two elements behave synergistically (O'Shea-Greenfield and Smale 1992), whereas if they are separated by more than 30 base pairs, they act independently. If they are spaced between 15 and 20 base pairs apart they continue to act synergistically, but interestingly the TSS is shifted to a position 25 base pairs downstream of the TATA box, regardless of the position of the initiator.

1.2.1.3 Downstream Promoter Element (DPE)

The DPE is unusual in that it is found downstream of the TSS, and is thus part of the 5' untranslated region (UTR) of the gene to which it belongs. It is a 5 base pair motif with the consensus sequence RGWYV, and is found in the +28 to +32 region relative to the TSS (Kutach and Kadonaga 2000; Jin et al. 2006). Most DPE-containing promoters are TATA-less and contain an initiator, with the DPE and initiator acting synergistically as a $TF_{II}D$ binding site. The DPE is unable to bind $TF_{II}D$ alone, and perturbation of the precise spacing between the DPE and initiator elements in a DPE-containing promoter drastically reduce initiation efficiency (Burke and Kadonaga 1996).

Although promoters exist with both DPE and TATA box elements, their function seems to be very similar, with both acting as binding sites for $TF_{II}D$. Their similarity is demonstrated by the fact that if transcription from a promoter is abrogated by mutations in its TATA box, transcriptional activity can be restored by the addition of a DPE in the appropriate location (Burke and Kadonaga 1996).

1.2.1.4 TFIIB Recognition Element (BRE)

This element is present in a subset of TATA-containing promoters, and is found immediately upstream of the TATA box, approximately in the -37 to -32 base pair range. It was originally discovered in archaea (Reiter, Hudepohl, and Zillig 1990; Hain et al. 1992), but its existence has also been demonstrated in humans (Lagrange et al. 1998). Its 7 base pair consensus sequence in humans is SSRCGCC, and binds to $TF_{II}B$ (Nikolov et al. 1995; Lagrange et al. 1996; Jin et al. 2006). Its precise function in humans is unclear, as there is evidence that it is involved in both transcriptional activation (Lagrange et al. 1998) and repression (Evans, Fairley, and Roberts 2001). However, it is the only known promoter element to date that binds a factor not associated with $TF_{II}D$ (apart from the MTE, whose binding protein is unknown, see section 1.2.1.5).

1.2.1.5 Motif ten element (MTE)

The MTE element was discovered relatively recently by a scan for over-represented motifs in *Drosophila* promoters (Ohler et al. 2002), and is conserved through mouse and human with a consensus of CSARCSSAACGS (Jin et al. 2006). *In vitro* transcription and luciferase reporter studies on promoters containing wild type and artificially-mutated MTEs demonstrated that it was indeed a functional promoter element, and that mutations could abrogate transcriptional efficiency (Lim et al. 2004). The MTE requires the presence of an initiator element, but is independent of TATA boxes and DPEs and can compensate for the removal of the latter two elements *in vitro* (Lim et al. 2004). The same study also demonstrated synergistic effects between the MTE and TATA box and between MTE and DPE. The factors that bind to this element have not yet been determined.

1.2.2 The Proximal Promoter

In general, an isolated core promoter can initiate transcription only at low levels (Lemon and Tjian 2000). The temporal and scalar control required to maintain robust gene expression is conferred by the binding sites in the proximal promoter. Variable as core promoter sequences are, the proximal promoter is even less well-defined. Functionally, it can be regarded as an array of binding sites used by TFs to relay signals to the basal machinery. There is no agreement as to how far upstream the core promoter extends, where one can draw a boundary between a promoter element and an enhancer element, and even where the core promoter ends and the proximal promoter begins. For example the CCAAT-box, a motif located 75-80 base pairs upstream of the TSS, has been variably classified as part of the core promoter (due to its relative invariability compared to other TFBS) and as part of the proximal promoter (due to its distance from the TSS).

The functional characteristics of the proximal promoter depend on the binding sites present in it and the relative spacing and clustering between them. Each TFBS can function individually when binding a TF, or can form a cluster with other sites that can bind multimeric TFs. Each TFBS or TFBS cluster can function as a modular

element, relaying separate signals to the transcription initiation complex either directly by DNA looping and protein-protein interactions, or indirectly via transcriptional cofactors, of which there are a large number (Chen 1999; Lemon and Tjian 2000). The TFBS complement of promoters will vary greatly depending on the characteristics of the gene it regulates. There are therefore few if any rules about the binding sites and relative positioning to be expected in a typical promoter. However, recent deletion studies of a set of promoters in the ENCODE regions has suggested that, on average, the promoter as far as 300 bases upstream tends to contain elements that promote transcription, whereas the -500 to -1000 base pairs contain more negative regulatory elements (Cooper et al. 2006).

1.3 Identifying promoters

For the mechanistic and sequence basis of promoters to be studied with any degree of confidence, it is essential that the promoters themselves be identified against the genomic background. Given the considerable length of the human genome and the economic and labour cost of functionally characterising the regulatory properties of that much sequence, *in silico* methods for predicting promoters have long been an important goal for the bioinformatics community. The development of such methods faces significant hurdles due to the functional and sequence heterogeneity of promoters. In parallel, efforts are also underway to design high-throughput experimental promoter screening methods that, even if unable to elucidate every single possible promoter in the genome, could return a robust training set of verified promoter sequences for use in furthering the *in silico* research.

1.3.1 Computational approaches

The array of binding sites for basal transcription apparatus and TFs described above may give the impression that overall promoter function is well-understood, and that searching for these binding sites is sufficient to identify promoters based only on their sequence. However, this is far from being the case. Binding sites are not fixed sequence motifs, but are usually tolerant of substitutions without necessarily losing function, provided the affinity of the TF to the site is not affected. TFBS can be described in terms of a position weight matrix (Bucher 1990), which describes the

probability of each base being any one of the four possible bases. Because of the looseness of many of these binding sites, and because a typical binding site is very short (typically 5-7 nucleotides, almost never exceeding 25 nucleotides), a given promoter will contain a great number of binding sites simply by chance. Only a small number will be functional. Indeed, any stretch of the genome regardless of whether it is regulatory or not is bound to contain these sequences by chance. There are a variety of databases available that contain the weight matrices of known TFs (Wingender et al. 1996; Sandelin et al. 2004). These are often used to scan putative promoter sequences for binding sites, but these must be considered highly provisional in the absence of experimental data confirming their functionality.

Promoters are generally located immediately 5' of their TSSs. As such, early promoter prediction algorithms were in reality TSS predictors that would look for the known promoter elements described above, such as the TATA box, and attempt to place a TSS using these elements as a guide. The common occurrence of these binding sites, and the fact that a minority of promoters contain any one of them, led to a very high false positive rate (Fickett and Hatzigeorgiou 1997). Since then, a whole range of different promoter predictors has been released, each using different computational methods such as artificial neural networks (ANN), various Markov models, relevance vector machines and statistical methods for comparing sequence. Sequence properties and criteria used as the basis for promoter and TSS prediction have included (see Table 1 for references);

- Presence of CpG islands
- TATA boxes and other core promoter motifs and their relative positions
- Increased clustering of TFBSs
- Combinations of TFBSs and core motifs in particular positional arrangements
- Motifs overrepresented in training sets of experimentally derived promoters
- Statistical properties of sequence composition
- Downstream first exons and donor splice sites
- Deep evolutionary conservation

While some of these tools initially reported promising results on small datasets, subsequent application to whole genome promoter prediction has yielded

disappointing results (Bajic et al. 2004). All tools tested on whole genome data to date suffer from one of two problems; a very low sensitivity measured by the number of known promoters predicted, or a high false positive rate (Table 1). In many cases they were not even as good at predicting known promoters as a simple scan for CpG islands. Indeed, non-CpG island-containing promoters are an area where most predictors perform particularly badly. Combining two different promoter prediction algorithms can improve the false positive rate, although any increase in sensitivity, as measured by the number of known promoters predicted, is only modest (Bajic et al. 2004).

With the advent of multiple vertebrate genomes, as well as multiple closely related non-vertebrate species such as *Drosophila* or yeast, evolutionary conservation is now becoming a common criterion for detecting functional elements (Ahituv et al. 2005; Dermitzakis, Reymond, and Antonarakis 2005; King et al. 2005; Siepel et al. 2005; Xie et al. 2005; Robertson et al. 2006). Promoters in general are more highly conserved than non-genic sequence, although the degree of conservation may be related to the functional classification of the gene (Iwama and Gojobori 2004; Suzuki et al. 2004). Such studies have tended to focus on the discovery of regulatory elements and motifs in general rather than restricting themselves to promoters per se. The existence of such highly conserved non-coding regions both as distinct elements and as shorter sequences within known elements is regarded as strong evidence of their functional significance. However, there is little agreement on what these functions might be, and currently no easy way of differentiating between possible different functions (e.g. some may be enhancers or LCRs, and others may be sequences involved in matrix attachment).

Program	Details	Sensitivity %	Ppv %	True positive cost	Reference
CpGProD (0.0)	Statistical rule-based system. Detects only CpG-island-related promoters	47.26 47.26	51.84 51.84	0.9290 0.9290	(Ponger and Mouchiroud 2002)
CpGProD (0.3)	Statistical rule-based system. Detects only CpG-island-related promoters	37.09 37.09	69.79 69.79	0.4329 0.4329	
DragonGSF	ANN, concept of CpG island combined with predictions of DragonPF	65.21 61.79	62.99 64.80	0.5876 0.5432	(Bajic and Seah 2003b; Bajic and Seah 2003a)
DragonPF (50%)	ANN, overlapping pentamer matrix models of promoters, exons and introns. Separate modules for promoters in G+C-rich and G+C-poor regions	56.05 53.85	21.30 32.23	3.6940 2.1032	(Bajic et al. 2003)
DragonPF (55%)	ANN, overlapping pentamer matrix models of promoters, exons and introns. Separate modules for promoters in G+C-rich and G+C-poor regions	67.65 64.68	19.68 30.43	4.0808 2.2863	
DragonPF (65%)	ANN, overlapping pentamer matrix models of promoters, exons and introns. Separate modules for promoters in G+C-rich and G+C-poor regions	80.93 77.28	15.05 24.62	5.6454 3.0611	
Eponine	Relevance vector machine based on a TATA-box motif in a G+C-rich domain	40.08 39.91	66.98 67.33	0.4929 0.4852	(Down and Hubbard 2002)
FirstEF	Quadratic discriminant analysis of promoters, first exons and first donor site. Uses concept of CpG island	80.98 79.41	35.18 39.37	1.8427 1.5400	(Davuluri, Grosse, and Zhang 2001)
FirstEF (CpG-)	Quadratic discriminant analysis of promoters, first exons and first donor site. Uses concept of CpG island	4.38 4.12	5.61 6.25	16.8408 15.0064	
FirstEF (CpG+)	Quadratic discriminant analysis of promoters, first exons and first donor site. Uses concept of CpG island	76.99 75.64	50.52 55.57	0.9793 0.7995	

Program	Details	Sensitivity %	Ppv %	True positive cost	Reference
NNPP2.2 (0.90)	Three time-delay ANNs trained to recognise TATA box and initiator, as well as their mutual distance	92.77 77.12	2.78 4.08	35.0159 23.5194	(Reese 2001)
NNPP2.2 (0.95)	Three time-delay ANNs trained to recognise TATA box and initiator, as well as their mutual distance	85.43 69.00	3.02 4.41	32.1452 21.6587	
NNPP2.2 (0.99)	Three time-delay ANNs trained to recognise TATA box and initiator, as well as their mutual distance	56.50 43.32	4.27 6.11	22.4452 15.3734	
Promoter 2.0	ANN trained to recognise a combination of four TFBSs (TATA box, CCAAT-box, GC-box, initiator) and their mutual distances	57.23 44.07	3.27 4.90	29.6203 19.4289	(Knudsen 1999)
McPromoter (+0.005)	ANN, interpolated Markov model, different physical properties of promoter regions and statistical properties of promoters versus non-promtoers	27.13 26.96	78.39 87.08	-	(Ohler et al. 2002)
McPromoter (-0.005)	ANN, interpolated Markov model, different physical properties of promoter regions and statistical properties of promoters versus non-promtoers	55.65 54.96	70.95 79.20	-	

Table 1. Data on the whole genome application of a representative set of promoter prediction algorithms. This was carried out by Bajic and colleagues, and the data was obtained from Bajic et al 2004. Some programs were run with several different parameters, and these are detailed in brackets underneath the program name. The top set of numbers in each cell shows the results without Repeatmasker, and the lower set with Repeatmasker in use. Further details on the algorithms can be found in Bajic et al 2004. Sensitivity is defined as the percentage of known promoters in the genome correctly predicted by the algorithm. The true positive cost is the number of false positives predicted for every true positive. McPromoter was only tested on chromosomes 4, 21 and 22, and no true positive cost was calculated. ANN = artificial neural network, ppv = positive predictive value.

1.3.2 Experimental approaches

The development of new technologies for genome-scale functional interrogation of non-coding DNA and the decreasing cost of doing large experiments has resulted in an increasing focus on scanning the genome for promoter elements in an unbiased manner, without necessarily relying on *in silico* predictions beforehand. The classical method for functionally characterising putative promoters has been to clone them into a reporter plasmid, transfect them into an *in vitro* model system (either cultured cells or model organisms) and then carry out nested deletions to determine the boundaries of the minimum sequence necessary to drive expression. However, this is a labour intensive procedure that required the determination of putative promoters beforehand, such as the presence of a confirmed TSS.

During the human genome project the 5' ends of genes, and hence TSSs, were annotated using evidence such as ESTs, cDNA libraries and gene prediction software (Collins et al. 2003; Consortium 2004a). These all have a certain degree of uncertainty associated with their designation of gene starts; for example it is difficult to guarantee that cDNAs in a library are indeed full length, as unlike the 3' end there are no sequence features that identify the 5' end of a cDNA. Various promoter-trapping technologies were also developed over the last 15 years to screen for promoters *de novo*. Initially, these were based on the gene trap vectors used to determine expression patterns in model organisms (Stanford, Cohn, and Cordes 2001). They functioned by integrating a retroviral-based reporter vector into a cell line, or in some cases a model organism, and detecting the expression of the reporter if integrated downstream of a promoter. Genomic DNA would then be prepared from positive clones, and the sequence flanking the integration rescued by PCR or restriction enzyme digestion followed by self-ligation. More advanced vectors and reporter enzymes then enabled the direct cloning of libraries of random genomic fragments followed by vector recovery and resequencing to identify putative promoters. The most successful of these systems to have been applied in a large-scale study was developed by Myers and colleagues at Stanford (Khambata-Ford et al. 2003), and a screen of a whole genome fragment library isolated 244 putative promoters that aligned to the 5' end of an annotated gene or to a CpG island. This was only 28% of all fragments isolated, and although a further 20% had some evidence of

promoter activity from the genome annotation (e.g. aligning upstream of a gene predicted by a single annotation program only) nearly half of the fragments recovered did not align anywhere near the start of a gene or any other sequence feature to suggest promoter activity. Thus systems such as these also seem to suffer from a high rate of noise and false positives. Interestingly, although 70% of all isolated putative promoters in this study did not align near a known TSS, 86% were capable of promoter activity in a reporter assay. This implies that either there are still a considerable number of genes that have not yet been discovered, or that many intergenic DNA sequences can function as promoters if placed in a context where they are accessible to the transcription machinery. Evidence of extensive transcription taking place outside annotated genes lends weight to the idea that, rather than being experimental noise, extraneous hits from experimental promoter screens may reflect this extra transcription.

There has been more success in the application of novel methods for capturing the 5' ends of processed mRNA transcripts, such as 5'-end serial analysis of gene expression (5' SAGE) and cap analysis of gene expression (CAGE) (Shiraki et al. 2003; Hashimoto et al. 2004). These make use of the Gppp cap at the 5' end of mRNA in order to capture transcripts with intact 5' ends. Biotinylated linkers containing a recognition site for a type IIs restriction enzyme (which can cleave several tens of bases away from its binding site) are used to purify short sequence tags from the start of the transcripts. These are then ligated together and sequenced at high throughput, and clusters of tags mapped to the reference genome point to TSSs. These techniques are capable of experimentally confirming TSSs more rigorously than before, and have cast doubt on the idea that one promoter necessarily contains one functional TSS (Carninci et al. 2006). A recent whole-genome analysis of multiple CAGE libraries from human and mouse reveal that promoters can be grouped into different classes depending on the profile of their TSSs. While some promoters have a tightly-defined single TSS as per the classical definitions, there are promoters with broadly-defined start sites spread over many tens of bases, with a dominant start site surrounded by minor start sites, and even with two or more highly-specific start sites (Carninci et al. 2006). Promoters with tightly defined start sites were more likely to contain TATA-boxes, and promoters with less well-defined initiation profiles were more likely to be in CpG islands (Carninci et al. 2006).

In the last few years, the development of ChIP-chip technology has been the most significant technical development in enabling the interrogation of protein-DNA interactions *in vivo* and on a genomic scale (Ren et al. 2000). ChIP (or chromatin immunoprecipitation) is a well-established technique for purifying DNA fragments that bind to particular proteins. Briefly, cells are treated with a chemical agent that cross-links any proteins bound to DNA covalently. The cells are lysed and the genomic material containing the cross-linked proteins is sheared into small fragments of 300-500 bases. An antibody is used to immunoprecipitate the protein of interest, thus also precipitating the DNA fragments bound to it. The cross-linking can be reversed by heat and acid hydrolysis, liberating the DNA fragments for analysis. When this technique was first developed, the analysis of the precipitated DNA fragments would be done by PCR amplification with primers targeted to specific regions. The recent innovation is to analyse all the precipitated DNA fragments at once by PCR-amplifying and fluorescently labelling it before hybridising it on to a microarray. In this way, enrichment for any given fragment can be detected over a control DNA preparation labelled with a different fluorophore. Given an appropriate antibody to a TF or other DNA binding protein, the extent of the genome that can be analysed for enrichment in a ChIP experiment, and hence binding of the protein of interest, is limited only by the coverage of the mircoarray. Extensive work has been carried out to map the action of TFs in *Saccharomyces cerevisiae*, and these have progressed to the point where one study has mapped 106 TFs across the whole yeast genome using antibodies to epitope-tagged TFs (Lee et al. 2002). The binding characteristics of a number of TFs have been mapped using microarrays covering a variety of genomic regions and elements. These include p53, Sp1 and c-Myc (Cawley et al. 2004), CREB (Euskirchen et al. 2004) and NFκB (Martone et al. 2003), which have been mapped on chromosome-scale tiling arrays. HNF (Odom et al. 2004) and c-Myc again (Li et al. 2003) have been studied genome-wide using arrays of PCR-amplified promoter fragments. All these studies have been important in understanding the regulatory connection between genes. The most interesting studies from the point of view of promoter discovery however have been using antibodies to components of the basal transcription machinery, such as $TAF_{II}D$ or RNA Pol II itself (Kim et al. 2005a), using a series of tiling arrays covering the whole genome. These have allowed true genome-scale examination of the assembly of pre-initiation complexes (PIC), and

hence the presence of promoters *in vivo*. The first genome-wide survey of active promoters in a cell line has recently been completed (Kim et al. 2005b), paving the way for such studies in cell lines of diverse tissue origins. Such studies will be invaluable in deciphering the regulatory logic behind the establishment of different tissues.

Initial whole-genome ChIP-chip surveys in a human cell line have indicated that a substantial number of promoters remain to be discovered (Kim et al. 2005b). While many of these appear to be alternative promoters to known genes, there is also evidence that a significant fraction come from novel transcriptional units. Many of these regions of PIC assembly also have other evidence of promoter function, such as the presence of ESTs and enrichment for putative promoter elements such as CpG islands. This ties in with evidence from expression microarray studies that there is extensive expression from regions outside the annotated protein coding gene set (Kapranov et al. 2002; Rinn et al. 2003; Cheng et al. 2005). The physiological importance of these transcripts is still unclear, but their existence suggests that there are entire classes of sequences that are capable of driving expression, whether cryptically or otherwise, that we cannot yet identify. The rate of "novel" fragments capable of promoter activity from large-scale promoter screens is also suggestive of this (Khambata-Ford et al. 2003). It may in part explain some of the difficulties in identifying promoters both *in silico* and *in vitro*.

1.4 Variation in promoter sequences

To a first approximation, promoters are subject to the same mutational forces as shape the rest of the genome, with the exception of cytosine deamination in constitutively unmethylated CpG island promoters. The spectrum of variation present in promoters encompasses SNPs, indels including transposable elements, microsatellites and other repeat length polymorphisms. Unlike in coding sequence, where classification of mutations as synonymous or non-synonymous is relatively trivial, it is impossible to determine the functional consequences of a promoter polymorphism from a simple examination of its sequence, due to the functional ambiguity of regulatory DNA in the absence of experimental data.

The pre-eminent method of testing the effect of polymorphism on promoter efficacy has long been the transient transfection reporter assay, where the variant promoter alleles are each cloned into a promoter-less plasmid containing a reporter gene such as CAT, firefly luciferase or GFP (Alam and Cook 1990). Each plasmid is then transfected into the *in vitro* model system of choice, usually a transformed cell line. A constitutively active control plasmid containing a separate reporter is often co-transfected with each allelic construct, in order to control for experimental variables such as transfection efficiency and enable direct comparison between the results for each allele. Polymorphisms in the many human promoters have been investigated in this way, usually because of some clinical interest in the downstream gene (Rockman and Wray 2002). A search on PubMed for papers detailing such experiments yields in excess of 300 papers at the time of writing. Many of these promoter assays have been accompanied by EMSA experiments or association studies linking a promoter variant to some disease phenotype (Rockman and Wray 2002). However, the wide variety of cell lines and experimental technologies used makes sophisticated meta-analyses of this body of work problematic, as each cell line contains its own complement of TFs. Only recently have such assays begun to be applied to larger sets of genes using the same cell lines, making the prospect of a global analysis of *in vitro* functional SNPs more plausible (Buckland et al. 2005). These studies suggest that 22% of promoters contain sequence variants that affect promoter strength in a reporter assay. However, this is likely to be an underestimate, as the small ethnically diverse panel used for SNP discovery in these papers may have led to an ascertainment bias away from rare SNPs. This is because the likelihood of detecting a SNP in a panel is proportional to its minor allele frequency, making rare SNPs unlikely to be detected in small panels. Carrying out such experiments remains labour-intensive, and with over 12 million human SNPs in dbSNP at the time of writing, testing every polymorphism in a putative promoter is still economically ambitious. As long as this remains the case, a computational method of functional prediction will remain desirable, and this will depend to a large extent on establishing representative experimental datasets of functional variation in humans.

A reasonable hypothesis would be that a SNP within a TFBS is likely to affect the binding of the associated TF, whereas one outside a binding site is more likely to be neutral. However, the short sequences of typical TFBS means that any given sequence

is very likely to contain a large number of sites, with only a small minority being functional *in vivo*. Discriminating between these functional sites and the background of false positives is currently very difficult without experimental data. Multiple lines of evidence can be used to gain more certainty of the importance of some sites. For instance, if a binding site is for a TF known to function as a multimer, either with itself or other factors, the coordinate presence of the binding sites at appropriate spacing would be indicative of functionality. Also, many binding sites have relatively loose weight matrices and can withstand base substitutions at many positions with only a modest effect on the affinity of the TF to the site. This means that the impact of functional polymorphisms can be drastic or subtle depending on the position weight matrix of the binding site in question. In contrast, polymorphisms outside of binding sites, whether predicted or experimentally confirmed, cannot necessarily be dismissed as non-functional, as they can affect the conformational properties of the DNA or the relative spacing of functional TFBS, thereby influencing their interactions with the Pol II complex (Rothenburg et al. 2001a).

A recent study predicted a set of 36 from 200 promoter SNPs would be functionally significant using comparative genomics and predicting the effect of the binding sites (Mottagui-Tabar et al. 2005). 7 out of the 10 SNPs tested in mobility shift assays showed an effect on TF binding, suggesting that it is possible to predict the effect of a SNP on the affinity of protein binding *in vitro* with moderate accuracy. However, it is still unclear how this translates into *in vivo* function, as only four SNPs were tested in luciferase reporter assays, and of these only two showed significant differences in promoter strength.

1.5 Natural variation in gene expression levels

There is now a significant body of evidence to indicate that heritable variation in gene expression between individuals is widespread. This has come largely from expression microarray studies in model organisms (Brem et al. 2002) and humans (Cheung et al. 2003; Monks et al. 2004). More recently, there have been several association studies that have identified SNPs that are associated with expression phenotype (Monks et al. 2004; Morley et al. 2004; Cheung et al. 2005; Deutsch et al. 2005; Stranger et al. 2005). These were done by large-scale genotyping of the SNPs across the genome

combined with expression arrays to measure variation in gene expression, followed by association analysis to find genes linked to expression phenotypes. Taken together, these studies suggest that 25-30% of functional regulatory variation acts in *cis*-, with the remainder acting in *trans*- (Pastinen, Ge, and Hudson 2006). A recently completed whole genome association analysis of expression phenotypes on the entire HapMap set of 210 parents has recently been completed (Stranger *et. al.* unpublished), giving the first truly whole genome picture of the extent of heritable gene expression phenotypes. It is often difficult to distinguish between *cis*- and *trans*- acting SNPs discovered in these experiments, especially as the definition of these terms is not universally agreed. Many would define a *cis*- variant as directly influencing the expression of the gene whose phenotype it is associated with. If it is in the promoter region it may influence the binding of TFs, or if it is in an enhancer element further upstream it can disrupt the normal interactions of the enhancer with the promoter. A *trans*- acting variant is often taken to be one that influences another gene, perhaps a TF, that itself regulates the gene whose expression phenotype is associated with the polymorphism. The definition of an association as *cis*- or *trans*- is often arbitrarily decided by the distance from the associated expression phenotype (Stranger et al define a *cis*- association as anything within 1 megabase of the expression phenotype). It is not uncommon for regulatory elements to be many tens or hundred of kilobases from the genes they modulate, such as in the case of the Shh gene that is regulated by an enhancer element 800 kb away from its TSS (Lettice et al. 2002). Without extra experimental information on the mode of action of the putative functional SNP, such distinctions are difficult to make. Indeed, it is not always clear whether the SNPs found in such studies are causative or just in linkage disequilibrium with the real causative polymorphisms. If a putative regulatory SNP arising from an association is of sufficient interest, further evidence of its functionality can be obtained from a reporter assay, by quantifying transcripts from each allele using a transcribed marker SNP or by measuring RNA pol II loading in a heterozygous individual. This confirmation can be important in the correct interpretation of association results on a gene-by-gene basis. An A/G polymorphism 308 base pairs upstream of the tumour necrosis factor (TNF) promoter has been repeatedly associated with susceptibility to a variety of infections diseases (McGuire et al. 1994; Shaw et al. 2001) but reporter assays have been unable to definitively confirm that the SNP impacts on promoter strength. Examination of Pol II loading using the haploChIP method showed that *in*

vivo it had no effect (Knight et al. 2003). Similar experiments on alleles in linkage disequilibrium with the TNF -308 SNP revealed differential pol II loading on another G/A SNP in the promoter of the LTA gene. This SNP was itself a marker for a haplotype of several polymorphisms in the LTA promoter (Knight et al. 2003). Investigation of the basis for the original association with a TNF SNP thus successfully redirected attention on a more likely candidate gene.

A crucial difference between *cis*- and *trans*- regulation is that *cis*-regulatory variants will influence only the copy of the gene on the same chromosome, whereas *trans*-acting variation will influence both copies. This would give rise to allele-specific expression, where expression from one member of an allelic pair has significantly higher expression than the other. This means that, given a method for differentiating between transcripts from each allele, the presence of *cis*- regulatory variations can be detected without having candidate SNPs to start with. The archetypal instance of allele-specific expression is imprinting, where one chromosomal copy is completely silenced, and expression of the gene is thus monoallelic. The major mechanism for imprinting involves the methylation of imprinting control regions, which in turn silence the expression of a number of imprinted genes in a cluster (Reik and Walter 2001; Strathdee, Sim, and Brown 2004). There are currently 48 known imprinted genes in human and 79 in mouse (Morison, Ramsay, and Spencer 2005), although it is thought that there may be a significant number still undiscovered. Several recent papers using SNP microarrays or RT-PCR have shown that allele-specific expression is common in the human genome outside of imprinted genes (Yan et al. 2002b; Bray et al. 2003; Lo et al. 2003; Pastinen et al. 2004). Hudson ***et. al.*** have surveyed dbEST and identified ESTs containing polymorphisms whose allele frequencies are known. Deviations in the proportions of ESTs for each allele in dbEST relative to their known allele frequencies are indicative of differential expression. Nearly 1000 genes were found with an allelic imbalance in EST representation (Ge et al. 2005). All this evidence has led to the well-accepted view that *cis*-regulatory variation is plentiful in the human genome, although the mechanistic basis for it remains poorly understood. Currently, experimental surveys of allele-specific expression have not generally been followed up with *in vitro* studies of particular variants, so whether they are due to promoter variation or variation in other elements remains to be determined.

1.6 Promoter polymorphisms in disease and evolution

The majority of known monogenic diseases involve mutations that affect the coding sequence of a gene, and hence severely impair its function *in vivo* (McKusick 1998). These diseases are generally rare, with the illnesses segregating in family pedigrees with clear mendelian inheritance patterns. Mutations such as these can explain only a tiny proportion of the genetic component of human disease, with the majority thought to be accounted for by the concerted influence of many loci with more modest effects. As the available resource of human SNPs continues to grow at a rapid pace, and the cost of genotyping assays falls, association studies involving large numbers of individuals are becoming more and more feasible. There is now a significant number of putative promoter SNPs associated with disease phenotypes including schizophrenia (Saito et al. 2001; Wonodi et al. 2005), asthma (Nakashima et al. 2006), bipolar disorder (Barrett et al. 2003) as well as many cancers (Elander, Soderkvist, and Fransen 2006; Park et al. 2006; Snoussi et al. 2006). Even diseases with very large environmental components, such as HIV, have shown these associations (Shin et al. 2000). In many cases, further experimental data have indicated an *in vitro* or *in vivo* effect on gene expression. Some detailed examples are reviewed by Knight (Knight 2005), and other examples include hypertension (Kumar et al. 2005; Li et al. 2006), ά-thalassemia (De Gobbi et al. 2006), coronary heart disease (Spiecker et al. 2004), systemic lupus erythematosus (Gibson et al. 2001) and osteoporosis (Garcia-Giralt et al. 2002; Garcia-Giralt et al. 2005). Changes in gene expression levels in general have been linked to disease phenotypes, particularly in cancer where they have been better-studied (Ross et al. 2000). It is also increasingly recognised that such changes can be caused not only by DNA sequence polymorphisms or non-synonymous mutations in TF genes but by epigenetic dysregulation (Baylin 2005). While there can be extensive transcription profile change between tumour tissue and normal tissue, aberrant methylation at key cancer-associated genes can cause expression level changes that then increase the risk of tumour formation (Yan et al. 2002a; Deng et al. 2004). These can consist of either one or both of hypermethylation of tumour suppressor genes (Herman et al. 1994) and hypomethylation of oncogenes (Feinberg and Vogelstein 1983), as well as global methylation changes that have more extensive effects such as re-activating latent retrotransposons that could then become mutagenic (Alves, Tatro, and Fanning 1996; Lin et al. 2001).

It is has long been proposed that evolution in regulatory sequence may account for a significant proportion of phenotypic evolution (King and Wilson 1975), but it is only with the advent of multiple genome sequences that this can be explored on a significant scale. Significant turnover in functional TFBSs between species has already been demonstrated, suggesting that the generation of new binding sites or the loss of old ones is not an unlikely event (Dermitzakis and Clark 2002). Regulatory sequence variation has been shown to have phenotypic consequences in multiple eukaryotic organisms from *Saccharomyces cerevisiae* (Fay et al. 2004) and *Drosophila melanogaster* (Rifkin, Kim, and White 2003) to primates (Enard et al. 2002) and humans (Pastinen and Hudson 2004; Knight 2005). This abundance of heritable *in vivo* expression differences is important from an evolutionary standpoint because functional regulatory polymorphisms with real physiological or morphological phenotypes will be visible to natural selection. This is especially likely when regulatory variants affect the expression of TFs with many downstream targets, with developmentally important regulators such as Hox genes being a good example (Carroll 2000). Evidence of regulatory variation leading to morphological change is available from model organisms. Mutations in an enhancer controlling the Hoxc8 gene between chicken and mouse have been shown to affect its spatial expression pattern, and hence the difference in thorax development between these two species (Belting, Shashikant, and Ruddle 1998). There are also known instances of balancing selection conserving the function of a regulatory element despite changes in sequence. A good example is the stripe 2 element (S2E) in *Drosophila* species, which regulates the *even-skipped* gene. The S2E sequence has diverged significantly between *Drosophila melanogaster* and *Drosophila pseudoobscura*, including gains and losses of several predicted binding sites for TFs. Despite this, both elements drive expression of a reporter in exactly the same way in Drosophila embryos (Ludwig et al. 2000). However, if chimeric enhancers are constructed containing 5' and 3' halves from each species, the pattern of reporter expression is disrupted (Ludwig et al. 2000). This indicates that the functional consequences of mutations in the S2E have been dampened by compensatory mutations in the same element.

Evidence of natural selection on promoter alleles has been detected in wild populations of teleost fish (Crawford, Segal, and Barnett 1999; Segal, Barnett, and

Crawford 1999) and Drosophila (Daborn et al. 2002; Lerman et al. 2003), as well as artificial selection on natural sequence variation during the domestication of maize (Wang et al. 1999). This demonstrates that selection can act on the raw material provided by *cis*-regulatory variation. Evidence from studies of *Drosophila melanogaster* and *Drosophila simulans* as well as hybrids of the two species suggests that the majority of lineage-specific gene expression differences can be explained by *cis*-regulatory variation rather than *trans* (Wittkopp, Haerum, and Clark 2004). In humans, the best evidence of selection on promoter variation is in genes involved in susceptibility to infection (Tournamille et al. 1995; Hamblin and Di Rienzo 2000; Bamshad et al. 2002). This is perhaps not surprising as infections have been a major selective force in human evolution, and remain one of the strongest agents of selection in the developing world.

1.7 Aims of this thesis

Despite the significant recent advances in discovering regulatory variation in the human genome, the mechanistic basis of much of this variation remains something of a black box. The complexity of eukaryotic transcriptional networks, the structural malleability of regulatory elements compared with coding regions and the context dependence of sequence variant function means that there is still no reliable way to predict what the effect is of introducing a quantitative change in the regulatory framework of the cell. The comprehensive testing of every possible regulatory permutation in the lab is still far from being technically or economically feasible. The most productive way to tackle this problem is to build *in silico* models based on representative experimental datasets.

Promoters have been a natural target for research into *cis*-regulation. Their importance in integrating regulatory signals to a single gene gives them a crucial role in gene expression. They are also easier to identify than enhancers or other distant elements, being generally restricted to the 5' ends of genes. There are several strategies available for studying the effect of promoter variation, and each has its own advantages and disadvantages. The closer the experiment is to studying expression variation in an *in vivo* system, the more physiologically relevant the data becomes. However, it also means that more factors come into play such as the epigenetic state

of the promoter, the chromatin environment and the presence of inducing factors, which may be either unknown or prohibitively increase the complexity of the experiments if elucidated. With experimental designs that remove these extra factors, such as *in vitro* transcription experiments or mobility shift assays, the link between the results and the genotypes will be much clearer, but the presence of any effects found *in vivo* is not confirmed.

A large number of promoter polymorphisms are known that can affect the rate of initiation in an *in vitro* reporter assay (Rockman and Wray 2002; Buckland et al. 2005). However, the majority have been studied because of a clinical interest in the downstream gene (Rockman and Wray 2002). Because the experiments were done in many different labs under widely varying experimental conditions and vector designs, they are not suitable as a stand-alone dataset for the analysis of promoter variation in general. There is also a bias towards promoters linked to diseases, and they may not be representative of promoter variation in the genome as a whole. Buckland and colleagues have published a series of papers containing reporter assay screens of promoter variation, using candidate promoters from a variety of sample sources (Hoogendoorn et al. 2003; Buckland et al. 2004a; Buckland et al. 2004b; Buckland et al. 2005). While some of these are also selected based on the types of genes they regulate, others are simply selected by chromosome. Together these are currently the largest coherent set of tested promoter polymorphisms.

The main aims of this thesis were threefold:

1. To build up a set of robustly-tested functional polymorphisms in human promoters
2. To use this set to assess the ability of current models of regulatory elements to predict functional promoter variation
3. To try and learn more about how *in vitro* promoter assays relate to *in vivo* gene expression

In this thesis, I explored the effect of promoter sequence variation on the efficiency of the promoter, as measured by luciferase reporter assays. Chromosome 22 was chosen as a model system for the genome as a whole, and there was no selection for genes apart from their absence from gene families (this was for practical reasons). The first

phase of the work involved the generation of a resource of promoter polymorphisms to be subsequently tested. This was done by resequencing all unique promoters on chromosome 22 from a panel of unrelated individuals. The resulting set of SNPs was analysed for haplotypes, and these were then cloned using a novel high-throughput strategy into luciferase reporter plasmids. Four transformed cell lines, HT1080, TE671, HEK293FT and HeLa were chosen as the *in vitro* model system for transient transfection of the cloned variant promoters. These experiments revealed a new set of promoter SNPs with functional consequences in these cell lines. The resulting collection of SNPs with assigned functional consequences was used to assess the ability of a variety of putative regulatory elements to retrospectively predict SNP functionality by looking for enrichment of functional SNPs in these elements. Whole genome expression microarrays were used to assess the TF expression profiles of these cells, enabling the analysis of the luciferase data with knowledge of the TFs present in each cell line. Tests were done to see if the action of functional SNPs could be accounted for by differential expression of TFs across cell lines. The concordance of promoter activity and endogenous gene expression in the same cell lines was also assessed in order to quantify how much of gene regulation takes place at the promoter itself versus upstream elements and epigenetic modifications. Finally an attempt was made to generate new motifs using the promoters of genes co-regulated across the four cell lines, in order to see how their performance would compare to motifs already known.