Studies of the effects of promoter sequence variation on gene expression in human chromosome 22

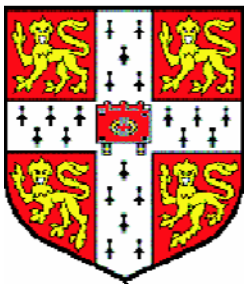
Jamil Bacha

Wolfson College

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

wellcome trust sanger institute



Declaration

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

This dissertation does not exceed the page limit specified by the Biology Degree Committee.

Acknowledgements

This thesis would not have been possible without the support, both scientific and personal, of a great many people. First and foremost, my deepest personal thanks go to my supervisor Dr. Ian Dunham for his unstinting support and mentorship throughout my time in the group. I would also like to thank Dr. John Collins and Dave Beare for their accumulated lab and computer wisdom, and all the members of Team 62 for making this a fun and unforgettable experience. Special thanks also go to Dr. Nick Luscombe and (nearly-Dr.) Juanma Vaquerizas at the EMBL-European Bioinformatics Institute for their invaluable intellectual contribution to the array analysis work, and particularly to Juanma for carrying out the quality control and linear modelling analyses on the array data.

Thank you to my thesis committee Dr. Alex Bateman and Dr. Dave Vetrie (Sanger Institute) and Dr. James Ajioka (University of Cambridge) for their intellectual input and guidance. A big thank you to Dr. Steven Leonard and Dr. Sarah Hunt for helping me mercilessly flog the SNP database until it did what it was told (eventually!), and to Nick Matthews, Jonathan "Bill" Bailey and the Sanger Institute Sequencing Centre for their hard work on the promoter and clone re-sequencing. Thank you to Dr. Barbara Stranger for running the association analysis of promoter SNPs and for our fun talks on the Sanger bus, Dr. Robert Andrews and Dr. Gregory Lefebvre for contributing the clustering of co-expressed genes, and Dr. Thomas Down for his invaluable assistance with the motif generation and analysis. Thank you to Andy "Wilb" Dunham and Andy Bentley of the ExoSeq group for teaching me how to tame the wild lab robot! Thank you to Dr. Manolis Dermitzakis, Dr. Ewan Birney, Dr. Thomas Down and Dr. Vardhman Rakyan for their very interesting scientific discussions and to my fellow graduate students for their not-so-scientific ones. Good luck guys!

My love and gratitude go to my family, without whom I would not have had the opportunity to embark on this journey and fulfil my childhood ambitions of a scientific career. Finally, special thanks to Dr. Davina Stevenson for her love and companionship through the highs and the lows … something I will always treasure.

Abstract

The molecular and physiological phenotype of a gene depends not only on the structure and properties of the protein it codes for, but on the regulation of the magnitude and timing of expression of that protein in the cell. The role of the promoter in gene regulation can be seen as an integrator of the numerous intra- and extra-cellular signals that influence the levels of transcription factors in the nucleus, with the output being the level of transcriptional initiation. The identification of transcription factor binding sites and promoter polymorphisms with real functional consequences continues to elude purely computational methods, and more experimental data is needed before this state of affairs is changed. In this project, I have re-sequenced the majority of promoters on human chromosome 22 from a panel of 48 unrelated individuals, generating a set of 807 promoter SNPs with associated genotype information. I then developed a novel high-throughput cloning strategy utilizing Gateway technology to produce a library of cloned promoter fragments, and applied this to generate a set of 293 promoter haplotypes from 84 different promoters. The functional significance of the promoter differences was assayed by luciferase reporter assays in HT1080, TE671, HEK293FT and HeLa cell lines. This revealed significant levels of sequence-dependent variation in promoter efficiency, with at least 22% of promoter SNPs having functional consequences. The performance of currently-known putative regulatory elements in retrospectively predicting functional variation was assessed, and found to be wanting. An expansion of upregulatory promoter mutations was noted in the population used, which has implications for the understanding of gene regulatory evolution. Analysis of the whole genome expression profiles of the four cell lines confirmed a qualitative correlation between promoter activity and *in vivo* gene expression, but also indicated that the presence of a known transcription factor binding site could often be ruled out as the mechanism for a functional promoter polymorphism. This study is the most detailed analysis to date of high throughput promoter assays, and is suitable for scaling up to genome-scale functional SNP discovery.

Contents

Abbreviations and Symbols

Abbreviations

ANN	Artificial neural network
ANOVA	Analysis of variance
BRE	$TF_{II}B$ recognition element
CAGE	Cap analysis of gene expression
CAT	Chloramphenicol acetyltransferase
CEPH	Centre d'Etude du Polymorphisme Humain
ChIP	Chromatin immunoprecipitation
DNA	Deoxyribose nucleic acid
DPE	Downstream promoter element
EMSA	Electrophoretic mobility shift assay
ENCODE	ENCyclopaedia Of DNA Elements
EST	Expressed sequence tag
GO	Gene ontology
Indel	Insertion/deletion polymorphism
LCR	Locus control region
LD	Linkage disequilibrium
MTE	Motif ten element
PCR	Polymerase chain reaction
PIC	Pre-initiation complex
Pol II	RNA polymerase II
RLU	Relative light units
RNA	Ribose nucleic acid
mRNA	Messenger RNA
RT-PCR	Reverse transcriptase PCR
SAGE	Serial analysis of gene expression
SELEX	Systematic Evolution of Ligands by EXponential enrichment
SNP	Single nucleotide polymorphism
TAF	TATA-associated factor

TBP	TATA-binding protein
TF	Transcription factor
TFBS	Transcription factor binding site
$TF_{II}D$	TBP-associated factor II D
TSS	Transcription start site
Tukey's HSD	Tukey's Honestly Significantly Different test
UTR	Un-translated region
VeGA	Vertebrate Gene Annotation

IUPAC Symbols for base positions

IUPAC Code	Meaning
A	A
C	C
G	G
T/U	T
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
N	G or A or T or C

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.

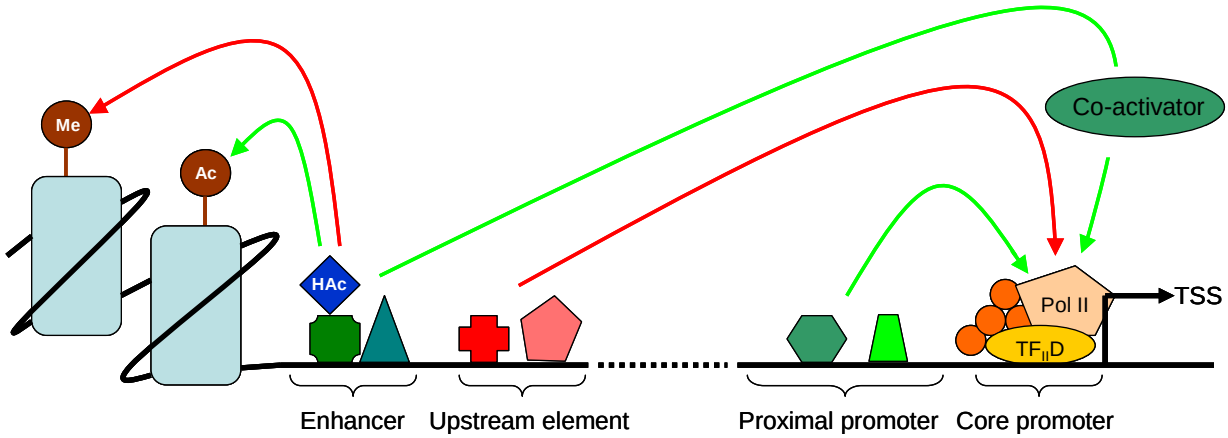


Figure 1. Diagrammatic representation of regulatory inputs into promoter function. The core promoter is the binding site for the basal transcription machinery, including RNA Pol II and $TF_{II}D$ and other general TFs. Both in the proximal promoter and in upstream enhancers and silencers are binding 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).

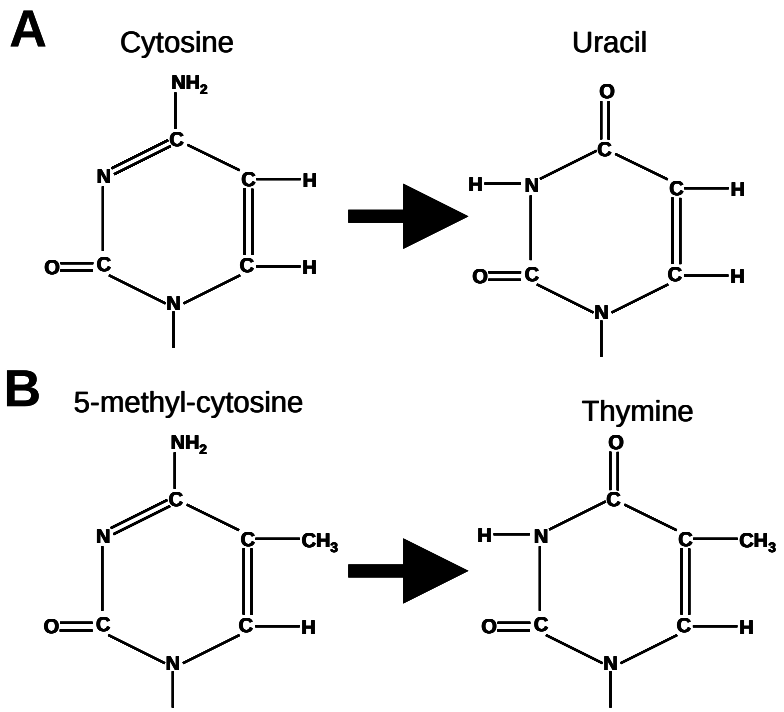


Figure 2. Deamination of cytosine and methylcytosine produce different bases. A) Cytosine bases are prone to spontaneous deamination, producing uracil as the resulting base. This is efficiently detected and repaired by the cellular repair machinery. B) Methyl-cytosine bases are prone to the same process, 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).

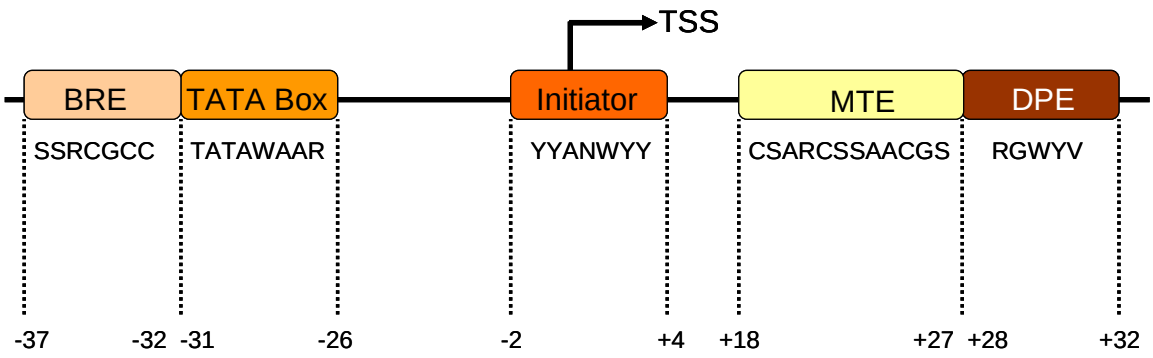


Figure 3. Known core promoter motifs in human promoters. The positions of the motifs are shown relative to the transcription start site (TSS), designated as base +1. Each of the elements is described in detail below. The consensus sequences are also shown in IUPAC ambiguity code notation. Figure adapted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

2 Materials and Methods

2.1 Common buffer formulae

Phosphate Buffered Saline (PBS)

36.65 g	Sodium chloride
11.80 g	Disodium hydrogen phosphate (Na_2HPO_4)
6.60 g	Sodium dihydrogen phosphate (NaH_2PO_4)
up to 5 l	Double-distilled water

10x Tris-Buffered EDTA (TBE)

109.0 g	Tris
55.6 g	Boric Acid
40 ml	0.5 M EDTA
up to 1 l	Double-distilled water
	… pH to 8.3

LB Broth

10 g	Tryptone
5 g	Yeast extract
10 g	Sodium chloride
up to 1 l	Double-distilled water
	… pH to 7.0

2x LB Broth

20 g	Tryptone
10 g	Yeast extract
10 g	Sodium chloride
up to 1 l	Double-distilled water
	… pH to 7.0

LB Agar

10 g	Tryptone
5 g	Yeast extract
5 g	Sodium chloride
up to 1 l	Double-distilled water
	… pH to 7.5
20 g	Agar

ExoSAP Buffer

20 ml	1 M Tris pH 8.0
10 ml	1 M Magnesium Chloride
70 ml	Double-distilled water

2.2 Cell Culture Protocols & Media

2.2.1 Media for HeLa and HT1080 cell lines

500 ml Modified Eagle's Medium (Sigma, #M2279)

10% FBS (Gibco #10270-106)

2 mM L-Glutamine

100 units ml^{-1} Penicillin

100 μg ml^{-1} Streptomycin

5% Non-essential amino acids (Gibco, #11140-035)

2.2.2 Media for TE671 and HEK293FT cell lines

500 ml Dulbecco's Modified Eagle's Medium (Sigma, #D5796)

10% FBS (Gibco #10270-106)

2 mM L-Glutamine

100 units ml^{-1} Penicillin

100 μg ml^{-1} Streptomycin

2.2.3 Passaging Cells

All cells were grown in a Galaxy R incubator (Scientific Laboratory Supplies) at 37°C, 5% CO_2 in 75 cm^2 culture flasks with 0.2 μm vent caps (Corning, #430641). They were passaged when between 80% and 95% confluent. HeLa, HT1080 and HEK293FT cells were split at 1:10, and TE671 cells at 1:6.

1. The media from the culture flask was decanted into 1% Virkon (Antec International, #330003).
2. The cells were washed twice with 10 ml of 1x PBS, and the wash decanted into 1% Virkon.
3. The cells were washed twice with 3 ml of 1x Trypsin-EDTA (Gibco, #25300-054), and the wash decanted into 1% Virkon.
4. The flask was incubated at 37°C, 5% CO_2 for 5 mins.
5. During the incubation, 15 ml of the appropriate cell culture medium was added to a new 75 cm^2 flask.
6. The cells are dislodged by sharply tapping the flask 2-5 times, and the cells suspended in 10 ml of the appropriate cell culture medium.

7. The appropriate volume of cell suspension was added to the new flask according to the recommended split ratios.

2.3 Chapter 3 Protocols

DNA samples of the 48 CEPH grandparents were a gift from Andrew Dunham, Wellcome Trust Sanger Institute, and were originally purchased from Coriell Cell Repositories. Samples used were:

Repository #	Cell line #	Age	Gender	Family #	Relation to proband
NA06985	GM06985	69	F	1341	Mat
NA06993	GM06993	74	M	1341	Mat
NA06994	GM06994	68	M	1340	Pat
NA07000	GM07000	66	F	1340	Pat
NA07002	GM07002	63	F	1333	Pat
NA07007	GM07007	95	M	1331	Pat
NA07016	GM07016	71	M	1331	Mat
NA07017	GM07017	61	M	1333	Mat
NA07022	GM07022	63	M	1340	Mat
NA07034	GM07034	71	M	1341	Pat
NA07049	GM07049	68	M	1333	Pat
NA07050	GM07050	62	F	1331	Mat
NA07055	GM07055	70	F	1341	Pat
NA07056	GM07056	65	F	1340	Mat
NA07340	GM07340	83	F	1331	Pat
NA07341	GM07341	61	F	1333	Mat
NA07345	GM07345	69	F	1345	Mat
NA11879	GM11879	66	M	1347	Pat
NA11880	GM11880	65	F	1347	Pat
NA11881	GM11881	62	M	1347	Mat
NA11882	GM11882	61	F	1347	Mat
NA11917	GM11917	66	M	1423	Pat
NA11918	GM11918	64	F	1423	Pat
NA11919	GM11919	67	M	1423	Mat
NA11920	GM11920	66	F	1423	Mat
NA11992	GM11992	86	M	1362	Pat
NA11993	GM11993	80	F	1362	Pat
NA11994	GM11994	80	M	1362	Mat
NA11995	GM11995	84	F	1362	Mat
NA12003	GM12003	97	M	1420	Pat
NA12004	GM12004	92	F	1420	Pat
NA12005	GM12005	77	M	1420	Mat
NA12006	GM12006	75	F	1420	Mat
NA12043	GM12043	74	M	1346	Pat

NA12044	GM12044	70	F	1346	Pat
NA12045	GM12045	74	M	1346	Mat
NA12144	GM12144	71	M	1334	Pat
NA12145	GM12145	70	F	1334	Pat
NA12146	GM12146	61	M	1334	Mat
NA12154	GM12154	92	M	1408	Pat
NA12155	GM12155	86	M	1408	Mat
NA12156	GM12156	81	F	1408	Mat
NA12236*	GM12236	86	F	1408	Pat
NA12239	GM12239	61	F	1334	Mat
NA12248	GM12248	89	M	1416	Pat
NA12249	GM12249	77	F	1416	Pat
NA12250	GM12250	66	M	1416	Mat
NA12251	GM12251	63	F	1416	Mat

* DNA sample no longer available from Coriell

2.3.1 Selection of promoters for re-sequencing

An in-house script (written by Dr. David Beare) was used to extract the genomic sequence from -2000 bases to +50 bases relative to the TSSs of all chromosome 22 genes with a confirmed 5' end, according to the latest published annotation (Collins et al. 2003).

NCBI BLAST was then used to map these sequences back against the human genome. The results were analysed manually, and promoter sequences that matched more than one location in the genome were eliminated.

2.3.2 Primer design

Primers were designed using Primer3 (Rozen and Skaletsky 2000). All parameters were used at default settings except for the ones in the table below:

Parameter	**Value used**
Primer optimum size	20
Primer minimum size	16
Primer maximum size	24
PRIMER_MAX_POLY_X	4
PRIMER_SELF_ANY	6.0
PRIMER_SELF_END	2.0
PRIMER_MIN_GC	18
PRIMER_MAX_GC	82.5
PRIMER_MIN_TM	50
PRIMER_MAX_TM	70

2.3.3 Optimisation of genomic PCR

All PCRs were carried out using the Hot Start Taq (Abgene #SP-0034) and associated reagents at their stock concentrations.

Standard protocol

A PCR reaction premix sufficient for the number of reactions to be carried out was prepared according to the following formula:

Reagent	1X / µl
10x Buffer	1.5
1 mM dNTPs	1.5
DMSO	0.75
ddH_2O	9.06
Taq	0.09
10 ng ml^{-1} DNA	0.5
Total premix	13.4
Primer (15 µM)	1.6

The following cycling protocol was used:

- 95°C for 15:00
- 95°C for 0:30, 60°C for 0:30, 72°C for 0:30 – 38 cycles
- 72°C for 10:00

Stepped activation

The same premix formula as the standard protocol was used, but the following cycling protocol was used:

- 95°C for 2:00, 60°C for 0:30, 72°C for 0:30 – 7 cycles
- 95°C for 0:30, 60°C for 0:30, 72°C for 0:30 – 31 cycles
- 72°C for 10:00

65°C Annealing

The same premix formula as the standard protocol was used, but the following cycling protocol was used:

- 95°C for 15:00
- 95°C for 0:30, 65°C for 0:30, 72°C for 0:30 – 38 cycles
- 72°C for 10:00

55°C Annealing

The same premix formula as the standard protocol was used, but the following cycling protocol was used:

- 95°C for 15:00
- 95°C for 0:30, 55°C for 0:30, 72°C for 0:30 – 38 cycles
- 72°C for 10:00

1.1 M Betaine/7% DMSO

A PCR reaction pre-mix sufficient for the number of reactions to be carried out was prepared according to the following formula:

Reagent	1X / µl
10x Buffer	1.5
1 mM dNTPs	1.5
DMSO	1.05
5M Betaine	3.3
ddH_2O	5.46
Taq	0.09
10 ng ml^{-1} DNA	0.5
Total premix	13.4
Primer (15 µM)	1.6

The following cycling protocol was used:

- 95°C for 15:00
- 95°C for 0:30, 60°C for 0:30, 72°C for 0:30 – 38 cycles
- 72°C for 10:00

2.3.4 High-throughput PCR of promoter fragments

1. Oligonucleotide primers were ordered from Illumina, and were supplied at a concentration of 30 µM. Polymerase and associated buffer was Hot Start Taq (Abgene #SP-0034).
2. A TECAN Genesis RSP150 robot was used to aliquot 8 µl of 3 µM primer into a batch of twelve 384-well PCR plates (Eppendorf, # 951020516), such that each plate was divided into 4 identical quadrants, each containing the same 96 primer pairs.
3. A PCR reaction pre-mix sufficient for the number of reactions to be carried out was prepared according to the following formula:

Reagent	**1X / µl**	**6000X / µl**
10x Taq Buffer	1.5	9000
1 mM dNTPs	1.5	9000
DMSO	0.75	4500
ddH_2O	5.86	35160
Hot Start Taq	0.09	540

4. For each 96-well plate of STSs to be sequenced, four 96-well microtitre plates (Greiner, # 650161) were filled with 145.5 µl of premix per well.
5. For each of the 48 DNA samples to be amplified 7.5 µl of 10 ng μl^{-1} DNA was added to each of 8 wells of premix in a column.
6. A TECAN Genesis RSP150 robot was used to aliquot 7 ul of premix/DNA solution into the 384-well PCR plates containing the pre-aliquoted primers, such that all the wells in each quadrant contained the same DNA sample.
7. The PCR plates were centrifuged at 1000 rpm for 1 min on an Eppendorf 5403 centrifuge.

8. PCR was carried out using the following reaction cycle on a thermocycler (MJ, #PTC-225):
    a. 95°C for 15 mins
    b. 95°C for 30 secs, 60/65°C for 30 secs, 72°C for 30 secs –> 38 cycles
    c. 72°C for 10 mins

2.3.5 Cleanup of PCR products

1. A premix of Shrimp Alkaline Phosphatase (USB, #70092X) and Exonuclease I (USB, #70073X) sufficient for the number of reactions to be cleaned was prepared, according to the following formula:
   1 ml ExoSAP buffer
   1 ml ddH_2O
   1 ml Shrimp Alkaline Phosphatase
   0.1 ml Exonuclease I
2. 2 µl was added to each PCR reaction and the plates centrifuged at 1000 rpm for 1 min on an Eppendorf 5403 centrifuge.
3. The PCR plates were incubated for 1 hour at 37°C and for 15 mins at 90°C on a thermocycler (MJ, #PTC-225), and stored at -20°C until sequencing.

2.3.6 Sequencing of PCR products

Cleaned PCR fragments were submitted to the Sanger Institute Sequencing Centre. They were sequenced from both ends using the di-deoxy chain terminator method (Sanger et al. 1977), with V3.1 Bigdye terminator chemistry (West et al. 2005). The resulting sequencing reactions were analyzed on 3730 ABI sequencing machines (Applied Biosystems, USA).

2.4 Chapter 4 Protocols

2.4.1 Creation of pools and design of oligos

The results of the haplotype predictions were analysed by eye, and a set of individuals chosen for each promoter such that the proportions of the different haplotypes present

was as close to equal as possible. 10 ng μl^{-1} solutions of the individual DNA samples were mixed in order to keep the concentration at that level.

Oligos for cloning the promoters into the Gateway vectors were designed by simply taking the sequence of the primers used for SNP-mining and adding the att-sites to the 5' ends.

2.4.2 PCR of promoters from pool templates

1. A PCR reaction premix sufficient for the number of reactions to be carried out was prepared according to the following formula:

Reagent	**1X / µl**
10x KOD Buffer	2
2 mM dNTPs	2
25 mM $MgSO_4$	0.8
DMSO	1
ddH_2O	8
KOD (2.5 U μl^{-1})	0.4

2. 14.2 µl of premix was aliquoted into 96-well PCR plates (ABgene, #AB-0800), and 0.8 µl of each primer mix was added to the reactions.
3. 5 µl of each template was added to the well containing the corresponding primers.
4. PCR was carried out using the following reaction cycle on a thermocycler (MJ, #PTC-225).
    a. 95°C for 4 mins
    b. 95°C for 1 min, 65°C for 30 secs, 72°C for 1 min –> 30 cycles (-0.3°C annealing temperature per cycle)
    c. 72°C for 5 mins
5. First round PCR products were diluted 1:200 in ddH_2O and used as templates for the second round of PCR.
6. A PCR reaction premix sufficient for the number of reactions to be carried out was prepared according to the following formula:

Reagent	1X / µl
10x KOD Buffer	2.5
2 mM dNTPs	2.5
25 mM $MgSO_4$	1
DMSO	1.25
ddH_2O	11.25
15 µM Primer	1
KOD (2.5 U μl^{-1})	0.5

7. 20 µl of premix was aliquoted into 96-well PCR plates.
8. 5 µl of template was added to each well.
9. PCR was carried out using the following reaction cycle on a thermocycler (MJ, #PTC-225).
    a. 95°C for 2 mins
    b. 95°C for 15 secs, 45°C for 30 secs, 68°C for 1 min – 5 cycles
    c. 95°C for 15 secs, 55°C for 30 secs, 68°C for 1 min – 20 cycles
    d. 68°C for 5 mins

2.4.3 Gateway cloning into pDONR223

The protocol included with the BP Clonase II kit (Invitrogen, #11789-020), was followed, except that reactions were all scaled down by half in order to conserve reagent.

1. A custom reaction buffer, called BP3 buffer, was made up and used instead of the included BP buffer. The formula for 5x BP3 buffer is:
   100 mM Tris-Cl, pH 7.5
   20 mM EDTA
   30 mM spermidine
   25% glycerol
   225 mM NaCl
2. A reaction premix sufficient for the number of reactions to be carried out was prepared according to the following formula:

Reagent	Volume (1X)
5X BP3 Buffer	2 µl
pDONR223	100 ng (minimum 1 µl)
ddH_2O	Up to 2 µl

3. 5 µl of premix per reaction was aliquoted into the wells of a 96-well PCR plate (ABgene, #AB-0800).
4. 4 µl of cleaned-up PCR insert was added to the premix, and mixed by pipetting.
5. 1 µl of BP clonase II was added to each reaction and mixed by pipetting
6. The reactions were incubated at 16^{o}C overnight on a thermocycler (MJ, #PTC-225).
7. 1 µl of proteinase K (included in the kit) was added to each reaction.
8. The reactions were incubated at 37^{o}C for 10 mins.

2.4.4 Transformation and preparation of pDONR223 haplotype libraries

1. 0.5 µl of each BP reaction was aliquoted into a 96-well PCR plate (Costar, #6511).
2. The plate was pre-chilled to -20^{o}C, and then placed in a metal heating block inside a benchtop cooler (StrataCooler) to equilibrate to 4^{o}C for 5 mins.
3. 10 µl of library-efficient DH5α cells (Invitrogen, #18263-012) were added to the plasmid and the plate incubated at 4^{o}C for 30 mins.
4. The cells were heat-shocked at 42^{o}C for 45 secs using a thermocycler (MJ, PTC-225).
5. The plate was placed back in the 4^{o}C heating block for 2 mins.
6. 90 µl of SOC media (Invitrogen, #15544-034) was added to each transformation.
7. The plates were incubated at 37^{o}C for 1 hour.
8. The transformations were plated on to Hybond-N+ nylon membranes (Amersham, #AMNK9655) laid on LB agar plates containing 100 ng ml^{-1} of spectinomycin (Sigma, #S-4014), and the plates incubated at 37^{o}C overnight.
9. Colonies were scraped into 10 ml of LB broth with a plastic spreader, and the cells pelleted using a Beckman centrifuge (J6-M6) centrifuge at 3000 rpm for 15 mins.
10. Plasmids were prepared from the cell pellets using the Qiaquick Spin Miniprep kit (Qiagen, #27104) as per manufacturer's instructions.

2.4.5 Gateway cloning into pGL3 Basic GW

The protocol included with the LR Clonase II kit (Invitrogen, #11791-020), was followed, except that reactions were all scaled down by half in order to conserve reagent.

1. A custom reaction buffer, called LR4 buffer, was made up and used instead of the included LR buffer. The formula for 5X LR4 buffer is:
   200 mM Tris-Cl, pH 7.5
   10 mM EDTA
   35 mM spermidine-HCl
   320 mM NaCl
   25% glycerol
2. A reaction premix sufficient for the number of reactions to be carried out was prepared according to the following formula:

Reagent	**Volume (1X)**
5X LR Buffer	2 µl
pGL3 Basic GW+	100 ng (minimum 1 µl)
TE	Up to 4 µl

3. 7 µl of premix per reaction was aliquoted into a 96-well PCR plate ABgene, #AB-0800.
4. 2 µl of prepared pDONR223 containing the inserts to be cloned was added to each reaction and mixed by pipetting.
5. 1 µl of LR clonase II was added to each reaction and mixed by pipetting.
6. The reactions were incubated at 16^{o}C overnight on a thermocycler (MJ, #PTC-225).
7. 1 µl of proteinase K (included in the kit) was added to each reaction.
8. The reactions were incubated at 37^{o}C for 10 mins.

2.4.6 Colony PCR of clones from pGL3 Basic GW haplotype libraries

PCR was carried out using the KOD Hot-start DNA polymerase kit (Novagen, #71086) and associated reagents

1. Colonies were picked from agar plates into 1 ml of LB broth containing 100 ng ml^{-1} ampicillin in a deep 96-well plate.

2. Cultures were incubated overnight in an Innova 4000 shaker incubator (New Brunswick Scientific) at 37°C, 275 rpm.
3. PCR templates were prepared by pipetting 50 µl of the overnight cultures into a pipette tip, expelling it back into the culture, and pipetting 50 µl ddH_2O several times using the same tip.
4. 100 µl of the cultures were mixed with 20 µl 50% glycerol and stored at -70°C to produce long term stocks.
5. A PCR reaction premix sufficient for the number of reactions to be carried out was prepared according to the following formula:

Reagent	**1X / µl**
10x KOD Buffer	1.5
2 mM dNTPs	1.5
25 mM $MgSO_4$	0.6
DMSO	0.75
ddH_2O	4.75
15 µM Primers	0.6
KOD (2.5 U μl^{-1})	0.3

   The primers used were RVPrimer 3 (CTAGCAAAATAGGCTGTCCC) and GLPrimer2 (CTTTATGTTTTTGGCGTCTTCCA), and were pre-designed by Promega to amplify across the multi-cloning site
6. 15 µl was aliquoted into the wells of a 96-well PCR plate (ABgene, #AB-0800).
7. 5 µl of each template was added to the reactions and mixed thoroughly by pipetting.
8. PCR was carried out using the following reaction cycle on a thermocycler (MJ, #PTC-225):
   a. 94°C for 2 mins
   b. 94°C for 30 secs, 60°C for 30 secs, 68°C for 1 min –> 25 cycles
   c. 68°C for 5 mins

2.4.7 Sequencing of colony PCR products

Cleaned fragments were sequenced from both ends using the di-deoxy chain terminator method (Sanger et al. 1977), with V3.1 Bigdye terminator chemistry (West

et al. 2005). The resulting sequencing reactions were analyzed on 3700 ABI sequencing machines (Applied Biosystems, USA).

2.4.8 Preparation of plasmids for high-throughput transfection

The Millipore Montage96 Plasmid prep kit (Millipore, #LSKP096) was used to prepare reporter plasmid for transfection. A modified version of the protocol was used as follows:

1. Ice scrapings from glycerol stocks of each plasmid were inoculated into 1 ml starter cultures of 2x LB broth in deep 96-well plates (Costar, #3961). Cultures were incubated for 6-8 hours at 37^{o}C, 275 rpm in a shaker incubator.
2. 20 µl of starter culture were transferred to fresh 1 ml cultures in a new plate, and incubated overnight at 37^{o}C, 275 rpm in a shaker incubator.
3. Cultures were centrifuged at 2600 rpm in a Sorvall RT7 centrifuge (RTH-250 rotor), and the supernatant decanted away.
4. Pellets were re-suspended in 130 µl of solution 1 (Millipore kit) using a pipette to ensure re-suspension.
5. 130 µl of lysis buffer (Millipore kit) was added to each well, and the plates shaken gently on a Stovall belly dancer for 1 min, and incubated at room temperature until 5 mins after addition of the lysis buffer.
6. 130 µl of neutralisation buffer (Millipore kit) was added to each well, and the plates shaken gently on a Stovall belly dancer for 2 mins.
7. Cell lysates were centrifuged at 2600 rpm for 15 mins.
8. The supernatants were transferred to a new plate and re-centrifuged at 2600 rpm for 15 mins.
9. The supernatants were transferred to a Multiscreen$_{96}$ lysis clearance plate (Millipore kit, #MANANLY), and the lysates filtered into a Multiscreen$_{96}$ plasmid plate (Millipore kit, #MANUPSD) using an eppendorf plate vacuum manifold at 0.27 bar (8 in Hg).
10. Lysates were filtered through the plasmid plates at 0.81 bar (24 inHg), with the filtrate directed to waste.
11. The wells were washed 5 times by adding 100 µl of HPLC-grade water and filtering at 0.81 bar, with the filtrate discarded each time.

12. 35 µl of Tris-HCl pH 8.0 was added to each well, and the purified plasmid re-suspended by shaking vigorously on a Sorvall belly dancer for 30 mins.

2.4.9 Co-transfection of cell lines with reporter plasmids

Just prior to setting up the transfection reactions, the cells to be transfected were split according to the protocol detailed above. All transfection experiments were done using cells at passages 3-6. The Effectene transfection reagent (Qiagen, #301427) was used to transfect the cells, and all reagents came from the kit unless otherwise stated.

1. The cells in the cell suspension were counted using a Neubauer 2.5 μm^2 haemocytometer.
2. The concentration of the suspension was adjusted with appropriate growth medium to 6.67 x 10^4 cells ml^{-1}, and 1 x 10^4 cells (150 µl) were seeded into the wells of 96-well cell culture plates (Falcon, #3072). Sufficient wells were seeded to carry out each transfection in quadruplicate, with 4 wells per plate to be used for negative control transfections.
3. 354 ng of each plasmid to be transfected was aliquoted into 96-well PCR plates, with the top wells of every 3rd column containing the same mass of pGL3 Basic (Promega, #E1741).
4. 71 ng of pRL-CMV (Promega, #E2261) was added to each well, and the volumes made up to a total of 47.5 µl with EC Buffer.
5. A dilution of 3.4 µl in 80 µl of Enhancer reagent was made in EC buffer, and 80 µl of this solution was added to each transfection and mixed by pipetting 6 times. The reactions were incubated at room temperature for 5 mins.
6. During the 5 min incubation, a 1:40 dilution of Effectene reagent in EC buffer was prepared. 85 µl of this solution was added to each transfection and mixed by pipetting 6 times, and the reactions incubated at room temperature for 10 mins.
7. 50 µl of each transfection reagent was pipetted into 4 wells pre-seeded with cells, and mixed by pipetting once.
8. The plates were incubated at 37^{o}C, 5% CO_2 for 48 hours.

2.4.10 Assay of firefly and renilla luciferase levels

All assays were carried out using the Dual-Luciferase Reporter Assay Kit (Promega, #E1960), and all reagents are from this kit unless otherwise stated.

1. The media from the transfections was aspirated into 1% Virkon.
2. The cells were washed once in 100 µl 1x PBS, with the wash solution aspirated into 1% Virkon.
3. 23 µl of 1x passive lysis buffer was added to each well, and the plates shaken vigorously for 30 mins on a Sorvall belly dancer.
4. 20 µl of the cell lysates were transferred to 96-well Optiplate luminometer plates (PerkinElmer, #P12-106-001).
5. The levels of firefly and renilla luciferase were assayed using a Berthold LB96V luminometer equipped with dual injectors, one for each of the two luciferase substrates. The injectors were programmed to dispense 30 µl of luciferase assay reagent II (LAR II) and 30 µl Stop & Glo reagent, with each injection followed by a 1.6 sec delay and a 10 sec measurement time.

2.5 Chapter 5 Protocols

2.5.1 Preparation of total RNA from cell lines

RNA samples were prepared from HT1080, TE671, HEK293FT and HeLa cells using the RNeasy mini kit (Qiagen #74104) according to manufacturer instructions.

2.5.2 Sample preparation and hybridisation on whole genome expression arrays

RNA samples for hybridisation were prepared according to the Affymetrix GeneChip Expression Analysis Manual (Affymetrix, California). 5 μg of total RNA per replicate per cell line was prepared and hybridised on Affymetrix U133 Plus 2.0 arrays as per manufacturers instructions. Hybridisations were carried out in an Affymetrix GeneChip oven overnight at 42^{o}C, 60 rpm. Array washes were done on an Affymetrix GeneChip Fluidics 450 Workstation, and the arrays scanned with an Affymetrix GeneChip Scanner. All protocols are outlined in detail in the Affymetrix GeneChip manual, and were followed without deviation. Data analysis was carried out using the Bioconductor package (Gentleman et al. 2004) in collaboration with Juanma Vaquerizas and the European Bioinformatics Institute (see section 5.2). Present, marginal and absent calls were made using the PANP algorithm (Warren et al. 2006).

3 SNP-mining of chromosome 22 promoters by re-sequencing

3.1 Introduction

In order to investigate the functional effect of promoter sequence polymorphism, the first step was to develop a resource of promoter SNPs that could then be used in functional experiments. At the time this project began, the HapMap project was still 2 years from completion (Consortium 2005b). Despite this, the human genome was already covered by a large number of SNPs discovered by many different studies using a variety of techniques. Many of these will have been located in promoters. However, simply knowing that SNPs exist in certain genomic positions in a population does not constitute a useful experimental tool unless the original samples can be obtained or the SNPs can be re-created by mutagenesis. What is required is a set of DNA samples of known genotypes that can be cloned as required.

There are two ways to establish a resource of promoter SNPs for subsequent experiments. The first is to use dbSNP to find promoters with known polymorphisms, and then to genotype them in a set of DNA samples and/or cell lines from different individuals. At the start of the project, dbSNP (then on build 119) contained 7,231,721 SNPs, on average one every 475 bases of the genome. However, because dbSNP holds the combined output of a wide range of SNP discovery studies using varying methods, populations and target regions, the distribution of the SNPs is not even across the genome. Using dbSNP as the sole source of polymorphisms for an experimental study means that not all SNPs will be detected, giving a misleading picture of variation in the tested region. In addition, it may also be necessary to design specific assays for every SNP depending on the genotyping platform to be used. The second method, and the one chosen for this project, is to re-sequence defined promoters from a panel of multiple individuals. This has the advantage of genotyping both known and novel SNPs. It also confirms the true sequence context of the polymorphisms (i.e. the consensus sequence of a promoter in a population of individuals may itself be different from the human genome reference). The extensive support infrastructure, large sequencing capacity and established high-throughput pipelines available at the Sanger Institute also make this method particularly feasible.

This project aimed to re-sequence all promoters on human chromosome 22. This chromosome was chosen as a model system for the rest of the genome because of its

high quality of annotation, high gene density and relatively small size. These factors have historically resulted in chromosome 22 being chosen to pilot large scale sequencing and functional studies (Dunham et al. 1999; Collins et al. 2004). A potential disadvantage to using a chromosome to functionally represent the genome is the possibility of a bias in functional classes of genes. A comparison of the Gene Ontology classes for genes on chromosome 22 against five random lists of 1000 genes showed no evidence of this (data not shown). Promoters that are duplicated or in low copy repeats were excluded from the project. This is because it would be very difficult to specifically PCR one copy instead of another. Bases where the copies have diverged from each other would appear as universally heterozygous SNPs, and any real SNPs found would not be assignable to one copy or another.

After selecting the target genes, the next step was to choose the population in which the SNP discovery phase would be carried out. Bearing in mind that the aim of this particular study was not only to discover and genotype SNPs but subsequently clone haplotypes and functionally interrogate individual SNPs, the selection of a panel that would maximise the number of SNPs discovered would not necessarily be ideal. There were two possible strategies to follow; either selecting a panel of individuals of diverse ethnic origin or a larger population from a single ethnic group. The ethnically diverse panel would likely yield more SNPs than the single population, as the individuals will be more diverged. However, as the SNPs would have been arising in parallel lineages prior to being placed in the same pool, the haplotypes would be more different from each other than would be the case if the population were ethnically homogeneous. This is equivalent to the effect of admixture on the linkage disequilibrium patterns in a population (Hartl and Clark 1997; Huttley et al. 1999; Pritchard and Przeworski 2001). When two or more previously isolated populations mix, linkage disequilibrium increases, particularly when the admixture is sudden. With a larger single population, there may be fewer SNPs discovered (depending on the relative sizes of the panels), but these SNPs will have been segregating in the same population, and there will have been more recombination between them. This increase in the number of combinations of alleles means that, when using the naturally occurring haplotypes to investigate the effect of polymorphism on gene expression, the effect of individual SNPs can be interrogated more easily. Another important consideration was that for a given number of individuals, an ethnically diverse panel

comprising a smaller set of samples from different populations would lead to a bias against the discovery of rare SNPs compared to a panel with no substructure. For example, in a panel of 10 individuals in a single population, the lowest minor allele frequency that could be determined by diploid resequencing would be 0.05 (a single heterozygote in the panel). If this panel was composed of 2 individuals each from 5 sub-populations, then the lowest determinable allele frequency for a lineage-specific SNP would be 0.25 (a single heterozygote out of the 2 individuals in the sub-population). Of course, SNPs with minor allele frequencies below these thresholds would still be discovered, but they would be disproportionately less likely to be found compared to more common alleles, and their true allele frequency would not be determinable below these values. For these reasons, it was decided that that a moderate-size panel of individuals from a single population would be used in order to obtain more haplotypes for the given number of SNPs.

The panel chosen for these experiments comprised of 48 unrelated individuals from the CEU pedigrees collected by the Centre d'Etude du Polymorphisme Humain (CEPH). These pedigrees are of families of European origin resident in Utah, in the United States of America. The 48 individuals chosen were grandparents from 12 families, with the maternal grandfather of one family replaced with one from a 13th family due to the unavailability of DNA samples. They were originally earmarked for genotyping in the then-nascent HapMap project. Since this study began, 17 individuals from this panel were dropped from the HapMap project due to poor viability of the transformed cell lines, and thus lack of availability of the corresponding DNA samples. The remaining 31 individuals still provide a good overlap with the HapMap data, which gives good scope for confirming a subset of SNPs found in this project. Using a similarly-sized panel from the Yoruba CEPH pedigrees (of African rather than European descent) may have increased the number of SNPs found, due to the larger genetic diversity in African populations (Przeworski, Hudson, and Di Rienzo 2000). However, this particular CEU panel was already being used in large-scale re-sequencing projects at the Sanger Institute. Thus there was a ready supply of the DNA samples available, and all necessary ethical approval and other regulatory procedures were already fulfilled. In addition, several panels of CEU individuals have been the subject of expression microarray analyses demonstrating substantial hereditary variation in gene expression (Monks et al. 2004; Morley et al.

2004; Cheung et al. 2005), which is good evidence that *cis*-regulatory variation is there to be found in this population.

The strategy used for re-sequencing promoters in this project is based on a SNP discovery pipeline used by the ExoSeq group (A. Dunham *et. al.*) at the Sanger Institute (Figure 4).

The genomic regions to be re-sequenced were divided into target fragments of around 500 base pairs each. Primers were designed to these fragments and used to PCR them from a panel of DNA samples from different individuals, with a separate PCR for each sample. These were then sequenced in both directions using the individual PCR primers as sequencing primers. The resulting 96 sequences for each fragment (2 sequences for each of the 48 PCRs) were then aligned *in silico*, and SNPs called using specialised SNP-finding algorithms based on sequence quality, relative peak height and confirmation by multiple sequence reads. A further primer test step was added to the strategy prior to the PCR and sequencing, in order to conserve resources.

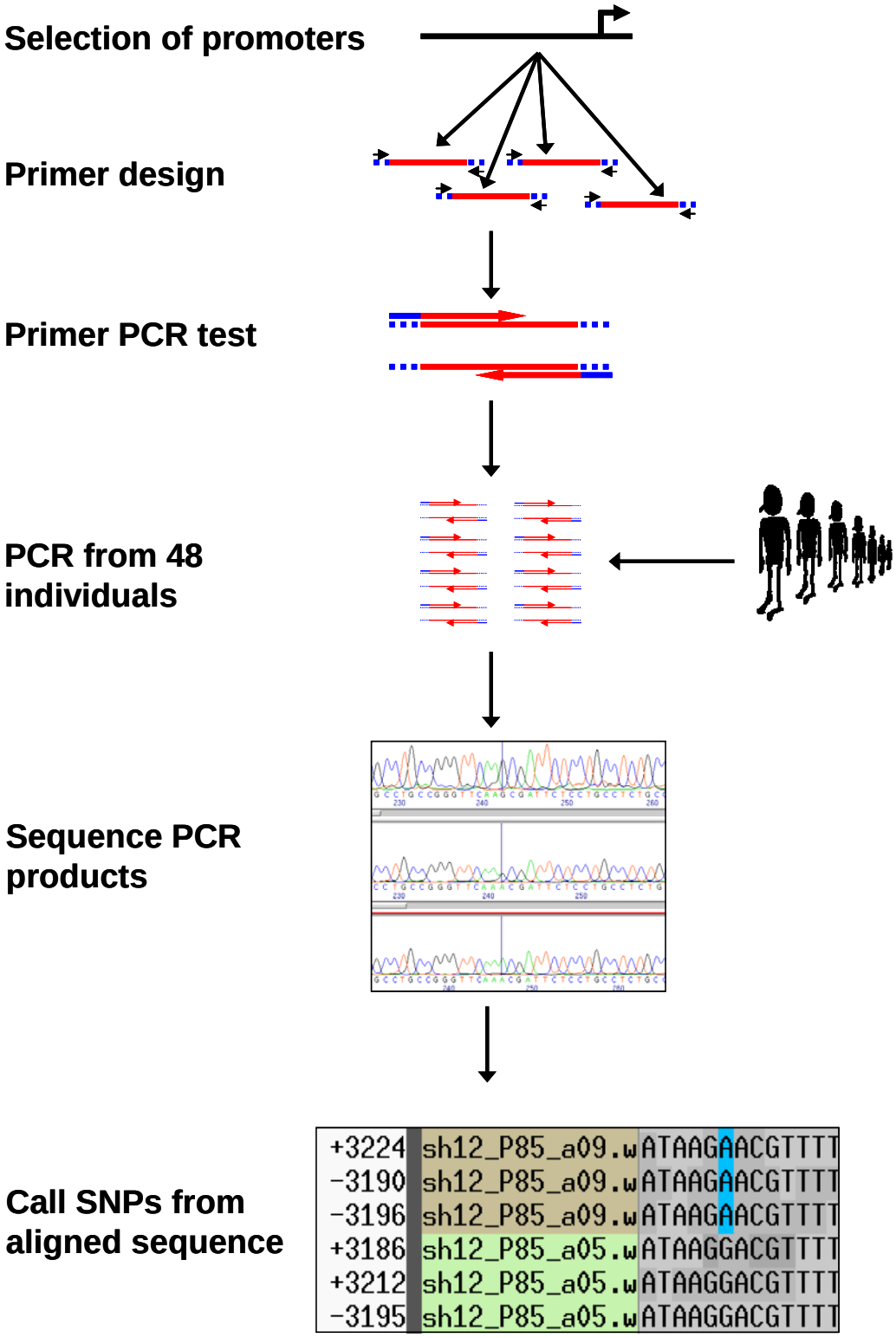


Figure 4. Flow diagram of the strategy used to mine the promoters of chromosome 22 for SNPs. Promoters were identified from the latest annotation and those in low copy repeat regions discarded. PCR primers were designed to amplify the promoters in 4 approximately equal segments, and the conditions for each primer pair optimised. Primers pairs that are successful were used to amplify the corresponding fragment from each of 48 unrelated individuals. The PCR products were sequenced and the sequences aligned and analysed computationally for evidence of SNPs.

3.2 Results

3.2.1 Selection of promoters for SNP-mining

The initial list of 393 candidate genes whose promoters were to be sequenced consisted of those with experimentally confirmed 5' ends according to the latest published annotation of chromosome 22 (Collins et al. 2003). This list excluded known pseudogenes and non-coding transcripts. Promoter sequences for each gene were identified in the human genome sequence (NCBI build 34) by finding each transcription start site (TSS) and extracting the sequence 2 kilobases upstream and 50 bases downstream.

The promoters were mapped back to the genome by BLAST, and the results analysed by eye in order to identify promoters which matched multiple places in the genome. This process eliminated 50 genes, leaving 343 candidates (appendix A). Of the genes eliminated, 19 belong to known gene families, with the remainder probably the result of isolated duplications. 20 genes from known gene families were not eliminated in this way, suggesting that the promoter sequences may have diverged sufficiently for them to be easily distinguishable.

3.2.2 Primer design

Each promoter was divided *in silico* into 4 adjacent target regions for PCR, and a unique pair of primers was designed for each (Figure 5). 100 base pairs either side of each target region was allowed for the placement of primers, in order to keep the total length of each product at no greater than 714 bases. This was considered to be both an easy size for PCR, and the length above which most sequencing traces would start to show marked decreases in quality.

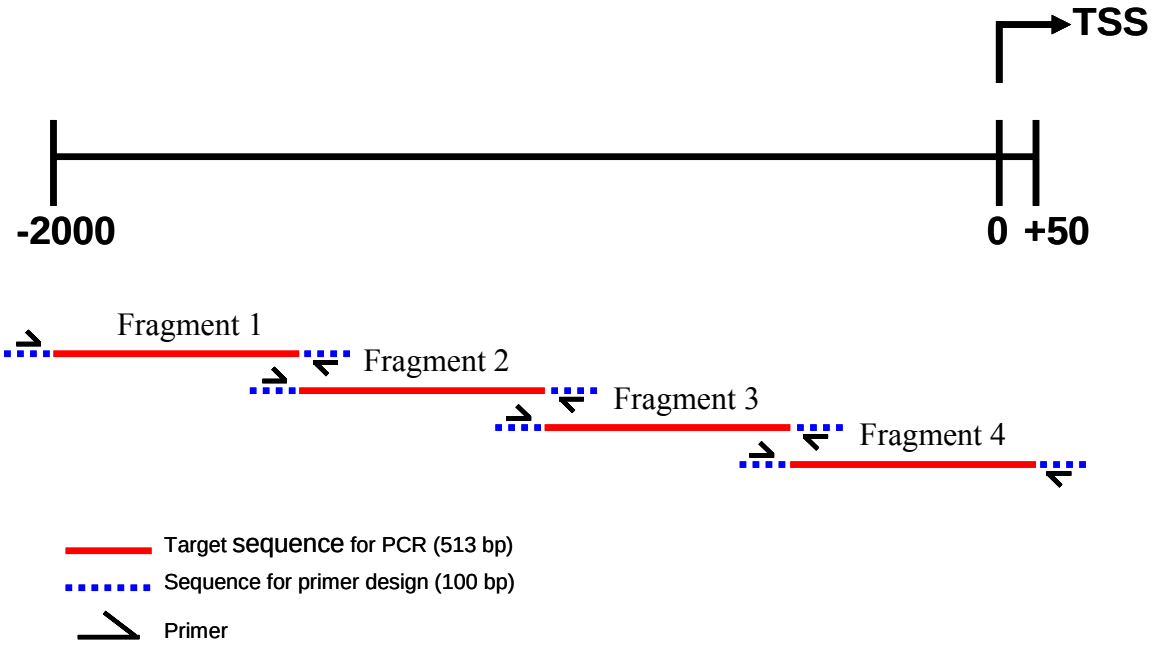


Figure 5. Schematic of the primer design strategy for re-sequencing promoters.

The primers were designed using Primer3, with some default parameters adjusted to aid primer design in GC-rich regions (see methods and materials). Primers for all 4 segments were successfully designed for 312 promoters, with the remaining 31 promoters missing 1, 2 or 3 primer pairs (Table 2).

# Fragments	Promoter coverage		# Promoters
4	1,2,3,4		312
3	1,2,3		2
	1,2,4		6
	1,3,4		9
	2,3,4		6
2	1,2		2
	3,4		1
1	1		5

Table 2. Coverage of the promoter sequences by successfully designed amplicons. Coverage is shown by listing the numbers of the amplicons designed as well as diagrammatically. Amplicon 1 is designated as the 5'-most fragment, and amplicon 4 the 3'-most.

3.2.3 Primer tests and PCR optimisation

Before amplifying the fragments from the 48-person CEPH panel, each pair of primers was tested in PCRs on standard genomic DNA under several sets of reaction conditions.

The sequencing pipeline that was to be used for these fragments was originally designed for very large genome-scale projects. It is currently being used by the ExoSeq group (A. Dunham et al) to re-sequence all exons in the human genome in the same panel of 48 individuals. As such, economic and technological considerations were a significant factor in the design of the experimental and computational components of the pipeline. Crucially for a relatively small project like this one, the *in silico* tracking system was not designed to cope with incomplete microtitre plates of PCR products, as it was assumed that these would not exist in a large project, and the high throughput fluid handling technologies on site would necessitate economically unfeasible waste of reagents and enzymes on empty wells. It was therefore important to keep the number of different conditions small, and thus the number of full plates of fragments per condition large. This would minimise the loss of any fragments left over after all full plates had been processed.

The pipeline was designed to use the same pair of primers for the PCR and sequencing steps. Ideally, it would be better to use an internal pair of primers for sequencing, as this would suppress the signal from any secondary products amplified by the PCR primers. This would double the number of primers required for each reaction, and for economic reasons was not implemented. This means that it is especially important that non-specific amplification is minimised as much as possible, and the cleanliness of the sequencing reactions was more dependent on the specificity of the PCR.

Initially, all primer pairs were tested using a standard protocol for genomic PCR that had been optimised by Bentley et al (unpublished). 892 (62%) gave clean bands with the standard protocol, 245 showed non-specific amplification and 228 showed weak

or no amplification (Table 3). The latter two categories were re-tested in new PCRs with different conditions designed to compensate for amplification problems.

#STSs	Annealing Temperature / ^{o}C				
	Standard protocol	**Non-specific**		**No product**	
		Stepped activation	*65^{o}C annealing*	*55^{o}C annealing*	*Betaine / DMSO*
Tested	1347	246	246	248	248
Successful	892	111	109	23	55
No product	228	not counted	not counted	120	164
Non-specific	245	not counted	not counted	105	29
Amplified	864	-	96	-	-

Table 3. Number of promoter fragments tested and successfully amplified in 5 different PCR conditions. The success of the PCRs was assessed by running the products on 1% agarose gels and manually inspecting the bands. The total number of amplicons tested in each condition is shown in the first row, and the primer test was designated successful if it showed a single band at the expected size, with no visible secondary bands. If no product was visible on the gel, the PCR was repeated using less stringent conditions (55^{o}C annealing) or additives to aid the processivity of the polymerase (betaine + DMSO). If multiple bands were visible on the gel, this was designated non-specific amplification, and the PCR was repeated using more stringent conditions (65^{o}C annealing) and by breaking up the polymerase activation step across the first 5 cycles rather than before the first cycle (stepped activation). The bottom row shows the number of amplicons processed through the sequencing pipeline using each condition.

3.2.4 PCR and sequencing of promoter fragments

A Tecan fluid handling robot was used to set up the PCRs. The primer pairs were grouped together according to their optimal annealing temperature in batches of 96, with each batch resulting in twelve 384-well plates of PCR products. The batches were quality-controlled after amplification by running samples from one plate from each batch on agarose gels to confirm that the PCR reactions had worked and the majority of products were present. The remaining primer and dNTPs in the reactions were removed with exonuclease I and shrimp alkaline phosphatase enzymes respectively, and the products were submitted to the Sanger Institute Sequencing Centre (SISC) for sequencing. The sequences were analysed using the ExoTrace analysis pipeline, and the SNPs automatically entered into the Sanger Institute's internal SNP database.

3.2.5 ExoTrace pipeline for sequence analysis and SNP detection

Prior to submission for sequencing, each plate of PCR products was assigned a barcode, containing information on the DNA sample and primer pair used in each well of the plate. This enabled each reaction to be tracked during the sequencing process, and resulted in each sequence trace being assigned to the correct individual and amplicon *in silico*.

The sequence traces were quality-checked and analysed for SNPs using ExoTrace. This is a set of algorithms and programs developed at the Sanger Institute by Dr. Steven Leonard (unpublished). There are two stages to the ExoTrace workflow; pre-processing and SNP calling.

The pre-processing stage uses raw sequence traces direct from the ABI sequencing machines, rather than those produced as a result of processing by the ABI software. This is because the ABI processing purposefully balances signal strengths across the four different channels, smoothes out peak shape and suppresses the signal in channels other than those of the called base. These processes mask the signals needed to reliably call heterozygous SNPs, and it is therefore desirable to avoid them. ExoTrace begins by applying a background correction to remove noise. It then applies a mobility shift to correct for the different rates at which the four fluorescent dyes move through the sequencing machine, which can cause overlapping peaks in the raw trace. Base-calling is carried out using PHRED (Ewing and Green 1998; Ewing et al. 1998), and the sequences aligned to their assigned reference by Crossmatch. Finally, the height of each peak is extracted for bases that align to the reference, giving a single value per base per channel.

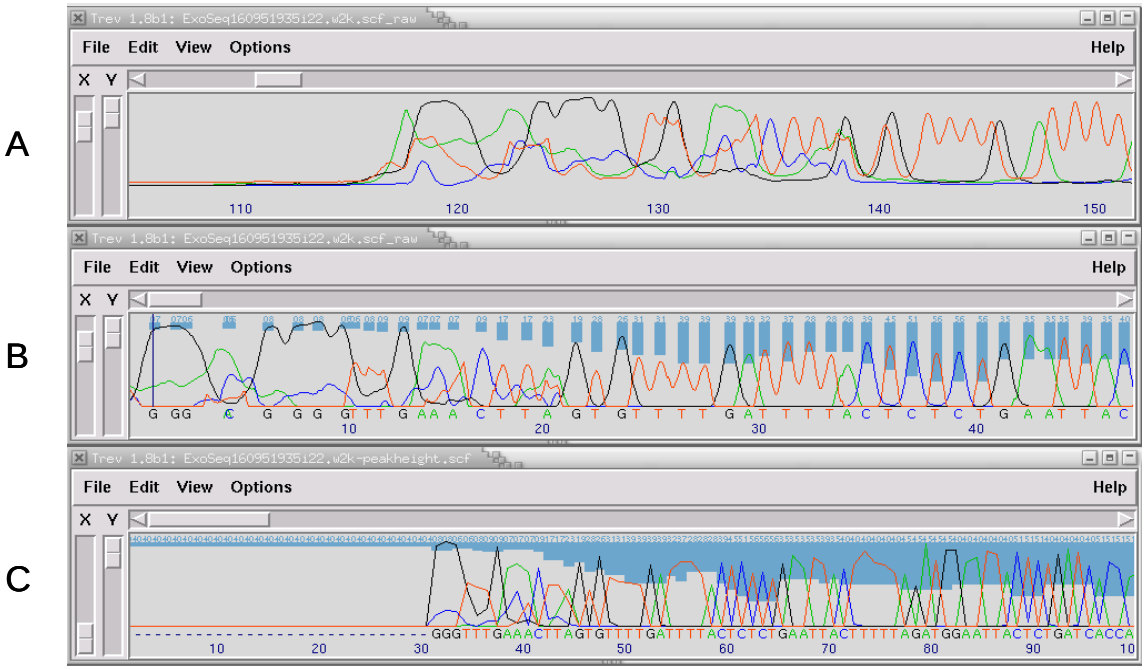


Figure 6. Pre-processing of raw sequence traces by ExoTrace prior to automated SNP calling. A) Raw unprocessed trace produced by the ABI 3730 sequencer. B) The same trace after background correction, mobility shift and base calling by PHRED. C) "Digitised" trace with a single value for each peak height. This figure was reproduced with permission from Dr. Steven Leonard.

In the SNP calling stage, individual reads are filtered according to whether they have sufficient signal strength and sequence quality, and whether they crossmatch to the reference sequence. Only sequence traces, and the bases within them, that align to the reference are used for SNP calling. Once the sequences are aligned, SNPs are called based on a comparison of expected and actual peak heights (Figure 7). Heterozygotes are called if the peak height of the reference base is around 50% of the expected value, and the height of the second highest peak is also 50%. Any heterozygotes must include the reference base as one of the two alleles. For a homozygous SNP to be called, the peak height in the reference channel must be small compared to the peak height of the called base, which must itself be at least 75% of the expected value. In both cases, if the peak height of the reference base is over 75%, no SNP is called. ExoTrace also requires that SNPs must be confirmed by sequence traces in both orientations. The only exceptions to this are if the called SNP matches one already present in dbSNP, or if all three genotypes are present among the aligned reads.

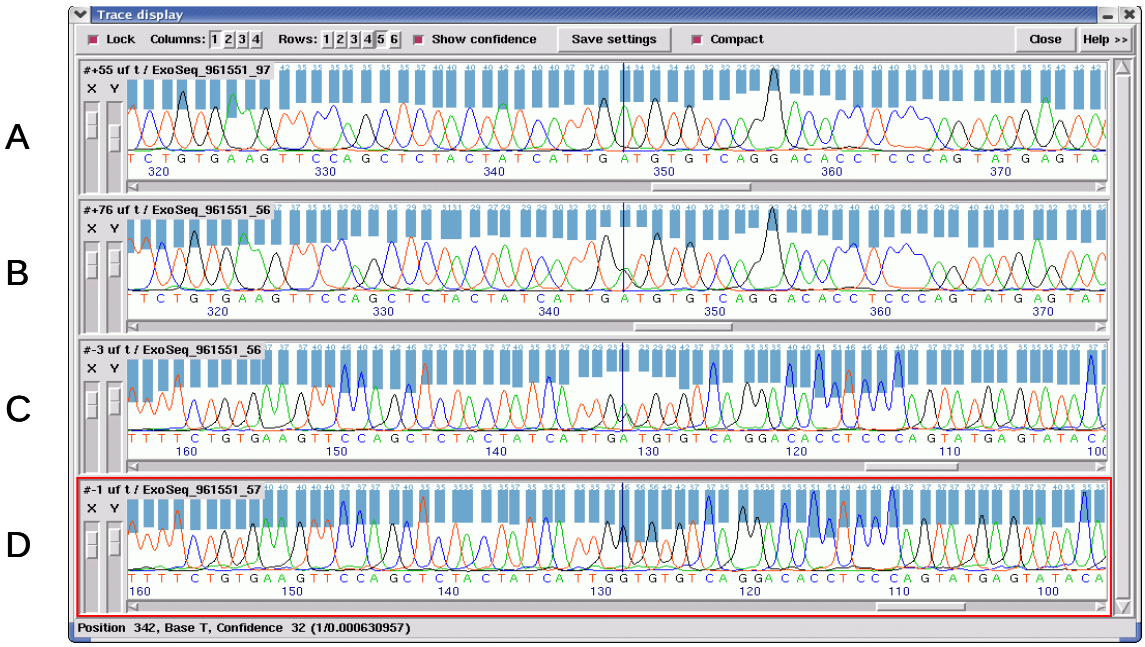


Figure 7. Four traces from a model SNP called by ExoTrace. A) Homozygous A. B and C) Heterozygous A/G. D) Homozygous G. Traces A and B are traces from the sense sequencing primer, and traces C and D from the antisense primer. This figure was reproduced with permission from Dr. Steven Leonard.

3.2.6 Second round of primer design

As more and more sequence from the promoter fragments was analysed, it was found that runs of single bases anywhere in an amplicon would usually cause a drastic drop in sequence quality when the polymerase processes through them. While the length that such runs have to be before they disrupt sequencing can vary, 8-10 bases seems to be size at which degredation of sequence traces becomes marked. Thus, sequencing traces from each end of the amplicon would be normal until the run of bases and practically unusable after it. This had the effect of masking SNPs anywhere in these amplicons from the ExoTrace software, because SNPs would only be detected in one direction and bidirectional confirmation is an important criterion for passing a SNP call.

A second round of primer design and re-sequencing was started, with a primer design strategy to compensate for this problem. Each promoter containing at least one run of 8 or more of the same base were selected for the second round. Primers were designed using Primer3 and the same parameters as the first round. Rather than split the promoters in to equal blocks of target sequences, they were split using the polyN runs

as boundaries (Figure 8). 129 promoters were found to contain polyNs, and the 558 new primer pairs were designed for them.

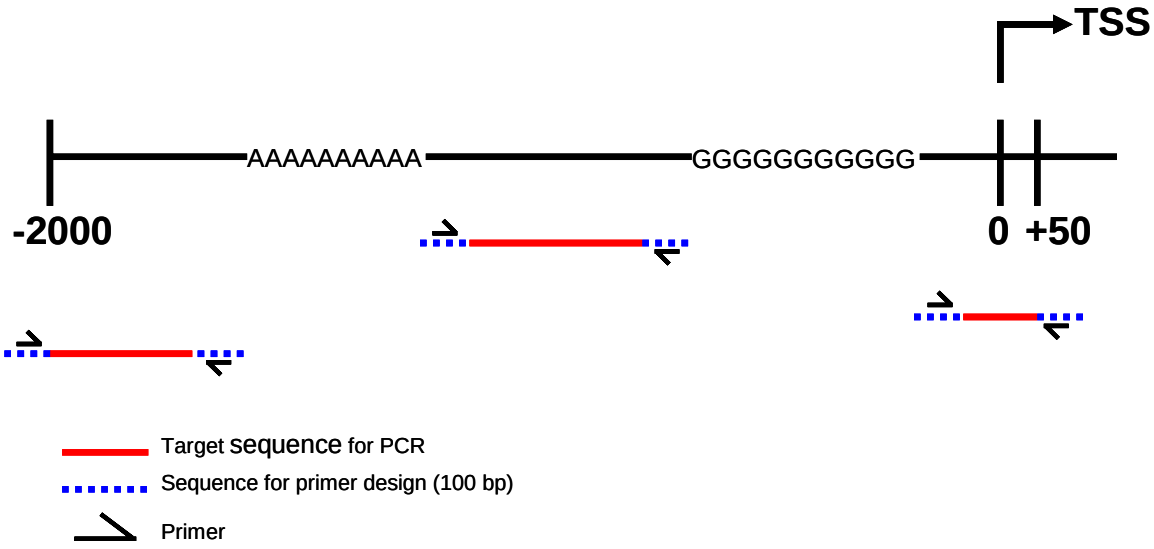


Figure 8. Schematic of the strategy for primer design around polyN motifs.

3.2.7 PCR tests of the second batch of primers

The second set of primer pairs was tested in two PCR conditions in parallel, with annealing temperatures of 60°C (standard protocol) and 65°C (increased stringency). The standard protocol successfully amplified 346 fragments, 26 more than the more stringent protocol. While the stringent protocol was able to clean up some reactions with non-specific amplification, this was more than made up for by the loss of products which had amplified well in the standard protocol. The standard protocol was therefore used to amplify those ampliconss that had passed the primer test, as well as an additional 38 amplicons with weak bands that were added to fill the final plate.

# Amplicons	Annealing Temperature / ^{o}C	
	60^{o}C	65^{o}C
Tested	558	558
Successful	346	320
Amplified	384	-

Table 4. Primer test results on the primer set designed around the polyN sequences.

3.2.8 Promoter sequencing results

In total, 1344 promoter fragments were amplified by PCR from each individual in the 48-person panel, requiring a total of 64,512 PCR reactions. These represented at least one fragment from 332 different promoters, or 96.8% of the original 343. 252 promoters (75%) returned at least one successfully sequenced amplicon. Of these, 131 (52%) had at least 75% of their sequence covered by successfully sequenced amplicons, and 208 (83%) were covered across at least 50% of their length.

Of all the amplicons submitted for sequencing, 1187 returned sequence of sufficiently high quality to be used for SNP calling (Table 5). The remainder failed due to poor quality traces (causing the amplicon to fail quality check) or because they did not crossmatch with the reference sequence (possibly due to slippage caused by low complexity sequence, or non-specific amplification leading to two sequences being present). Due to time constraints, amplicons that failed along the pipeline for any reason were not repeated.

# STSs	Primer set 1	Primer set 2	Total
Total	960	384	1344
Failed Quality check	83	74	157
Analysed for SNPs	877	310	1187

Table 5. Sequencing quality of the amplicons submitted for sequencing.

The initial round of primer design and re-sequencing yielded 630 SNPs that passed the ExoTrace criteria. The second round of re-sequencing added another 177 new SNPs, as well as re-confirming 92 that had been found in the first round. This gave a total of 807 SNPs. At the time the SNP discovery was first completed, 508 of the 807 SNPs (62.9%) were not present in dbSNP. However, in the latest version of dbSNP (build 125) that has now decreased to 26%. The SNPs were distributed evenly across the 2 kb promoter sequences, apart from two noticeable drops in SNP number around the overlaps between fragments from the first primer set (Figure 9). This is likely to be due to the relatively poor sequence quality near the ends of sequence traces, and it seems that the overlap of the amplicons in this case was not sufficient to completely compensate for this effect.

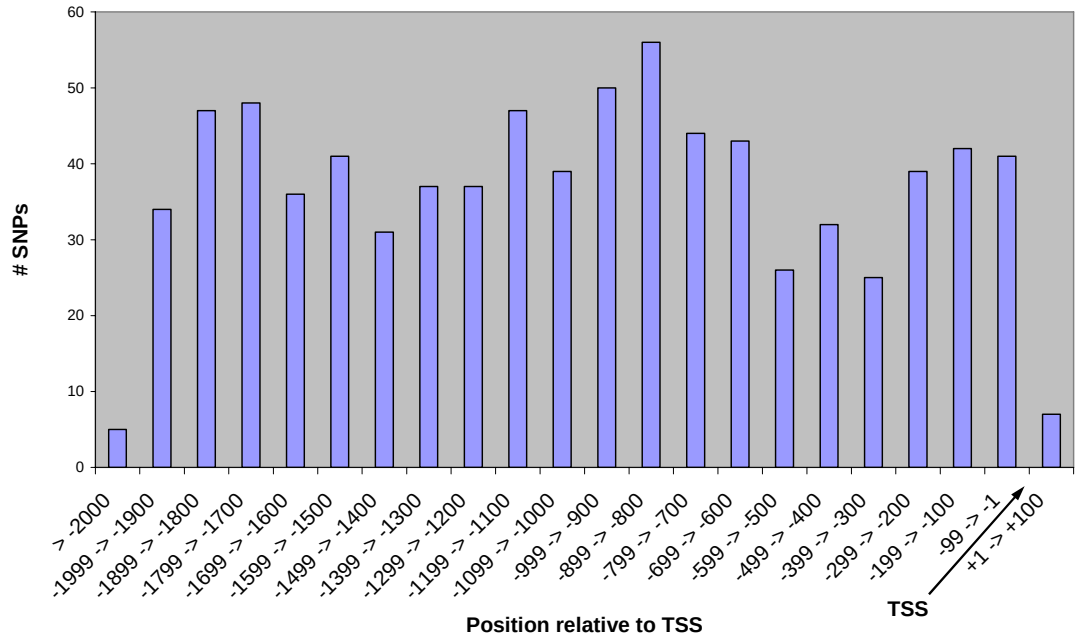


Figure 9. Distribution of SNPs relative to the transcription start site (TSS).

All SNPs were submitted to the Sanger Institute SNP database, and will subsequently be submitted automatically to dbSNP by an automated submission pipeline in place at Sanger. I also created a custom MySQL database for the purpose of this study. This made it far easier to carry out analyses and data manipulations, as the database structure was much simpler and was not constrained by the need to fit in with a laboratory pipeline. All SNPs found are listed in appendix B. An example of data from one of the promoters is shown in Figure 10.

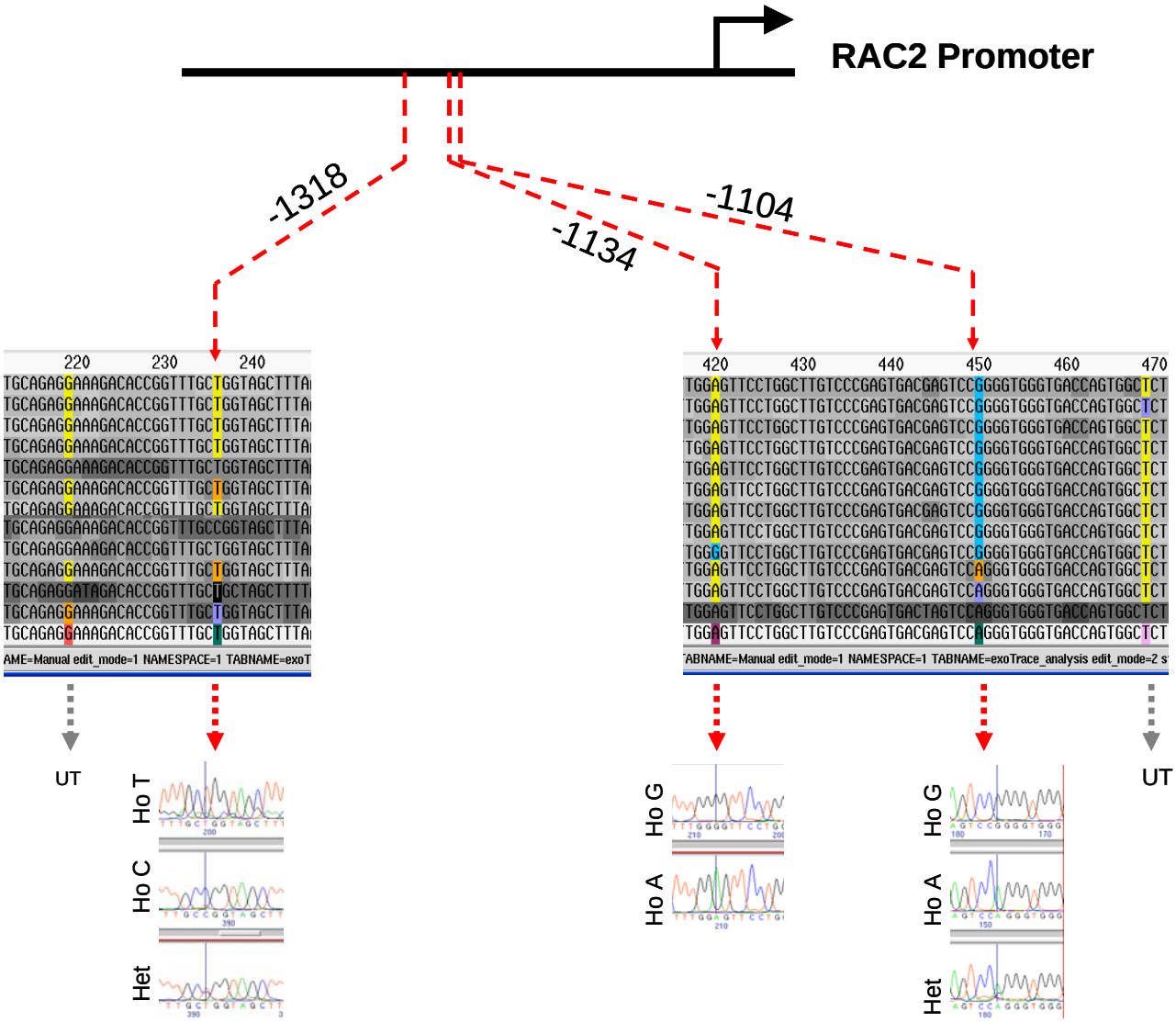


Figure 10. Schematic of the SNP-finding process using the RAC2 promoter as an example. A) Three SNPs were found in this promoter; a C/T SNP at -1318, an A/G SNP at -1134 and a third A/G SNP at -1104 from the TSS. B) All successfully sequenced PCR products were aligned and ExoTrace was used to detect putative SNPs based on the criteria outlined in section 3.2.5. Here, five ExoTrace calls are shown as columns of colured bases on the alignment. C) In this example, three of the five ExoTrace calls fulfilled the criteria (red dashed arrows) and were confirmed as SNPs, whereas two failed due to lack of bi-directional confirmation of putative variant calls (UT/grey dashed arrows).

3.2.9 Distribution of SNP types and allele frequencies

The minor allele frequency of each SNP was calculated by counting the homozygous and heterozygous calls on each of the 48 samples sequenced. This gives a frequency resolution of 1/96, or 0.0104, and means that alleles as rare as 0.01 minor allele frequency can be detected. This assumes that all 48 samples were amplified

successfully and sequenced to good quality. In practice, there is often a loss of a small number of samples due to stochastic failures in sequencing or PCR, meaning that many SNPs are called from fewer than 48 samples (and in some cases substantially fewer). This would be expected to push the minor allele frequency distribution in favour of common SNPs. As would be expected, the majority of SNPs found in the promoter re-sequencing had small minor allele frequencies (Figure 11).

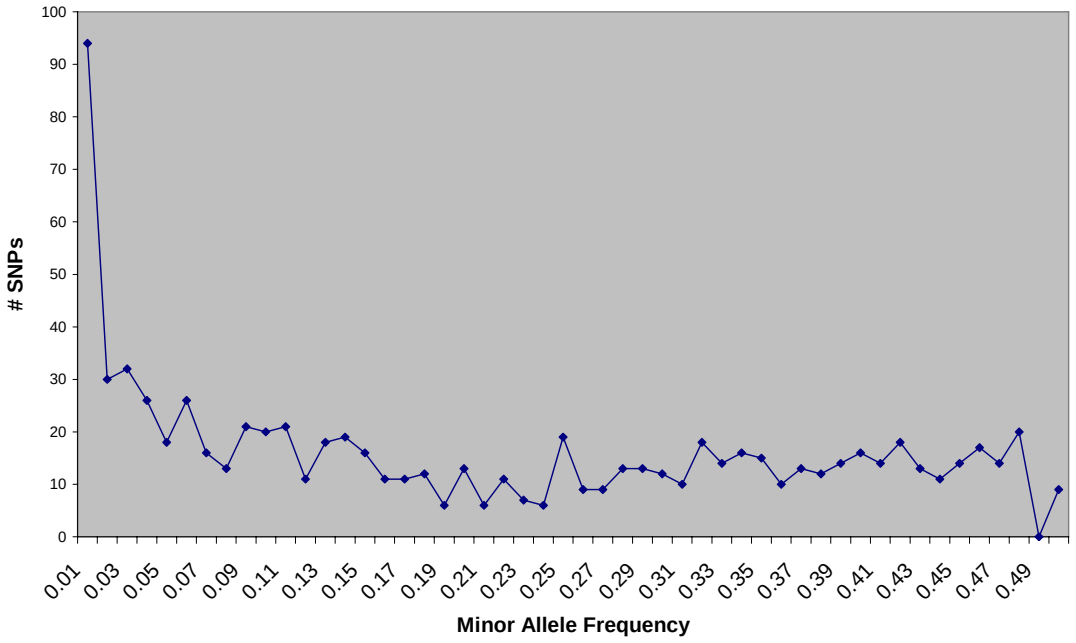


Figure 11. Distribution of the minor allele frequencies of chromosome 22 promoter SNPs.

The distribution of SNP types was compared to a control set from chromosome 22 as a whole in order to see whether there are any differences in the kind of SNPs found in promoters relative to what would be expected. The control set was made up of all SNPs in dbSNP build 125 from chromosome 22 that could be aligned to the chimpanzee genome. This was to enable later use of the chimp sequence to infer direction (see section 3.2.12). The proportions of the different SNP types did not deviate significantly (p-value = 0.19, χ^2) from that expected in the whole genome (Figure 12 A and B). This was somewhat surprising, as an under-representation of C/T SNPs due to the lack of methyl-cytosine deamination at promoters may have been expected. There was a small increase in C/G SNPs at the expense of transitions, consistent with higher GC content, but this was very small and not significant. The SNPs were divided into two sets according to their presence in CpG islands,

according to the CpG island annotation on the UCSC genome browser (NCBI build35). This revealed a marked difference in the distributions of SNPs in CpG islands relative to chromosome 22, with far fewer A/T SNPs and transitions. This may be due to a combination of lack of cytosine methylation and elevated GC content. However, the distribution of promoter SNPs outside CpG islands is not significantly different from that of promoter SNPs generally, or from chromosome 22 as a whole (p-value = 0.53, χ^2).

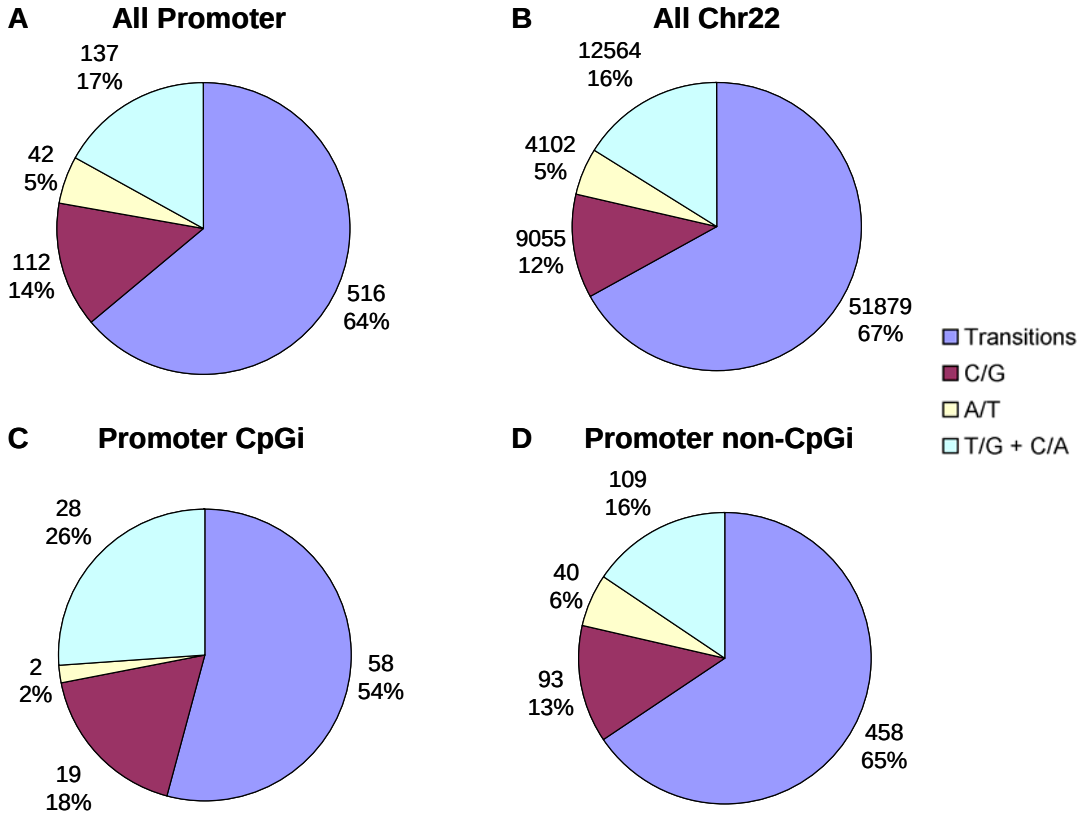


Figure 12. Distributions of the SNP alleles relative to chromosome 22 and to CpG islands in promoters. A) All SNPs from the promoter re-sequencing dataset. B) SNPs from chromosome 22 (Collins et al.). C) Promoters SNPs within CpG islands (according to the UCSC genome browser). D) Promoter SNPs outside CpG islands.

3.2.10 Comparison of polymorphic promoters with downstream gene function

If promoter sequence polymorphism has an effect on the level of gene expression, then one can hypothesise than some functional classes of genes would be more tolerant of such changes than others. For example, genes involved in crucial processes such as cell cycle control or DNA damage repair might be hypothesised to have lover

mutation rates at their promoters compared to other genes such as extracellular receptors due to purifying selection eliminating variation in the former. A recent study has found evidence that genes are preferentially located in mutational hot or cold spots depending on their function (Chuang and Li 2004). In order to test this idea, the Gene Ontology (GO) terms associated with genes having polymorphic promoters was compared to those for the genome as a whole. Five different lists of 1000 randomly selected human genes were generated by Juanma Vaquerizas at the European Bioinformatics Institute to use as comparisons with the list of genes with polymorphic promoters discovered in this project. The FatiGO tool (Al-Shahrour, Diaz-Uriarte, and Dopazo 2004) was used to compare these lists of genes across all GO hierarchies and levels. No significant over- or under-representations of GO terms were found for any level of the GO structure (Figure 13).

Figure 13. Comparison of the Gene Ontology terms for genes with polymorphic promoters (orange bars) against a list of 1000 randomly selected genes (green bars). This analysis was carried out using the FatiGO web tool, and was repeated for five different control lists of 1000 random genes. In all cases, no significant differences were found between the functional categories of genes with polymorphic promoters and the control sets. The three categories with the most significant differences are shown for each of the three GO heirarchies. Raw p-values are shown on the left-hand numerical column, and adjusted p-values on the right-hand column.

3.2.11 Analysis of the genomic context of promoter SNPs

Once can hypothesise that SNPs that affect gene expression levels do so because they disrupt a sequence element important to the regulation of that gene. While the difficulty of identifying such elements from sequence has been discussed, one could tentatively predict SNPs with potential regulatory function by seeing which ones co-localise with motifs of putative functional importance. Data on putative regulatory elements in the chromosome 22 promoters was downloaded from their respective databases or the UCSC or Ensembl genome browsers, and entered into the custom database containing the SNP data. The positions of all 807 SNPs were analysed for co-localisation with motifs of potential regulatory significance using MySQL search queries (Table 6). The details of each element type examined are below:

Phastcons regions: The Phastcons program identifies sequences within a cross-species sequence alignment that are highly conserved (Siepel et al. 2005). This data was obtained from the UCSC genome browser, and is for a multiple alignment of 5 vertebrates (human, mouse, rat, chicken, and *Fugu rubripes*).

cisRed motifs: The cisRED database (Robertson et al. 2006) holds a large collection of putative regulatory motifs discovered using a pipeline that incorporates three previously developed motif-finding algorithms, CONSENSUS (Hertz and Stormo 1999), MEME (Bailey and Elkan 1994) and MotifSampler (Thijs et al. 2002). The data was obtained using BioMart from the Ensembl genome browser.

Transcription factor binding sites (TRANSFAC): The TRANSFAC database of TFBS matrices is the largest and one of the most well-established databases of binding sites. It is a proprietary database with a reduced-data version available to the public. MATCH 2.1 Public (Kel et al. 2003) was used to scan the promoter sequences for binding sites, using the pre-set parameters designed to minimise false positives. Genomic coordinates for the binding sites were then calculated using the offset of the binding site from the known promoter start and end coordinates.

Transcription factor binding sites (JASPAR): JASPAR is a manually curated database of TFBS matrices (Sandelin et al. 2004). It only contains binding sites based on experimental evidence (such as SELEX experiments) and has a relatively small collection of non-redundant binding sites, in contrast to TRANSFAC which contains considerable redundancy and unverified sites. The JASPAR matrix set was downloaded from the web and promoter sequences were scanned using the MotifScanner program (Aerts et al. 2003) and a threshold of -6. Low quality motifs that hit the promoters more than 200 times were eliminated using a custom script. Genomic coordinates for the binding sites were then calculated using the offset of the binding site from the known promoter start and end coordinates.

Conserved TFBS: These represent TFBSs as defined by the TRANSFAC binding site matrices, and which are conserved between human, mouse and rat. All conserved TFBS sites on chromosome 22 were downloaded from the UCSC genome browser

Putative quadruplex sites: These are short purine-rich sequences that are capable of forming quadruplex loop structures within a single strand of DNA. They have been shown experimentally to be important in *cis*-regulation in at least one case (Seenisamy et al. 2004), and their pattern of distribution across the genome suggests that a certain proportion have some *in vivo* function (Huppert and Balasubramanian 2005). The coordinates for all putative quadruplex sites on chromosome 22 were provided by Julian Huppert.

	# SNPs	% SNPs	Observed / Expected SNPs	p-value (χ^2)
All SNPs	807	100	n/a	n/a
phastcons regions	40	4.96	0.67	7.87E-03
cisRED motifs	21	2.60	0.72	1.22E-01
TFBS (TRANSFAC)	36	4.46	0.94	6.94E-01
TFBS (JASPAR)	41	5.1	3.72	4.60E-07
Conserved TFBS	9	1.12	1.67	1.20E-01
Quadruplex sites	6	0.74	0.56	1.52E-01
SNPs in putative regulatory regions	130	16.1	0.94	4.55E-01

Table 6. Co-localisation of SNPs with putative regulatory sites motifs. The number of SNPs within the boundaries of an element in each functional category was calculated using a mySQL database. The majority of the coordinates were either downloaded from the UCSC or Ensembl genome browsers, while the TRANSFAC and JASPAR TFBS analyses were done *de novo* on the promoter sequences. The ratio between the number of SNPs observed in each functional category relative to the number of SNPs expected given the proportion of the promoters covered by the elements is shown, and the significance of this shown by p-value from a χ^2 test.

A total of 130 (16.1%) SNPs were found to be in a region of the genome that may be involved in transcriptional regulation (Table 6). In terms of the individual functional categories, this ranged from 6 (0.7%) to 68 (8.4%) SNPs. Some SNPs were found in multiple categories, and hence the total number of SNPs is less than the sum of the individual categories. It could be proposed that if these putative elements were really functional, then they may be less polymorphic than the surrounding promoter sequence due to possible purifying selection. This was tested by calculating the percentage of the total promoter sequence that was covered by each element category, and comparing the number of SNPs in each category with the number that would be expected if the SNPs were distributed randomly across the promoter using the χ^2 test. This showed that overall, putative functional elements were not any less polymorphic than would be expected by chance (Table 6). Only one of the categories, ultra-conserved elements from the phastcons track in the UCSC genome browser, showed a significant under-representation of SNPs. However, as these elements are defined by conservation, this was not surprising.

In addition to the above motifs, the SNPs were checked for regulatory potential using the 5x regulatory potential score (King et al. 2005) on the UCSC genome browser. This score is based on the similarity of conservation patterns in a training set of experimentally verified regulatory elements compared to a control set of non-regulatory ancestral repeat sequences, and has been computed from alignments of human with chimp, mouse, rat and dog. The score for each base represents a 100 base pair window centred on that base. 239 SNPs (29.5%) had scores greater than 0.01, which indicates that the base is in a sequence with very similar alignment patterns to known regulatory motifs. 73 of these were also present in at least one putative regulatory motif.

When combining these different analyses, 296 SNPs (36.7%) emerge as having some evidence of regulatory potential, whether because of its location in a putative regulatory motif or its regulatory potential score.

3.2.12 Evolutionary analysis of the SNPs using the primate genomes

In order to determine the directionality of the nucleotide changes, the draft chimp and macaque genomes were used to root each SNP. GALAXY 2.1 (Giardine et al. 2005) was used to extract the ancestral alleles from pre-computed alignments of the human genome to the chimp genome (Consortium 2005a) and, where there was no alignment to chimp, the macaque genome. 780 SNPs (96.7%) were accounted for in this way, with the remainder lying in areas not covered by these alignments. This is significantly better than the 80% of human SNPs rooted on publication of the draft chimpanzee genome (Consortium 2005a), reflecting the contribution of the macaque genome and possibly some improvement in the quality of the chimpanzee sequence since publication. The major allele in human is ancestral in 559 SNPs and derived in 199 SNPs. 10 SNPs are present in the alignment but have no corresponding base in chimp, possibly representing insertions in the human lineage or deletions in the chimp lineage. For 12 (1.5%) SNPs neither allele matched the chimp base. This may be due to an error in the chimp genome sequence or orientation of the chimp contig, although it is not impossible that some can be due to the base changing in both species. In total, 39 SNPs (4.8%) could not be rooted with either genome, slightly higher than the rate seen in previous comparisons (Dermitzakis *et al*, unpublished).

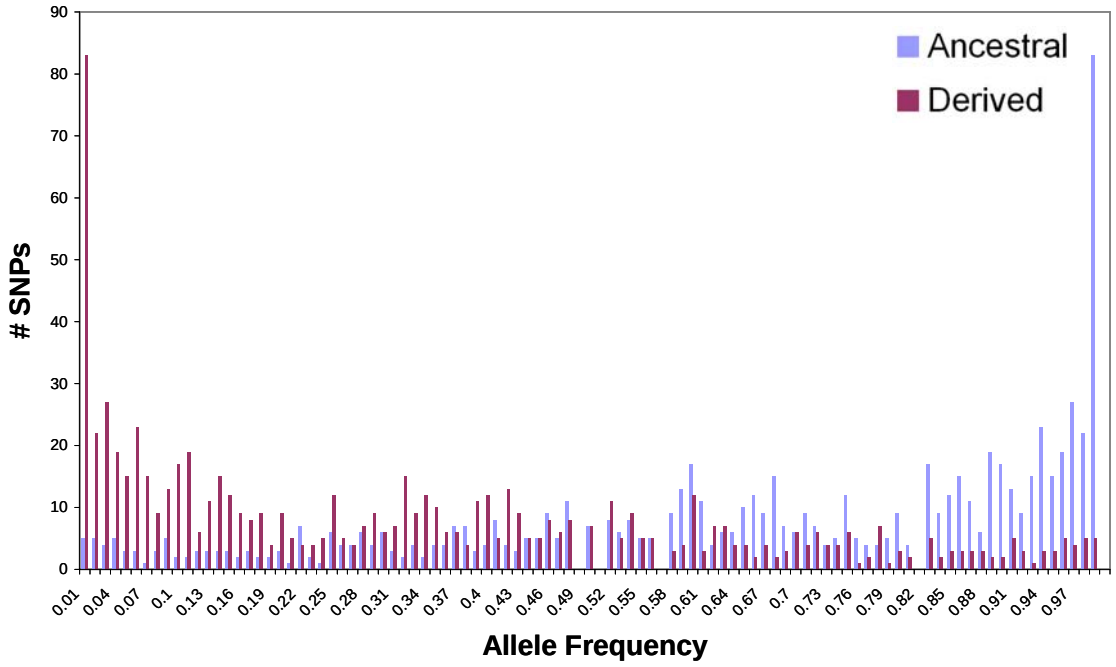


Figure 14. Allele frequency spectrum for ancestral and derived alleles rooted with the chimpanzee and macaque genomes. The two distributions are symmetrical due to the relationship between the two allele frequencies (i.e. one frequency is 1 minus the other frequency). There is a marked bias of derived alleles towards low allele frequencies, with most ancestral alleles being common.

In 185 of the 244 successfully rooted SNPs in putative regulatory elements or a high 5x regulatory potential score, the major allele was ancestral, and in 59 it was derived. This is not significantly different from the proportions for promoter SNPs as a whole ($p = 0.56$, Fisher's exact test).

The spectrum of mutations in promoters was compared to that for chromosome 22 as a whole, in order to determine whether there were any differences in the mutational processes operating at promoters compared with the rest of the genome. The genomic coordinates of all SNPs on chromosome 22 were downloaded from dbSNP and rooted with GALAXY in the same was as the promoter SNPs. 77600 SNPs were successfully rooted using the chimp and macaque genomes. A matrix was then constructed of all possible mutations and the number of such changes in chromosome 22 promoters and in the chromosome as a whole (Table 7).

		Derived Allele			
		A	**G**	**C**	**T**
Ancestral Allele	**A**		82 (10.8) *8929 (11.5)*	15 (2.0) *2742 (3.5)*	11 (1.5) *2070 (2.7)*
	G	164 (21.7) *17219 (22.2)*		52 (6.7) *4532 (5.8)*	39 (5.1) *3527 (4.5)*
	C	36 (4.8) *3587 (4.6)*	49 (6.6) *4523 (5.8)*		163 (21.5) *17080 (22.0)*
	T	22 (2.9) *2032 (2.6)*	27 (3.6) *2708 (3.5)*	97 (12.8) *8651 (11.1)*	

Table 7. Matrix of promoter SNP alleles including the direction of the mutations. The direction of each SNP is from the allele on the row to the allele on the column. The top row of each cell denotes the number of promoter SNPs, with the percentage of the total in brackets. The bottom row (in italics) denotes the same numbers but for the whole of chromosome 22. All mutations are shown as + strand mutations. As it is not in fact possible to determine which strand in a base pair has mutated, it is necessary to combine the numbers of SNPs from reciprocal pairs to get a truer reflection of the proportions of different mutations. Reciprocal pairs are shaded in the same colour above.

There were no striking differences between the proportion of each type of SNP between promoters and chromosome 22, although there were large differences between the proportions of SNPs within each category. In order to gain a clearer picture of any differences, the forward and reverse mutation rates for each SNP type were compared for the two categories. This was done by combining SNPs that were reciprocal to each other (for example, an A to G mutation on a given strand is equivalent to a T to C mutation on the opposite strand, so the two were added together). This resulted in six mutation classes rather than eight, as A to T and C to G SNPs cannot be differentiated from their reciprocals even with primate genomes. (Table 8). Each category was tested for a significant deviation from its expected proportion on chromosome 22 by using the χ^2 test, and by calculating the expected SNP number as being the same proportion as the same category in the rooted SNP list. No significant difference in any of the mutation categories was found between promoters and chromosome 22 overall (Table 8). Surprisingly, no decrease in C to T mutation was seen. This would have been expected, as it is known that methylated cytosines in CpG dinucleotides mutate to thymine by deamination at an accelerated rate, and that promoter sequences tend to be unmethylated in the human genome.

Mutation	# SNPs Observed	% mutations in Chr22	# SNPs Expected	p-value (χ^2)
All promoter SNPs				
C->T | G->A	327	44.2	335	0.578
C->A | G->T	75	9.1	69	0.480
C->G | G->C	101	11.7	88	0.152
A->T | T->A	33	5.3	40	0.254
A->C | T->G	42	7.0	53	0.112
A->G | T->C	179	22.7	171	0.515
Promoter SNPs within 500 bp of TSS				
C->T | G->A	70	44.2	76	3.22E-01
C->A | G->T	18	9.1	16	5.73E-01
C->G | G->C	32	11.7	20	5.15E-03
A->T | T->A	9	5.3	9	9.61E-01
A->C | T->G	10	7.0	12	5.22E-01
A->G | T->C	34	22.7	39	3.46E-01
Promoter SNPs in CpG islands				
C->T | G->A	44	44.2	46	7.62E-01
C->A | G->T	16	9.1	9	2.52E-02
C->G | G->C	18	11.7	12	6.64E-02
A->T | T->A	2	5.3	5	1.29E-01
A->C | T->G	10	7.0	7	2.86E-01
A->G | T->C	13	22.7	23	1.50E-02

Table 8. Comparison of directional changes in chromosome 22 promoters with the distribution of the same changes in chromosome 22 as a whole. The chromosome 22 distributions were used to calculate the expected number of promoters SNPs in each category, and the χ^2 test was used to assess the significance of the departure from the expected value for each mutation type.

Recent work at the Sanger Institute has quantified the degree of methylation at promoters, and discovered a methylation trough around TSSs that extends approximately 1 kb upstream and downstream (Beck et al unpublished). Relatively highly methylated DNA in the 5' half of the sequenced promoters may therefore have been masking a decrease in C to T mutations proximal to the TSS. To check for this effect, the analysis was repeated using only SNPs within 500 base pairs of the TSS. Again, no decrease was detected, although a significant overrepresentation of C/G SNPs was detected (Table 8). This may be due to elevated GC content at promoters in general, which would be expected to raise the number of C/G SNPs relative to all other mutation classes. Finally, the analysis was repeated a third time using promoter

SNPs within CpG islands. Even this analysis failed to show a significant under-representation of C to T mutations. This was a real surprise, as CpG islands are thought to arise from precisely this mutation bias. However, two biases were detected in this category of SNPs; a marked over-representation of C to A and G to T changes, apparently at the expense of A to G and T to C mutations (Table 8). This again can be explained by elevated GC content, which would be expected to be higher in CpG islands than even promoters as a whole. An increase in C/G SNPs was also detected in CpG islands, but this fell just short of statistical significance.

3.2.13 Association of promoter SNPs with gene expression levels

The lab of Dr. Manolis Dermitzakis at the Sanger Institute has recently carried out whole genome expression studies of all individuals in the HapMap Project using Illumina array technology (Stranger *et al*, unpublished). The aim of that study was to find SNPs that are associated with polymorphic gene expression levels, using the HapMap SNPs as their SNP resource. As 31 of the 48 individuals in my study overlapped with HapMap, it was possible to investigate the association of the promoter SNPs in each polymorphic promoter with the expression levels of the gene it regulates.

A script developed in the Dermitzakis lab was used to run an association analysis between all promoter SNPs found in this project and the expression levels of the downstream genes. Genotypes for the 31 individuals for which expression data was available were extracted and parsed into the appropriate format using a custom perl script, and the data passed to Barbara Stranger in the Dermitzakis lab where the association script was run. Multiple testing was corrected for using the Bonferroni correction method.

Only one promoter SNP was significantly associated with an expression phenotype in the downstream gene. This was a C/T SNP 1747 upstream of the TSS of the *SNAP29* gene. The *SNAP29* protein is involved in intracellular vesicle trafficking in neurons, and truncation of the protein has been linked to severe neurocutaneous abnormalities (Sprecher et al. 2005). Interestingly, previous studies have reported a significant association between another SNP in the *SNAP29* promoter and schizophrenia (Saito et

al. 2001; Wonodi et al. 2005). This was an A/G SNP at -849 bases from the TSS, and is present in dbSNP as rs165596. The G allele in this SNP was found to be significantly overrepresented in schizophrenia patients relative to control groups. This SNP was not detected in the promoter SNP mining as the amplified fragment in which it would be located failed to return usable sequence. The -1747 C/T SNP is novel and has never been reported before. However, the relationship between the two SNPs could still be determined because rs165596 was genotyped in the HapMap project.

Genotypes for 6 SNPs in a window of approximately 9kb to the -1747 C/T SNP, including rs165596, were downloaded from the HapMap dataset for the 31 individuals overlapping with SNP-mining panel used here. HaploView was then used to predict the haplotypes present in this region (Figure 15). The total of 7 SNPs were present in only 3 haplotypes across the 9kb window, showing tight linkage disequilibrium. Haplotypes 1 and 2 were much more common than haplotype 3, with frequencies of 0.5 and 0.42. These contained A and G alleles at rs165596 respectively, and both carried the common C allele at -1747 C/T. The third haplotype had a frequency of 0.08, and was formed by the mutation at -1747 C/T taking place in the background of haplotype 2 (Figure 15). The C allele at -1747 C/T segregates with haplotypes 1 and 2, and hence with either allele of rs165596 almost equally. However, the T allele at -1747 C/T segregates exclusively with the G allele at rs165596 according to this data (Figure 15).

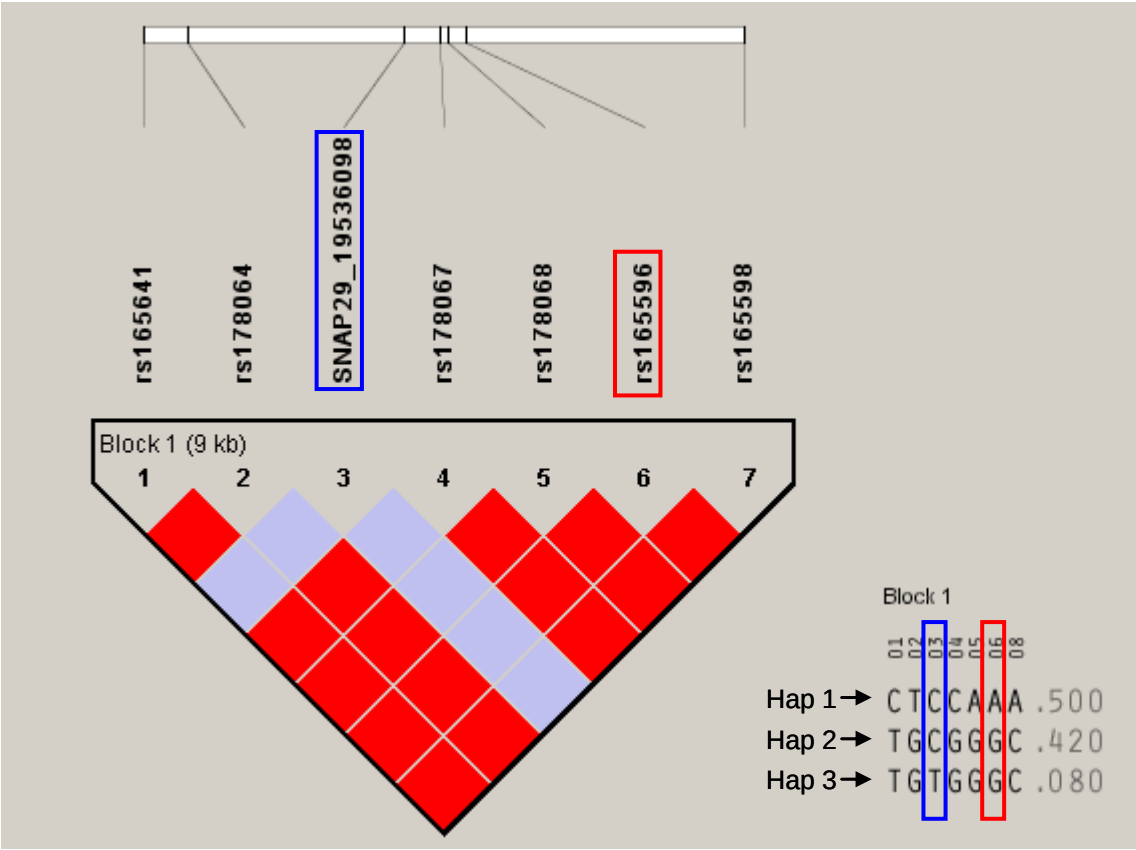


Figure 15. Linkage of the T allele of the novel -1747 *SNAP29* SNP with the G allele of rs165596.

The A/G SNP designated rs165596 has never been tested in a functional assay for effects on promoter activity, nor has an association with an expression phenotype ever been shown. It is therefore possible that this SNP is not causative but is in fact in LD with a functionally active SNP. To test the possibility that this is the case, and that the -1747 C/T SNP is a candidate for the real functional variant, the expression levels of *SNAP29* in the 31 individuals were recovered from the Stranger et al dataset, and the average expression level for each of the three possible genotypes at each SNP plotted (Figure 16). rs165596 was not associated with any change in *SNAP29* expression, whereas -1747 C/T showed a decrease in *SNAP29* expression associated with the rare T allele (Figure 16). This suggests that rs165596 is not the causative SNP in the schizophrenia association, but is in LD with another functional variant. It also suggests that -1747 C/T is a good candidate for that functional variant, and that it may contribute to schizophrenia susceptibility by causing a decrease in *SNAP29* expression. As the T allele is associated with the G allele at rs165596, the

overrepresentation of G alleles in schizophrenic patients may have been caused by its linkage to the T allele.

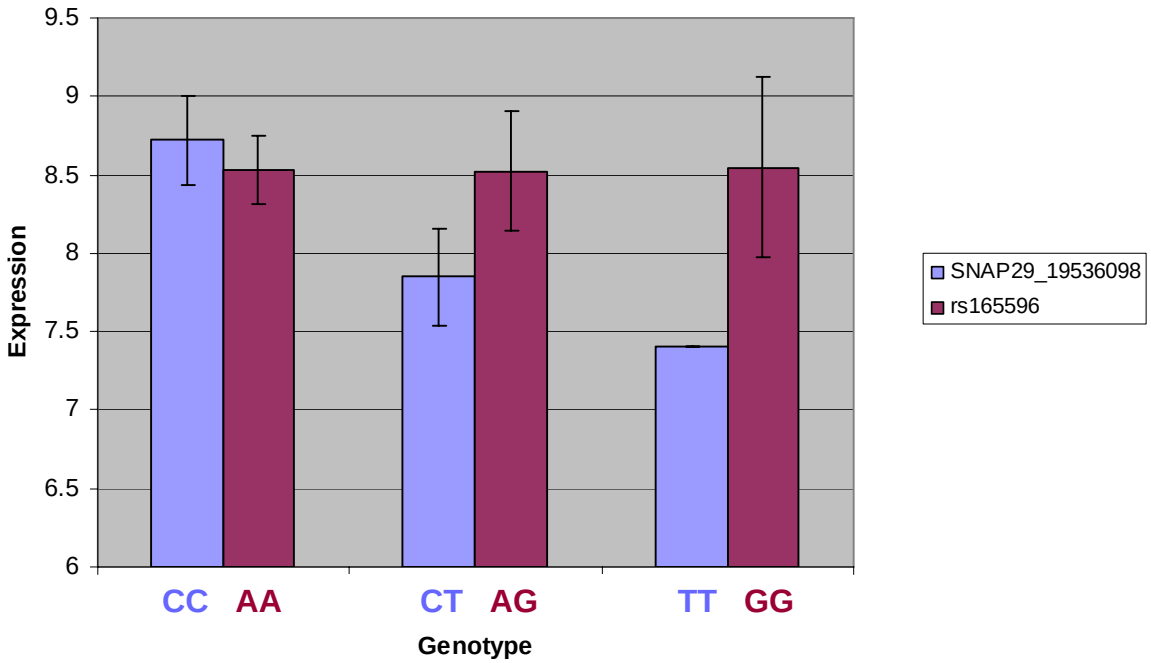


Figure 16. Association of the genotypes at the -1747 *SNAP29* SNP and rs165596 with *SNAP29* expression

3.3 Conclusions

In this chapter, the successful creation of a resource of genotyped promoter SNPs was described. This consisted of 807 SNPs with an estimated minor allele frequency of at least 0.01. The 1187 successfully sequenced amplicons totalled 680,510 bases of sequence. Once overlaps were taken into account, the total sequence coverage was 513,087 base pairs. This gave a SNP ascertainment rate of 1 SNP per 636 bases or 1.57 SNPs per kb. This compares to a rate of 0.93 SNPs per kb for SNPs from genomic clone overlaps in chromosome 22 (Dawson et al. 2001) and 0.52 SNPs per kb for data from the SNP Consortium produced from whole genome shotgun re-sequencing (Sachidanandam et al. 2001). Neither of these two datasets can be used to predict the number of SNPs expected from this study, as the methodologies are very different and unlikely to match the ascertainment of targeted re-sequencing. More recently, the ENCODE consortium has re-sequenced 10 regions of ~500 kb each from subsets of individuals from the HapMap panels. Re-sequencing of 16 unrelated Caucasians from the CEPH families by PCR from diploid samples resulted in an ascertainment rate of 4.86 SNPs per kb, markedly higher than that found for promoters. The difference is likely due to two factors; increased thoroughness of the re-sequencing itself (e.g. repeating of failed PCR and sequencing reactions from individual samples) and the inclusion of intergenic and intronic DNA which is likely to be under less selective constraint, and hence to contain more SNPs than putative regulatory regions such as promoters.

The most valid way to assess the rate of promoter SNP ascertainment is to compare it to other re-sequencing projects using the same number of individuals from the same population. The only major project currently using the same 48-person panel is the Sanger Institute ExoSeq project, which aims to mine exons across the human genome for SNPs by re-sequence. While data from this project has yet to be published, they report a rate of 9.27 SNPs per kb. This is slightly under six times as high as the rate from the promoter re-sequencing. As the ExoSeq project is a long term project with a team dedicated to its completion, they were able to repeat failed PCRs or sequencing reactions, and this would increase the ascertainment rate. Although the aim of the ExoSeq project is to re-sequence exons, their primer design pipeline allows 125 bases of flanking sequence around each exon, thus including a significant amount of intron

sequence. This in fact accounts for a large proportion of the SNPs discovered, and because introns are thought to be under less selective constraint than promoters, this would have driven up the number of SNPs found per kilobase relative to the promoter project. Also, exons are likely to contain far less low complexity sequence than promoters, making them easier to sequence and thus easier for ExoTrace to detect SNPs. A smaller study by T. Eades at the Sanger Institute is using this panel to re-sequence non-coding regions that are highly conserved between humans and mice. This has yielded 54 SNPs from 40 kilobases of sequence, a rate of 1 SNP per 740 bases or 1.35 SNPs per kb, somewhat lower than the rate for promoters. This is more likely due to the pre-selection of conserved sequences that will naturally contain fewer polymorphisms rather than a reflection of the relative SNP ascertainments of the two studies.

The overall minor allele frequency distribution was biased towards rare alleles, in accordance with what is generally expected of SNP distributions under neutral evolutionary conditions (Hartl and Clark 1997; Rockman and Wray 2002). However, there was also a statistically significant bias away from rare alleles compared to what would be expected from this panel. 25% of promoter SNPs had a minor allele frequency of 0.05 or lower, compared with 36% for data produced by the ExoSeq project ($p = 2.45 \times 10^{-11}$). While there are differences in the selective forces to which promoter and exonic SNPs are subject, the difference may again reflect a greater attrition rate in the promoter re-sequencing compared to ExoSeq for the reasons detailed above. 46% of HapMap SNPs had a minor allele frequency of 0.05 or less (Consortium 2005b), but the panel used for that project was far larger, and so the sensitivity to rare SNPs cannot be compared. In summary, the number of SNPs discovered in this promoter re-sequencing project falls short of the potential afforded by the 48-person CEPH panel, and this could have been improved upon by more repeats and optimization of failed PCR and sequencing reactions. Nevertheless, it is significantly higher than ascertainment from large scale SNP discovery projects, and is thus offers an improved resource for studying the functional effects of promoter variation.

Comparison of the distributions of different SNP types revealed no significant difference between promoters and chromosome 22, despite the known lack of

methylation at promoters which would have been proposed to influence the SNP distribution. Analysis of the rooted polymorphisms confirmed that most C/T SNPs are caused by a cytosine mutating to a thymine rather than the reverse, but still failed to show that this process happened any less frequently at promoters than in the rest of the chromosome. Restriction of the analysis to SNPs within 500 bases of the TSS, where lack of methylation is the most marked (Eckhardt et al, unpublished) did reveal a significant excess of C/G mutations, but this is more consistent with elevated GC content than with a methylation-related phenomenon. Indeed, even when only the rooted SNPs in CpG islands were analysed, the expected bias away from C to T mutations does not arise. A significant over-representation of C to A and G to T changes at the expense of A to G and T to C was observed, again consistent with elevated GC content leading to more G and C from which mutations can arise. While there was also an excess of GC SNPs in CpG islands, this fell just short of statistical significance. A possible explanation for these findings is that methyl-cytosine deamination is a relatively ancient process, dating as far back as the onset of DNA methylation in the mammalian lineage. As such, many of the methyl-cytosines in the human genome may have long since mutated to thymine and become the dominant alleles if not becoming completely fixed. As the number of CpG dinucleotides remaining in the human genome is relatively low (only 20% of the level expected), the rate of C/T SNP generation by methyl-cytosine deamination may have dropped significantly over evolutionary time. The lack of a bias away from these mutations in promoters may therefore reflect a corresponding drop in the rate of methyl-cytosine deamination in the wider genome, rather than signifying that promoters are methylated.

16.1% of the promoter SNPs in this study were found within putative regulatory elements. The precise figure is probably not meaningful, as the overall total was greatly influenced by the two TFBS databases, and the number of these elements found varies greatly with the parameters used. More importantly, there was no significant under-representation of SNPs in these elements overall. Such a bias might have been expected if the majority of these elements represented real functional sites that might be susceptible to purifying selection. Examination of individual categories showed only one with fewer SNPs than would be expected given the base coverage of the elements. However, this was the phastcons category, which is highly conserved by

definition and therefore almost certain to contain fewer SNPs regardless of any functional implications. Given the equal distributions of SNPs between these elements and promoters overall, there is no sign from the SNP data alone that these elements are predictive of functional SNPs *a priori*.

The lack of association between promoter SNPs and expression phenotypes as determined by Stranger et al was disappointing, although not entirely unexpected given the relative lack of power of the small overlapping set of individuals. The single SNP that was associated, located in the promoter of the *SNAP29* gene, did potentially shed new light on the mechanistic basis for an observed association with schizophrenia, and suggested that the C/T SNP at -1747 from the *SNAP29* TSS is a more likely candidate as the causative variant than the previously published A/G SNP, rs165596. This is not conclusive however, and further work is needed to demonstrate this more rigorously. An easy way to increase the power of the association is to genotype both the published A/G SNP and the -1747 C/T SNP in the remaining HapMap individuals and repeat the association using the expression data now available. Interestingly, the previously published association was found in Europeans but not in Africans, although rs165596 is common in both populations. However, -1747 C/T was rare in the panel tested here, suggesting that it may be a relatively recent lineage specific mutation. If -1747 C/T is absent in African populations (a question that could also be answered by typing the entire HapMap panel including the Yoruban population), this would be further evidence for its case as the causative mutation. Eventually, it would be necessary to carry out a case/control study with a panel of schizophrenia patients and controls, and see whether the T allele is overrepresented in affected individuals. The Sanger Institute has recently obtained a set of DNA samples from schizophrenia patients, so in fact this study may be easily achievable subject to time and resources.

4 High-throughput cloning and reporter assays on a promoter haplotype library

4.1 Introduction

In chapter 3 the discovery and genotyping of promoter SNPs from chromosome 22 by re-sequencing was described. In this chapter, the aim was to use this SNP resource as a tool to study the role of natural sequence variation on the level of activity of promoters. This required the isolation of individual promoter haplotypes from their diploid partners, and the measurement of the activity of each haplotype. SNPs that are found to alter promoter activity could then be examined for characteristics that could distinguish them from those that are found to be functionally neutral. There are two overall strategies for studying the effect of promoter polymorphism on gene expression. The first method is to somehow assay allele-specific expression in heterozygous individuals or cell lines *in vivo*, isolating the haplotypes by measuring them separately rather than by physically separating them into different assays. This can either be by differentiating between allelic transcripts using a transcribed SNP as a marker (Pastinen, Ge, and Hudson 2006), or by quantitatively assaying RNA Pol II loading on the promoters using the haploChIP method (Knight et al. 2003). The second approach is to clone individual promoter fragments carrying different alleles into a reporter plasmid (luciferase being the reporter of choice) followed by assays for that reporter in transiently-transfected cell lines (see chapter 3). The main advantage of using heterozygotes *in vivo* is that any effects discovered are more biologically relevant, as the two promoter variants are in their native chromatin contexts and exposed to identical TF backgrounds. However, the disadvantage is that the range of variation that can be tested is dependent on the number of different heterozygotes that can be found for a given promoter (and the presence of suitable markers in the case of allelic transcript assays). This will vary depending on the population history of the DNA sequence under study, and hence on the frequencies of the SNPs present and the extent of linkage disequilibrium. It may be very difficult to isolate individual SNPs from other variants on the promoter, or from polymorphisms in distant regulatory elements such as enhancers, making it potentially difficult to identify the relative importance of each polymorphism to any functional variation discovered. In contrast, cloning promoter fragments into an *in vitro* system allows the effect of promoter sequence variation to be studied in the absence of other *in vivo* regulatory inputs that may confound such effects. Indeed, positive or negative inputs from upstream regulators or chromatin may exert so much influence that they would mask subtle

effects of regulatory SNPs. It also enables each promoter haplotype to be tested in isolation, eliminating the need for heterozygotes and making it relatively easy to test every available haplotype, and even to mutate the promoter *in vitro*. However, the degree to which *in vitro* findings translate into real biological phenotypes is difficult to determine. Because the effect of promoter variation is highly context-dependent, it would be an impossible task to assay every possible combination of *in vivo* conditions in which it could be found. Despite these caveats, there is plenty of evidence to suggest that *in vitro* promoter studies often do translate to an *in vivo* effect (Rockman and Wray 2002), and that cloned promoter fragments contain many of the elements that lead to regulated function *in vivo* (Cooper et al. 2006).

In this project, the *in vitro* reporter assay approach was used to test a subset of the promoter SNPs discovered in chapter 3 for functional effects. The classical strategy for cloning the different haplotypes is to identify individuals homozygous for each one, amplify the promoter by PCR from each individual and clone it directly into a reporter plasmid either using PCR primers containing restriction enzyme sites or by blunt-end or TA cloning. This requires the ready availability of homozygous individuals or a separate round of cloning and sequencing of clones for each heterozygote in order to separate the two haplotypes, and is very labour- and time-intensive. Instead, a novel high-throughput cloning strategy was developed for this project that enables the cloning of a large number of promoter haplotypes in parallel, and takes advantage of the large sequencing capacity available at the Sanger Institute. Rather than attempting to isolate each haplotype at the beginning of the process by choosing the PCR template and by cloning of single heterozytoes, all haplotypes are amplified and cloned simultaneously in one batch, and the clones are separated at the end by screening clones from the resulting libraries. The method implements the Gateway cloning technology by Invitrogen that uses modified enzymes from the bacteriophage λ recombination system to move fragments directly between plasmids, without the need for restriction enzyme digestion, insert purification and re-ligation. This not only cuts down on the time needed for each reaction, but nearly eliminates much of the insert loss observed during more conventional cloning of promoter fragments, both in preliminary experiments for this project and by other labs (Buckland et al. 2005). The degree to which steps in the procedure need to be repeated is thus greatly reduced. To create libraries of cloned haplotypes, the strategy

involved the PCR of promoters from a mixed pool of DNA fragments representing the haplotypes to be cloned. These would then be cloned into a holding vector using Gateway, creating a resource of plasmid mixes for long-term storage. The mixes would be recombined into a luciferase reporter plasmid by Gateway cloning, and libraries of clones would be screened by sequencing to find haplotypes for functional testing in a luciferase assay.

The Gateway system is based on the use of the enzymes from the bacteriophage λ recombination system. These enzymes are responsible for integrating the λ DNA into the genome of *E. coli*, and facilitate the switch between lytic and lysogenic life cycles. The λ genome contains genes that code for two recombinases, λ integrase (Int) and λ Int and Excisionase (Xis), which catalyse the integration and excision of the bacteriophage along with *E. coli*-coded cofactors (Landy 1989; Ptashne 1992). During integration, Int causes recombination between the circular viral DNA and the *E. coli* genome at specific attachment sites (att sites) on both molecules. These are the attB (*E. coli*) and attP (bacteriophage λ) respectively (Weisberg and Landy 1983; Landy 1989). While they are not identical by sequence, they share a 15 base pair motif where recombination occurs. The result is the integration of the λ genome, and the generation of two different att sites at each end of the integrated λ called attL and attR. In order for lambda to excise, Xis reverses the integration process by catalyzing recombination between the attL and attR sites, resulting in the original attP-containing λ virus and the attB site in the *E. coli* genome. In Gateway cloning, the inserts to be cloned are flanked by two attB sites, and a modified Int enzyme and associated cofactors (BP clonase) causes recombination with a pair of attP sites in the target vector (Figure 17). The core sequences where recombination occurs are different between the two attB and attP sites, ensuring that each site can only recombine with its intended partner. Gateway cloning is thus directional. The insert is now in a plasmid and flanked by attL sites generated during the recombination (Figure 17). In order to transfer the insert to a target vector, that vector must contain a pair of attR sites. In the presence of the holding vector and the target vector, a modified Xis enzyme will catalyse a recombination event between the attL sites in the source vector and the attR sites in the destination vector. The insert is thus shuttled into the target vector, and the DNA between the attR sites in the target vector is moved into the source vector.

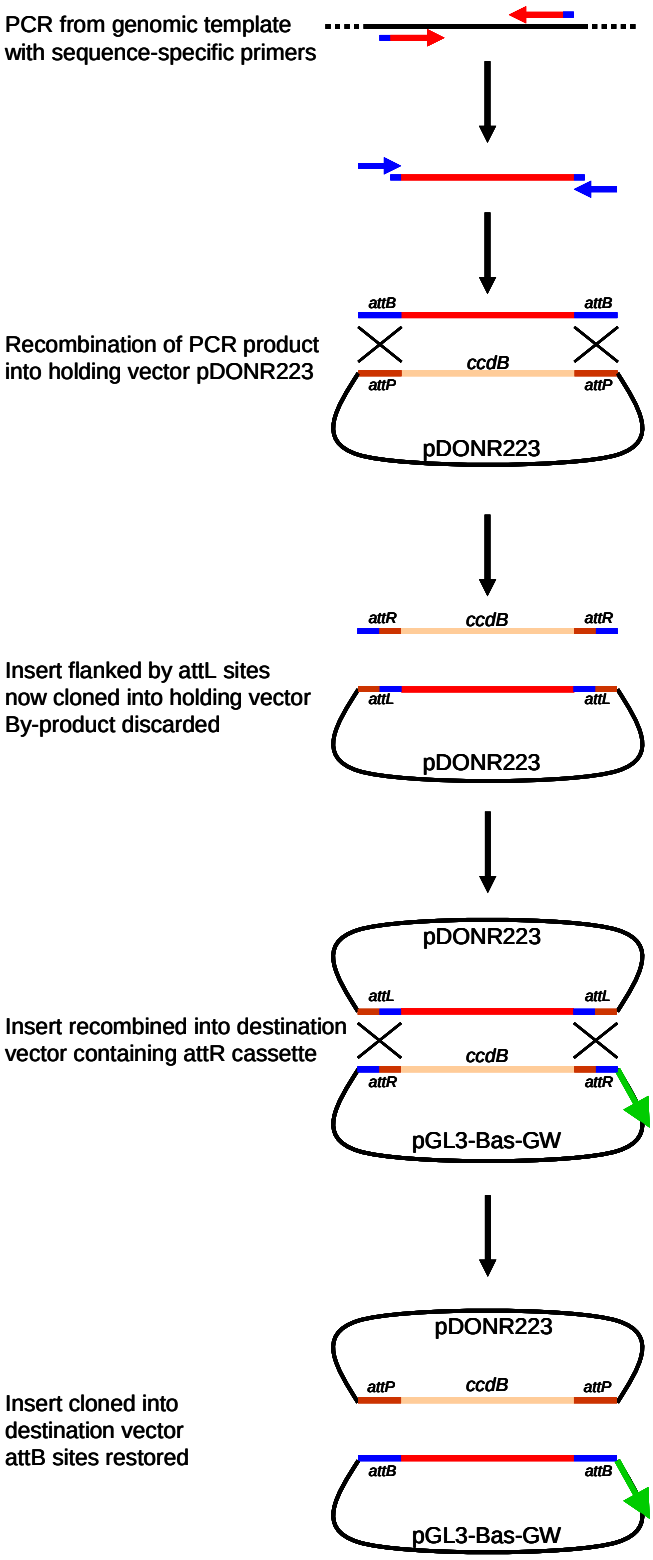


Figure 17. Cloning a PCR fragment using the Gateway cloning technology by Invitrogen.

In addition to the recombination mechanism, the other significant part of the Gateway technology is the selection system. Plasmids designed to receive inserts in a BP

reaction contain a cassette between the two attP sites that contains the ccdB gene. The ccdB gene product halts the growth of most *E. coli* strains by disrupting *E. coli* DNA topoisomerase II (Bernard and Couturier 1992). This acts as a negative selection marker, so that when a recombination reaction containing the insert and recipient plasmid are transformed into *E. coli*, those cells that take up unrecombined plasmid do not grow, meaning that only plasmids that have successfully received the insert and thus discarded the ccdB gene form colonies. In order to recombine the insert into a destination plasmid using LR clonase, that plasmid must have two characteristics; it must be made Gateway-compatible by cloning in a ccdB-containing cassette flanked by attR sites, and it must also carry a different antibiotic resistance gene to the one on the donor plasmid. The presence of dual selection markers ensures that only *E. coli* transformed with the recombined destination vector form colonies. Both unrecombined and recombined donor plasmids will be selected against by antibiotic and unrecombined recipient plasmids by ccdB.

Reporter assays on variant promoters are capable of detecting sequence-dependent functional variation on two levels. Individual promoter polymorphisms can each have an effect on promoter activity, or multiple SNPs can act synergistically to cause a functional difference between haplotypes. Where the line is drawn between these two factors depends somewhat on the sensitivity of the assay being used (i.e. SNPs that seem to act synergistically but show no effect individually may be escaping detection because their individual effects exist but are below the sensitivity of the assay). In order to be able to resolve the action of individual polymorphisms, it is necessary to maximise the number of combinations of alleles tested. Ideally the study of a polymorphic promoter would test every possible combination. However, this would almost certainly require extensive *in vitro* mutagenesis, as linkage disequilibrium across the promoter would make it unlikely that all possible combinations would be found in a natural population, particularly in promoters with 3 or more polymorphisms. While this may be possible in a study of one or two promoters, it is prohibitive when dealing with many promoters, as is the case in this project. It was therefore necessary to rely on the haplotypes present in the panel of individuals, and to try and clone as many of them as possible into a reporter vector, hence the importance of a robust high-throughput cloning strategy. The relatively deep re-

sequencing to generate the SNP panel also helped maximise the combinations of alleles.

In this project, all cloned haplotypes were tested independently on a set of four transformed human cell lines; HT1080, TE671, HEK293FT and HeLa. These are derived from fibrosarcoma, medulloblastoma, embryonic kidney and cervical carcinomas respectively and thus represent a range of human tissues. Because of the context-dependence of the functionality of promoter polymorphism, a broad range of cell types was chosen to maximize the number of functional SNPs discovered. All of these lines have been previously used in reporter assays (Hoogendoorn et al. 2003; Trinklein et al. 2003; Buckland et al. 2005; Kim et al. 2005a; Cooper et al. 2006), and have proved to be amenable to transient transfection with a range of commercially available reagents. HEK293FT and TE671 have been used previously in large scale studies of promoter variation, and revealed that 26% of functional SNPs in promoters active in both cell lines had cell-specific effects on promoter activity (Buckland et al. 2005).

4.2 Results

4.2.1 Experimental strategy

The two distinguishing and novel features of the cloning strategy used in this study are the cloning of mixed pools of inserts followed by recovery of clones by sequencing a clone library, and the use of Gateway cloning technology rather than conventional cloning. Instead of cloning each predicted haplotype individually by searching for homozygotes, pools of DNA samples from multiple individuals were created with each haplotype being represented by at least one chromosome in a diploid DNA sample (Figure 18). These pools were used as templates for PCR reactions using primer pairs with sequence-specific 3' ends of ~20 bases and 5' linker sequences containing part of the attB1 and attB2 sites. This was followed by a second round of PCR using universal primers to the linker region of the first round primers, and containing the remainder of the attB sites. Thus the two-round PCR for each promoter produced a mixture of products amplified from the different samples in the template pool. These were cloned into the Gateway-compatible plasmid pDONR223, yielding mixtures of plasmids containing each haplotype amplified from the PCR. pDONR223 is essentially a holding vector and does not contain a reporter gene, instead functioning as way to store the haplotype libraries in a form that could be easily and rapidly cloned as needed. The pGL3 Basic promoter-less luciferase reporter plasmid was modified to make it Gateway-compatible by inserting a cassette containing the ccdB gene flanked by attR sites. The promoter haplotypes in each pDONR223 mix were transferred to the modified reporter plasmid with LR clonase. Libraries of colonies were made using the resulting clone mix, and this was screened by PCR and sequencing of the inserts to identify which clones contained which haplotypes.

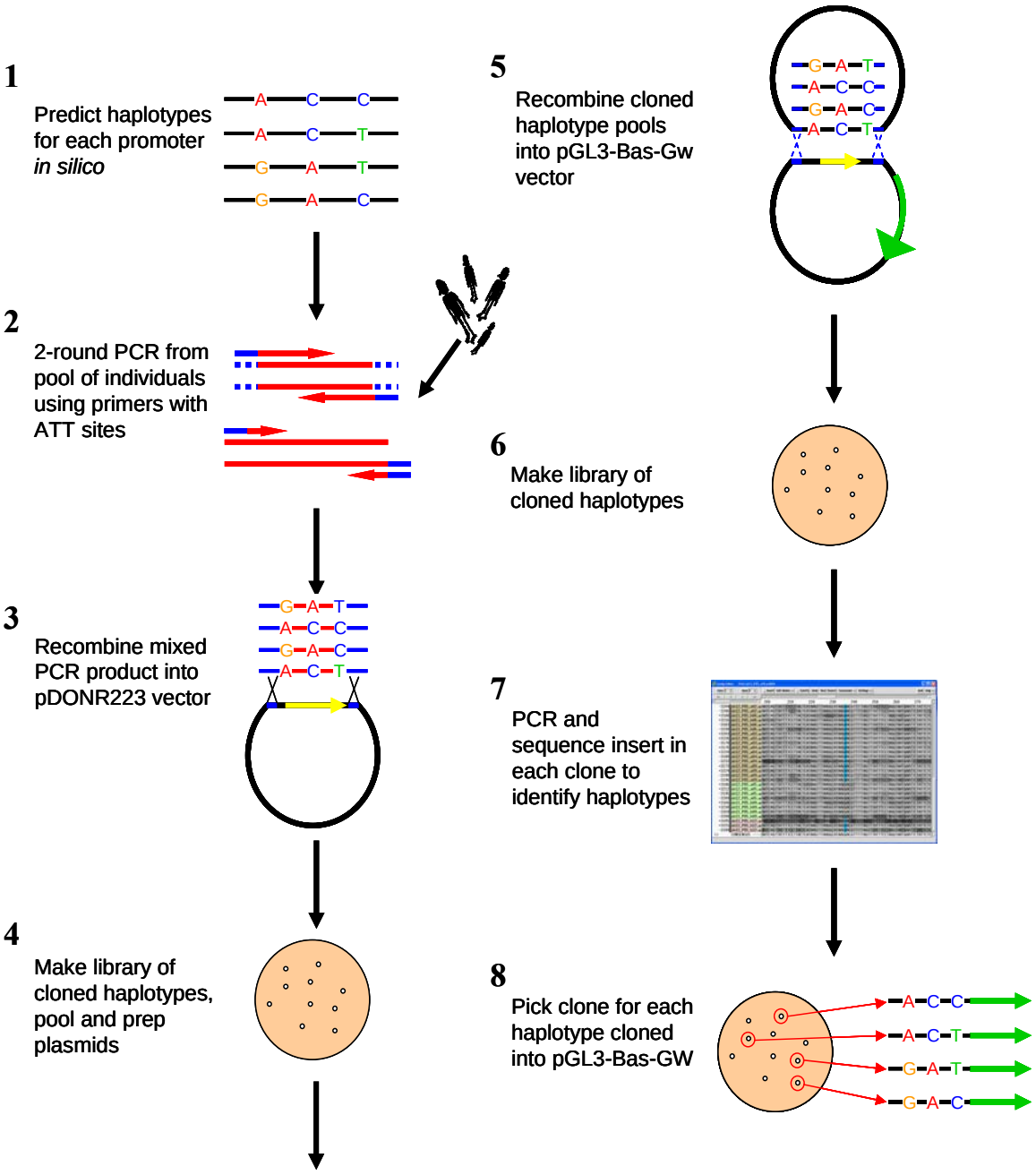


Figure 18. High-throughput strategy for cloning promoter haplotypes into luciferase reporter vectors using Gateway technology.

The choice of firefly luciferase, and particularly the pGL3 series from Promega, as the reporter to use in this project was mainly due to its sensitivity, large linear dynamic range and proven suitability for quantitative signal determination (Buckland et al. 2005). Luciferase expression driven by the cloned promoter fragments was assayed using Promega's Dual Luciferase reporter assay system, enabling direct comparisons between the expression levels from different reporter constructs. The system is based on the use of two reporter plasmids. The first plasmid is the pGL3 Basic plasmid

described above, with a luciferase cloned from the firefly *Photinus pyralis*, and into which the promoter haplotypes to be tested have been cloned. The second plasmid contains an active promoter, such as SV40 or other viral promoter, and constitutively expresses a second reporter. This is another luciferase, this time from the sea pansy *Renilla reniformis*. The two luciferases have very similar optical spectra, but require chemically distinct substrates. This allows the signal for each luciferase to be measured independently in the same well of a microtitre plate by the addition of the appropriate substrates and quenching reagents. The co-transfection of each pGL3-cloned promoter haplotype with the same pRL control plasmid enables internal normalisation for experimental variables such as transfection efficiency variation, and allows the signals from each haplotype to be compared directly.

4.2.2 Modification of pGL3-Basic to confer compatibility with Gateway technology

Before the promoter variation can be tested experimentally, the luciferase reporter vector pGL3 Basic must be made compatible with the Gateway system. This involved the insertion of an acceptor cassette into the multi-cloning site of the vector (Figure 19). This cassette is available from Invitrogen in 3 different reading frames for use in cases where the protein product will be expressed. In this case, the frame is not relevant, as promoters are not restricted to a particular frame relative to the coding sequence (the frame is set by the translation start site, not the transcription start site (TSS)). The cassette used was the RfC.1 version. It contains the ccdB gene for post-recombination negative selection as well as a chloramphenicol resistance gene to enable selection of modified plasmids once the cassette is cloned in. The cassette was blunt-end cloned into the *Sac*I site of pGL3-Basic by digesting and gel-purifying the plasmid, removing terminal phosphates and ligating in the cassette, which is provided with terminal phosphates to facilitate ligation. The ligations were transformed into JM109 competent cells and selected with cloramphenicol on LB agar plates. The ccdB gene was not toxic to JM109 because this E. coli strain carries the F episome. This contains the ccdA gene, which counteracts ccdB and thus allows the plasmid to grow where it would otherwise be negatively selected in strains without the F episome.

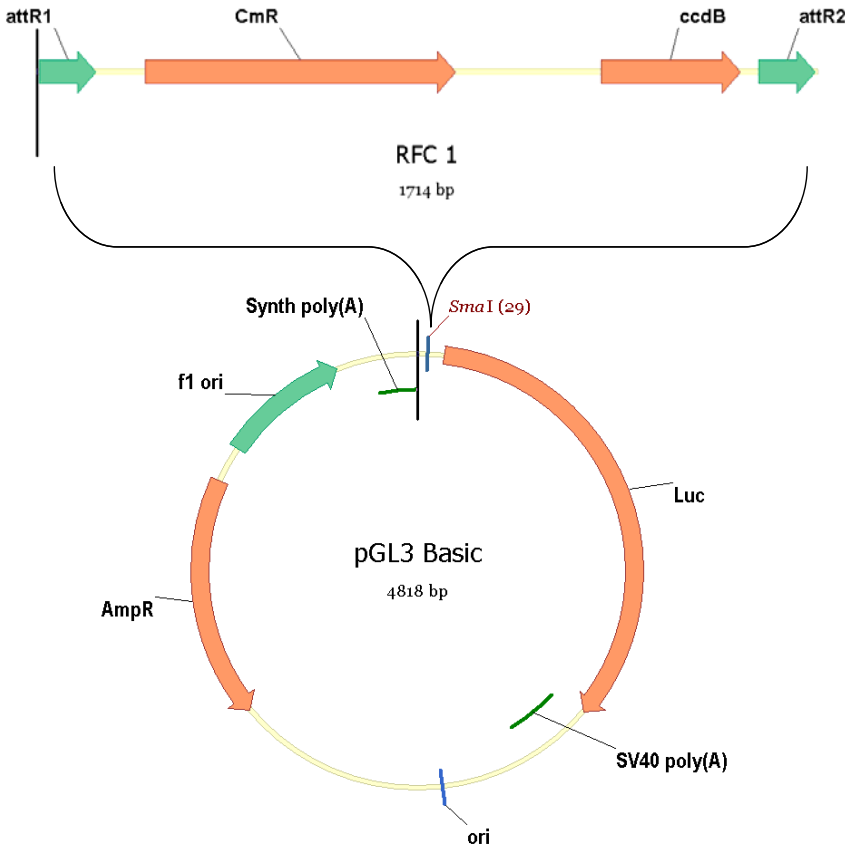


Figure 19. Modification of the pGL3 Basic reporter vector by the insertion of a Gateway acceptor cassette into the multi-cloning site (MCS). The MCS contains a *Sma*I restriction site that leaves blunt ends when digested.

Because Gateway cloning is a directional process, it is important that the cassette is inserted in the correct orientation, and thus that the two attR sites are correctly position relative to each other. Use of a plasmid with the incorrect orientation would result in the promoter being cloned in the wrong direction, leading to no reporter expression. Several clones were screened for insert orientation by carrying out colony PCR across the insertion site and end-sequencing the products using the pGL3-specific sequencing primers RVprimer3 (CTAGCAAAATAGGCTGTCCC) and GLprimer2 (CTTTATGTTTTTGGCGTCTTCCA). These were designed to amplify and/or sequence across the multi-cloning site of the pGL3 series of vectors. The two attR sites on the cassette differ by one nucleotide; if the cassette is cloned correctly the 5' end attR site should contain a run of 6 adenines, whereas in the 3' end that run is interrupted by a cytosine at third position. Clones containing the cassette in both the forward and reverse orientations were identified and one of each was successfully prepared from cultures of a single colony. While the plasmid containing the cassette in reverse was not used in this project, it was prepared due to its potential use in

investigating the bi-directionality of promoters or as a means of preparing negative controls for unidirectional promoters.

The process of cloning promoter fragments into the modified pGL3 vector (or indeed any Gateway-compatible plasmid) results in 169 bases from the ATT sites being present between the 3' end of the cloned fragment and the translation start site of the luciferase gene. This raised the possibility that the ability of cloned promoters to drive expression of the reporter may have been abrogated. Several test promoter fragments from the cloned set were amplified from standard genomic DNA and cloned into the modified Gateway vector, and were shown to be able to drive significant luciferase expression in HeLa cells (data not shown).

4.2.3 Selection of target fragments for cloning and functional testing

Despite the many papers investigating specific promoters for functional polymorphisms, attempts to clone and analyse promoter haplotypes in large numbers using classical restriction enzyme based methods or TA cloning have generally had high attrition rates (Buckland et al. 2005). My initial attempts to clone haplotypes from the highly polymorphic 2kb of the PDGFB promoter repeatedly failed with few clones being produced despite a wide range of methods and conditions attempted. Colony PCR of these colonies showed that they often contained a variety of insert sizes (and frequently no insert at all) despite the use of a single PCR product of defined size in the cloning reactions. This implied an inherent tendency in the PDGFB promoter for rearrangement and deletion when cloned, even when the recA$^-$ *E. coli* strain XL-10 Gold was used to minimise this. The reduction of the target fragment size only marginally improved the success rate. These results were replicated in a larger set of 10 promoters, with several attempts being required to obtain even one correctly-cloned insert. Other groups have also had problems with high-throughput cloning using various cloning methods such as TA cloning (Buckland et al. 2005). The extent of this phenomenon varies between promoters, but when it occurs it can require extensive optimization of the cloning strategy, and effectively precludes the study of those promoters in a high-throughput pipeline.

The high-throughput Gateway method developed improved the yield of successful clones by between 2- and 4-fold for the full-length clones, but they were still prone to rearrangements and insert-less clones. However, the smaller the promoter fragments cloned, the better the efficiency became, and ~500 base pair fragments were 100% successful in the clones tested (Figure 20).

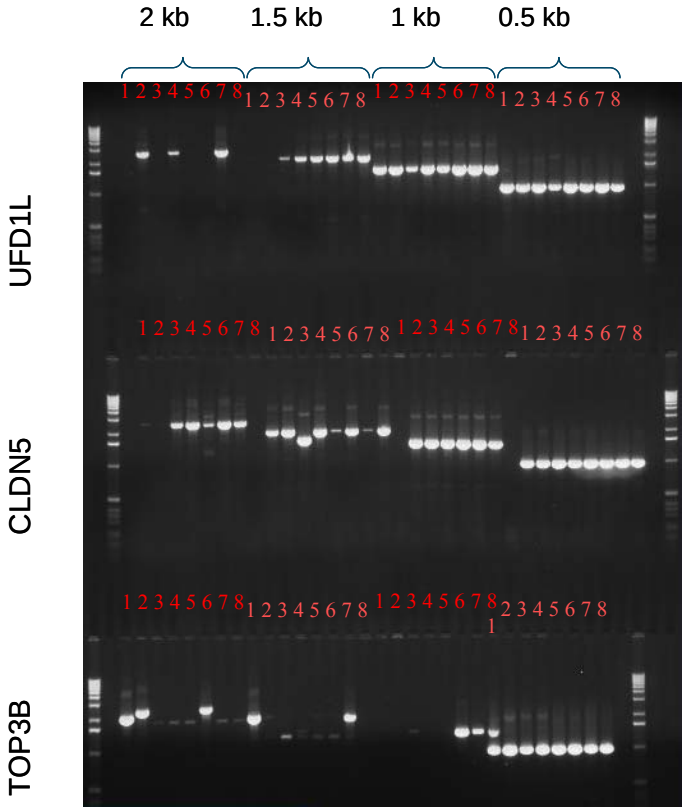


Figure 20. Cloning of four different promoter fragment sizes from the UF1DL, CLDN5 and TOP3B promoters using the high-throughput Gateway method. The fragment sizes tested extended for approximately 2 kb, 1.5 kb, 1 kb and 0.5kb upstream of the annotated TSS. The fragments were amplified for cloning by using the appropriate combinations of the 5' and 3' primers from the primer pairs used for the re-sequencing. Each promoter was cloned using the Gateway method into pGL3-Basic-GW. 8 clones per fragment were screened for insert presence and integrity using colony PCR across the insertion site, and the PCR products run on a 1% agarose gel. The performance of the cloning method increases significantly with decreasing fragment size, with the 0.5 kb fragments having 100% success in this test. Note that lane 1 of the 0.5 kb fragment of the TOP3B promoter was mis-loaded on the gel into the same well as lane 8 of the 1 kb fragment.

For the purposes of functionally testing promoter SNPs, it was decided that only the proximal ~500 base pair fragments would be targeted. This decision was motivated by the highly increased efficiency of cloning the small fragments relative to the full 2kb ones, as the strategy proposed here fundamentally relies on the ability to generate and sequence large numbers of clones containing variants of otherwise identical inserts. While a large number of SNPs would not be tested in this approach, it was likely that

many functional variants would be close to the TSS. Rockman and Wray surveyed functional promoter polymorphisms in the literature up to the end of 2001. A histogram of the positions of the SNPs they described, plotted from the data in their paper (Rockman and Wray 2002), showed a prominent peak centred in the first 100 bases upstream of the TSS and trailing away until around 500 bases upstream (Figure 21). While ascertainment and publication bias may be a significant factor in producing this peak, it demonstrates that there is ample functional variation to be found in these regions. Another consideration was the then-unpublished observations by Cooper that the -500 to -1000 bases relative to the TSS often contained negative regulatory elements (Cooper et al. 2006), and where this was the case this might have suppressed promoter activity *in vitro*, and possibly masked the action of more proximal SNPs.

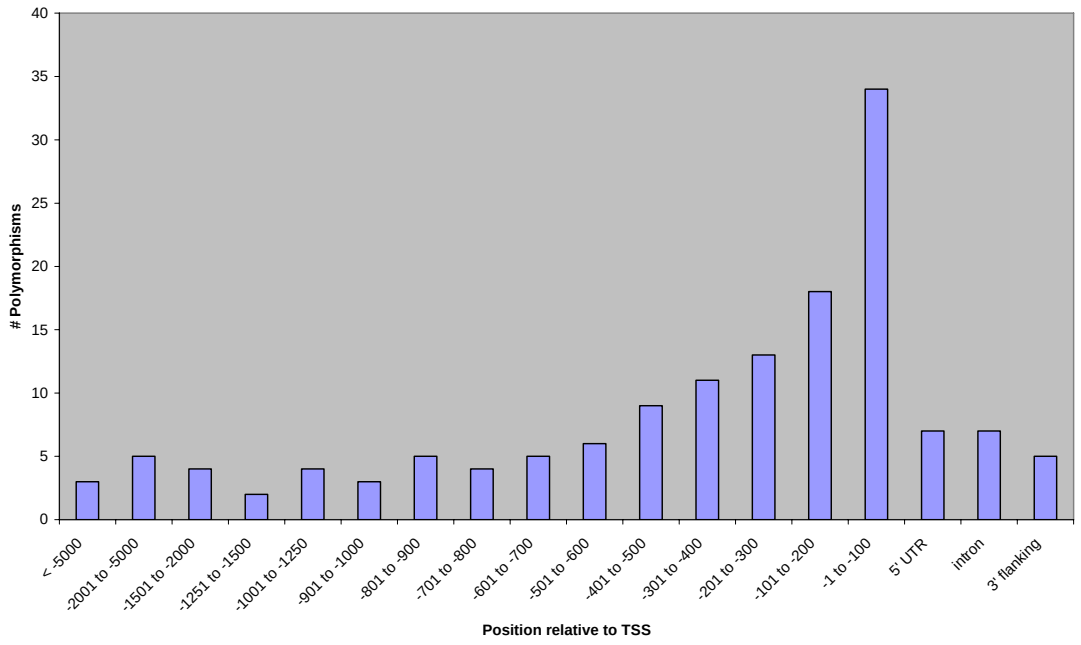


Figure 21. Profile of the numbers of experimentally verified promoter polymorphisms present in the literature. Data were taken from the supplementary material of Rockman and Wray 2002.

4.2.4 Prediction of promoter haplotypes

In order to select the appropriate DNA samples to construct template pools for promoter fragment PCRs, it is necessary to know the haplotypes present in each

individual. However, the promoter SNP resource described in Chapter 3 consists of unphased genotype data, so the haplotypes present had to be inferred from the genotypes. There are several programs designed to do this with different methods, including LCZC (Lin et al. 2002), HAPLOTYPER (Niu et al. 2002), HaploView (Barrett et al. 2005) and Phase 2.1 (Stephens, Smith, and Donnelly 2001; Stephens and Scheet 2005). The latter was chosen for this study due to its superior performance compared to LCZC and HAPLOTYPER and the suitability of the program to automation by scripting, which was not possible with HaploView.

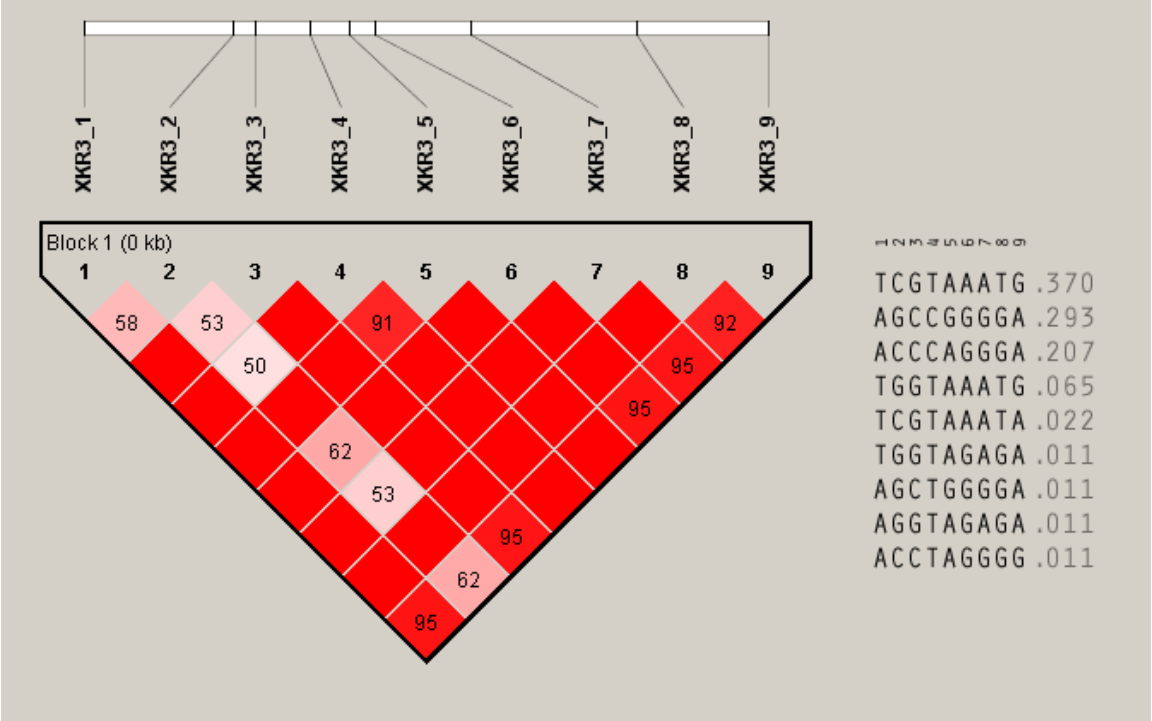


Figure 22. Prediction of haplotypes in the XKR3 promoter using the genotypes produced in the re-sequencing. This was the promoter with the highest number of SNPs in the fragment targeted for cloning, with 9 SNPs and 9 haplotypes. The coloured boxes between each SNP pair are a measure of the degree of linkage disequilibrium between them. The shade of red used is an indication of the D' measure for that SNP pair, with deeper shades signifying higher D'. The numbers in the boxes are the D' scores represented as a percentage, and empty boxes denote a D' of 1 (or 100 in this representation). The haplotypes predicted are shown on the right, along with the frequency of each haplotype in the population tested. This figure was plotted using HaploView for visualisation purposes, but all haplotype predictions were done using PHASE 2.1. In this case, the predicted haplotypes and frequencies are the same.

The genotypes for each SNP called from the re-sequencing were extracted from the ExoTrace-aligned contigs using a custom perl script. This called a second script written by Steven Leonard to interrogate the contigs, and then parsed the genotypes by promoter and wrote them in a format ready for Phase analysis. The fragment to be

cloned for each promoter was the same as the 3'-most of the four tiled fragments for which primers were designed in the re-sequencing. SNPs that fell outside these fragments were excluded from the analysis. This means that some promoters containing polymorphisms were not tested for functionality, as the polymorphisms fell outside the regions targeted for cloning. 127 promoters contained polymorphisms in the target regions. Analysis of these sequences revealed a total of 247 SNPs in 359 haplotypes (Figure 23). However, after the completion of the cloning stage of this project, it was subsequently discovered that not all SNPs had been mapped to the correct genome positions. This was due to a computational error in the Sanger Institute SNP database that was beyond my control. This resulted in some promoters appearing non-polymorphic in the target fragment because the SNPs had been incorrectly mapped to the fragment immediately upstream. Thus, only 109 promoters were selected for haplotype cloning and functional testing.

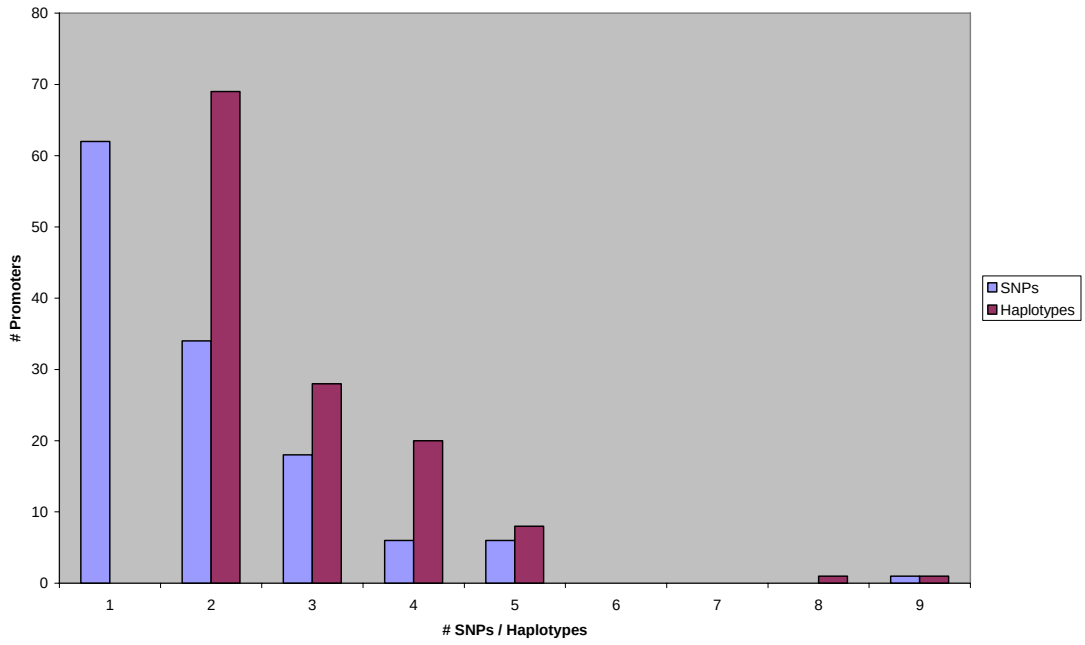


Figure 23. The numbers of SNPs and haplotypes present in the promoter fragments targeted for cloning and functional testing.

4.2.5 Construction of DNA pools and PCR of promoter fragments

The distribution of predicted haplotypes among the individuals was examined by eye for each promoter, in order to find the smallest set of DNA samples that would contain at least one of each haplotype. The aim was to have as close to an equal

representation of every haplotype in the pool as possible given the genotypes present, thus equalising the probability of recovering haplotypes from the pool that are common or rare in the population. Samples with incomplete genotypes from the re-sequencing, and thus with genotypes inferred by Phase 2.1 rather than experimentally confirmed, were avoided. Homozygous samples were chosen in preference to heterozygous ones where possible, in order to minimize the possibility that a heterozygote was miscalled by ExoTrace.

The resulting pools were used as PCR templates to create the mixed promoter inserts, with the first round amplified using sequence-specific primers carrying a short adapter sequence at the 5' end, and the second round with universal primers covering the 3' ends of the attB recombination sites (Figure 24). PCR was carried out using KOD polymerase. This is a proof-reading polymerase with a very low rate of error compared to standard polymerases such as Taq, helping to minimise the possibility of false SNPs being introduced into clones as a result of polymerase error.

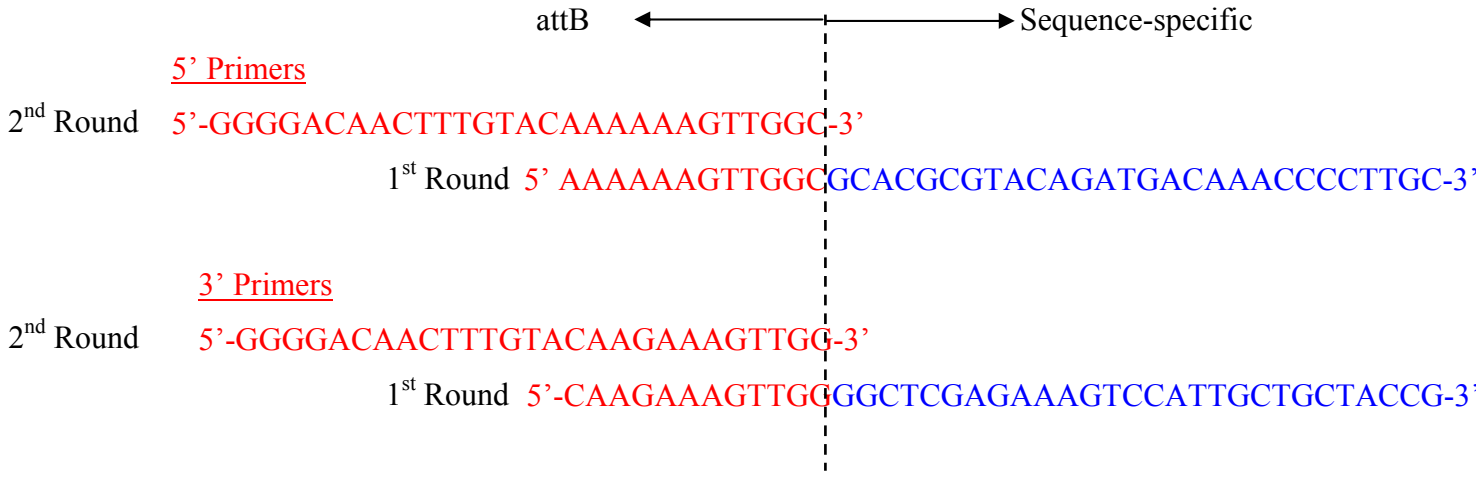


Figure 24. Primer design strategy for inserting attB sites upstream of promoter fragments by 2-round PCR. The first round primers contained a ~20-mer sequence-specific 3' section (blue), followed by linkers at the 5' end (red). The second round primers were universal and designed to anneal to the linker sequences.

The PCR reactions were run on 1% agarose gels, and the promoter fragments excised from the gel and purified using Qiagen's gel extraction kit. All PCRs were successful and produced fragments of the expected size. This was expected as polymorphisms

could only have been found during the re-sequencing if these fragments were amenable to PCR.

4.2.6 Creation of haplotype libraries

Each insert pool DNA was recombined into the kanamycin-resistant pDONR223 plasmid using BP clonase, and the recombination products transformed into DH5α cells. The transformed libraries were plated on to kanamycin-containing agar plates overlaid with nylon membranes. Colonies were harvested by scraping them into LB broth and pelleting the cells in a centrifuge. DNA was prepared directly from the pooled colonies. While the number of colonies produced for each library varied, almost all produced a minimum of several hundred colonies. Only one library (CRYBA4) failed to produce significant numbers of colonies despite repeated attempts.

The plasmid preps from these libraries now contained a mixture of inserts representing each haplotype in the original PCR template pool. This insert mix was cloned into pGL3-Bas-GW using LR clonase, and the products again transformed in DH5α cells and selected with ampicillin. These were plated on LB agar plates to produce libraries of promoter haplotypes in the luciferase reporter plasmid.

4.2.7 Screening haplotype libraries by sequencing

Colonies from each promoter library were screened by carrying out colony PCR across the insert site and sequencing the PCR product. The colonies were picked and cultured overnight in order to prepare glycerol stocks in 96 well plates for long term storage and the templates for the PCR. The number of colonies to be picked for each library was determined according to the following formula…

$$\text{Number of colonies} = \ln(1-x) / \ln(1-y)$$

… where x is the probability of finding at least one clone containing a haplotype of abundance y in the original pool of DNA samples. This assumes that the progress of each haplotype in a library is purely a function of their starting proportions in the PCR

template pool. For each library, the number of clones was calculated for a 98% probability of finding the least abundant haplotype in each pool. 1.4 times this number of colonies was picked for each library, in order to allow for failures in PCR and sequencing of the clones during screening. 10 promoters failed to produce as many colonies as required by these criteria. Of these, all colonies were picked from 6 of them, with the remaining 4 being discarded as they only produced 6 colonies or less and were regarded as having failed at the LR cloning stage.

	Promoters passed		**Promoters failed**	
	Number	**Percent total**	**Number**	**Percent total**
Total	109	100	-	-
PCR	109	100	-	-
BP library	108	99.1	1	0.9
LR library	102	93.6	6	5.5
Clone integrity	84	77.1	18	16.5

The PCR products were sequenced with 4 sequencing primers; 2 insert-specific primers identical to the ones used in the first round PCR, and RVPrimer3 and GLPrimer2. This increased the coverage of each sequenced product and also allowed for confirmation of insert orientation by comparing the sense and antisense sequences from each pair of primers.

1413 colonies from 102 promoters in total were sequenced. Promoters where at least two distinct haplotypes were confirmed by sequencing and cloned in the correct orientation were taken forward to the functional experiment stage.

4.2.8 Reasons for attrition at each cloning step

Due to time constraints, the causes of promoters failing during the cloning process were not investigated in detail, and only limited optimization was attempted on any failures (such as modifying the ratios in cloning reactions). A moderate level of attrition was considered acceptable given the stated aim of developing a high-throughput cloning strategy. 18 promoters were discarded from the final set due to a lack of sequence confirmation of the haplotypes.

During the construction of the haplotype libraries, colleagues in the lab uncovered a problem with the system that compromised the complete directionality of the cloning process. It emerged that the two ATT primers were not sufficiently different to each other to avoid occasional mispriming, and that a subset of the products of the 1st round of PCR would either have been primed with two of the same primer, or with the primers reversed relative to the target insert. In the former case, the products would fail to clone in the BP step and would never be visible. However, the latter case resulted in small but significant number of clones having been inserted in the wrong orientation. In the completed promoter libraries, there were 4 cases where multiple haplotypes were recovered but only 1 haplotype was confirmed in the correct orientation, with the remainder either cloned in the wrong orientation or with poor sequence coverage. 24 promoters had lost at least one haplotype due to lack of a confirmed clone, but still had at least 2 confirmed haplotypes and were thus included in the final test set.

4.2.9 Successfully cloned promoter SNPs

The 84 promoters with multiple confirmed clones yielded a total of 293 haplotypes. These contained a total of 228 polymorphisms. The cloned polymorphisms are listed in appendix C, and the haplotypes in appendix D. As well as 207 SNPs, these included 6 variable microsatellite repeats, 14 indels of at least 1 base pair and 1 hypervariable region that contained a complex pattern of CA and CG repeats and was impossible to resolve further. These non-SNP polymorphisms were not detectable in the re-sequencing, as ExoTrace is not capable of handling indels or polymorphisms with more than two alleles. Manual re-inspection of the re-sequenced promoter fragments showed that these indels were indeed present, but allele frequency data were not obtainable due to the difficulty of reliably calling heterozygous non-SNP polymorphism. 127 (55.7%) SNPs were already present in dbSNP. More significantly, 57 of the 207 SNPs (27.5%) were not present in the initial re-sequencing data. There are two possible sources for these new SNPs; either they are rare SNPs that were missed in re-sequencing due to the failure of one or more sequence reads and poor sequence quality, or they are polymerase errors artificially introduced by the two-

round PCR. It is difficult to be certain of which cause produced any given polymorphism.

The error rate of KOD polymerase was recently calculated as 1 base every 28.9 kb for a 25-cycle PCR reaction (Bethel et al unpublished observations). This corresponds to 1 base in 13.1 kb for a 55-cycle PCR as used for the promoter fragments, assuming that the error rate increases linearly with the number of cycles. While the use of DMSO (as is the case here) tends to increase the error rate of most polymerases, KOD polymerase can be used with up to 5% DMSO with no decrease in fidelity. DMSO is often used as a PCR additive to improve amplification of GC-rich regions, which promoters often are. 863 clones with complete sequence coverage and confirmed positive orientation were the source of the cloned SNP set. The average size of a PCR fragment is 575 bases. From these figures, 38 polymerase errors might be expected. The number of unexplained base differences discovered here is higher than would be expected, though not entirely inconsistent with this rate of error (57 novel SNPs corresponds to an error rate of 1 per 8.7 kb).

However, there is evidence to suggest that a fraction of these extra SNPs may be real. 7 (12%) of the 57 "new" SNPs matched a dbSNP entry with the same alleles, seemingly confirming that they are true SNPs. This is markedly lower than the 61% of cloned SNPs that match dbSNP overall. This in itself is not necessarily evidence that the majority of these new SNPs are errors, as there is considerable scope for an ascertainment bias in the re-sequencing data that would under represent rare SNPs. Sequencing failure for one individual from the 48-person panel can potentially mask a SNP with a minor allele frequency as high as 0.02 (if the minor allele was represented by a single homozygote). Of the SNPs discovered in the promoter re-sequencing that were already present in dbSNP, only 26/595 (4.4%) had minor allele frequencies of 0.02 or under. In contrast, 98/212 (46.2%) of those not previously in dbSNP had a MAF in this range. Rare SNPs are therefore much less likely to be present in dbSNP, and the low rate of matches to dbSNP in the "new" cloned variants is not necessarily indicative of a high error rate.

There are more extra SNPs near the ends of promoter fragments, with a particularly prominent peak at the extreme 5' end (Figure 25). These are areas where SNP

ascertainment by re-sequencing is most likely to fail, as the beginnings and ends of sequencing reads are often poor quality. The requirement for 2 reads also increases the difficulty of SNP ascertainment, as the antisense read may not reach the very ends of the product. The peak at the 5' end of the fragment also corresponds to where the sequence-specific sequencing primer hybridizes. This would have made SNPs in this area impossible to detect from re-sequencing PCR products, but they are detectable using vector primers in cloned fragments. Another possibility is that synthesis errors in a subset of the primer molecules have introduced base changes that were then detected in a small number of clones. 37 promoters in total contained new SNPs among their cloned haplotypes. This phenomenon has been observed by colleagues in the lab in separate experiments. Of these, 13 (35%) had more than one new SNP; 11 promoters had two, 1 promoter had three, and 1 had four. Such clustering of "novel" SNPs seems unlikely if all of them were PCR errors, and it is possible that where a promoter contains multiple unexplained SNPs some of them are in fact real.

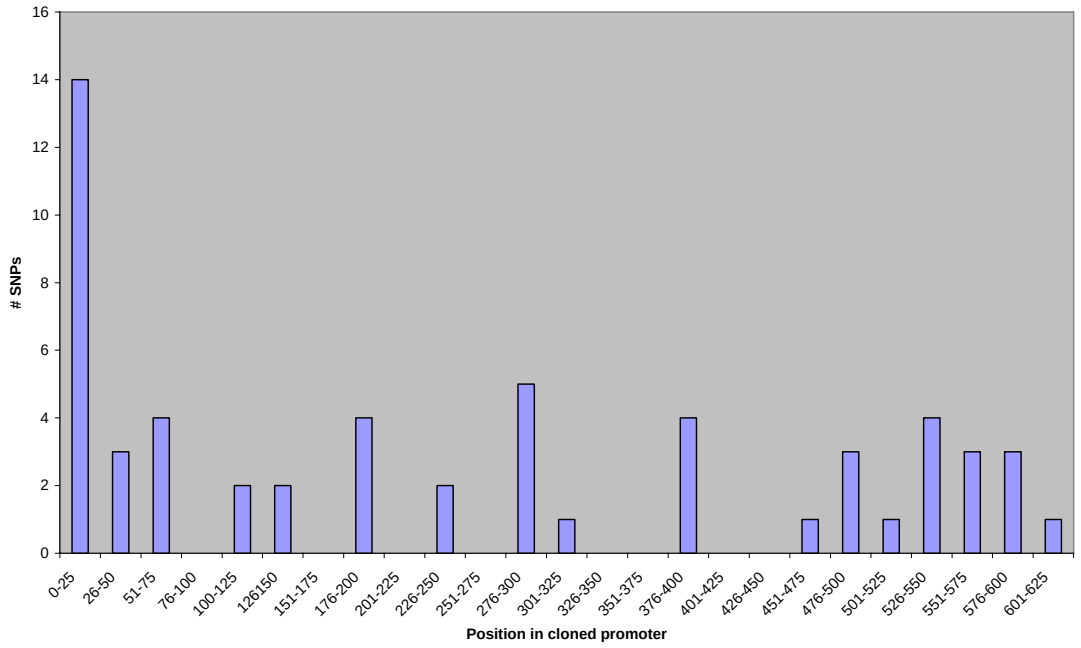


Figure 25. Distribution of SNPs in cloned promoter fragments that were not found in the re-sequencing data.

The only way to be certain which of these SNPs are real would be experimental confirmation. Either the re-sequencing could be repeated with optimised conditions to

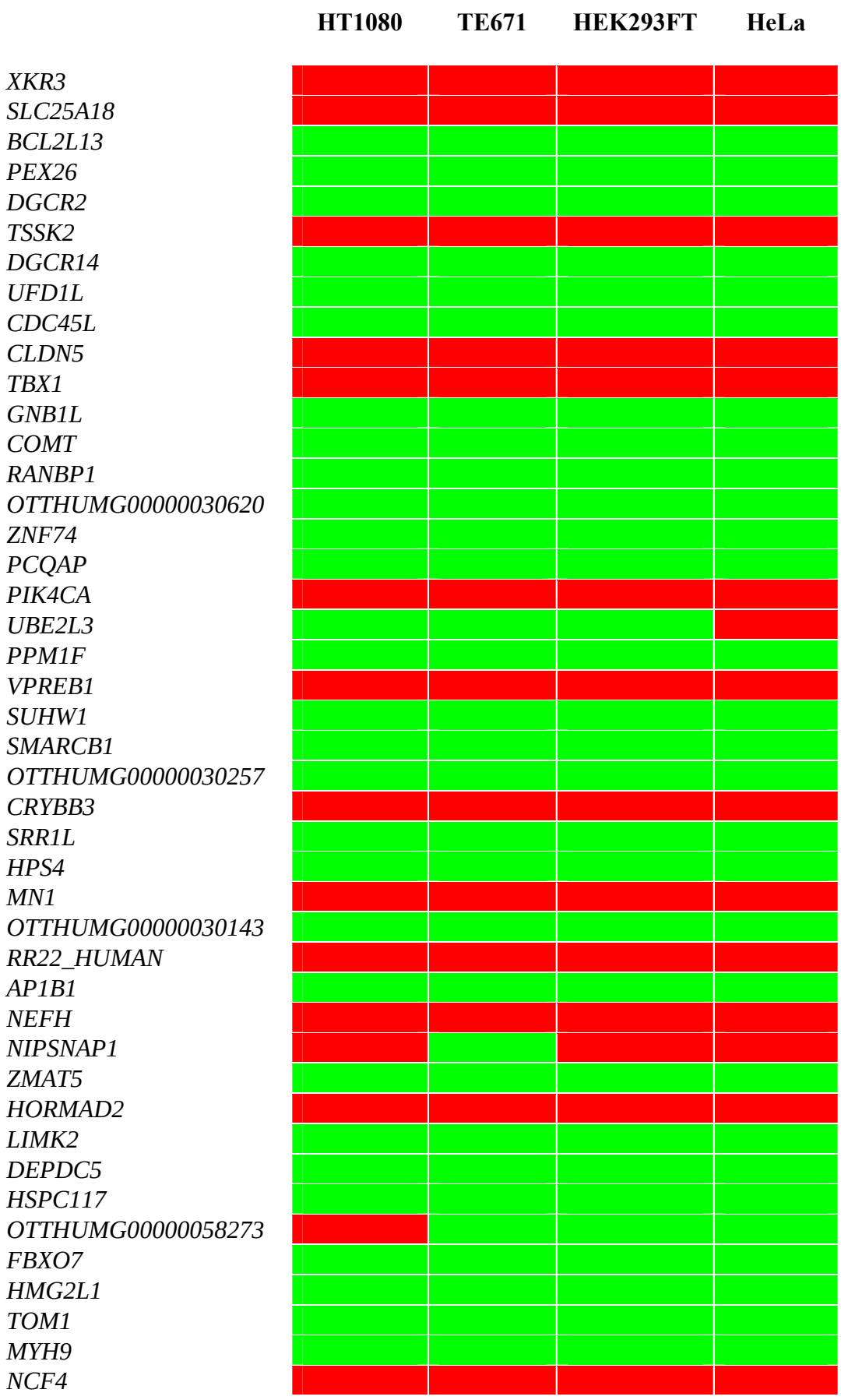
ensure success, or preferably a genotyping assay could be designed to confirm the genotype of the SNPs in each cell with less chance of being affected by surrounding DNA that is less tractable to sequencing. However, due to time constraints this was not possible.

For the purposes of examining the mechanistic aspects of promoter function, it can be argued that any polymorphism between promoter haplotypes may be informative regardless of whether that change is present in natural populations or introduced during the experiment. The creation of non-natural promoter haplotypes by *in vitro* mutagenesis for subsequent analysis in reporter assays is after all not unusual. These polymorphisms were therefore included in the luciferase assays and subsequent analyses related to the mechanism of action of promoters (e.g. context analysis of functional changes). For evolutionary analyses and any others relating to the prevalence of different SNP-types in the population these SNPs were excluded. This was because the presence of false SNPs may lead to erroneous conclusions being drawn, and in any case parameters such as minor allele frequency were not obtainable for SNPs that were not discovered by re-sequencing.

4.2.10 Functional testing of promoter haplotypes with luciferase assays

The library of 293 haplotypes cloned into luciferase reporter plasmids was transfected into HT1080, TE671, HEK293FT and HeLa cells in order to test each promoter for sequence-dependent promoter efficacy variation. A version of QIagen's high-throughput transfection protocol using the Effectene transfection reagent was used, with modifications to improve liquid handling during the procedure. The cells were transfected in 96-well microtitre plate format, with 4 technical replicate wells per haplotype per experiment. One set of 4 wells per plate contained a negative control pGL3 Basic without a promoter cloned into it. Each technical replicate was internally normalised against the *Renilla* control plasmid, and resulting readings expressed relative to the mean of the internally-normalised pGL3 Basic transfections. Two biological replicates of each cell line were transfected with two different plasmid preparations of the cloned promoters, in order to better control for stochastic effects caused by a particular plasmid prep or batch of cells. All cell lines were transfected between passages 3 and 6 after thawing from liquid N_2.

The results showed that the promoter constructs drove levels of luciferase expression that spread 2 orders of magnitude, from promoters that showed no activity to those with several hundred times background level. Determining an exact threshold below which a promoter is deemed inactive is to some extent an arbitrary process. While guidance can be sought from previously published work (Buckland et al. 2005; Cooper et al. 2006), each assay system in use will have its own sensitivity and dynamic range, so the thresholds may not be directly transferable. Here, a promoter is deemed to be active if at least one haplotype had an activity at least 7 times higher than the promoter-less plasmid. Other groups have used different criteria, with the only other large scale studies usually aiming for a threshold of 10x background (Buckland et al. 2005). The lower value of 7x was chosen here because it was observed that promoter constructs with over 10x background activity in one biological replicate sometimes dipped below that threshold in the other replicate, but were clearly still active. In addition, manual inspection of the results suggested that luciferase activity patterns for promoters below this were less reproducible. Using this threshold, each cell "expressed" between 50 and 55 promoters (Table 9). A total of 60 (71.4%) promoters were active in at least one cell line. 12 promoters showed cell-specific activity using the 7-fold cutoff. Cell specific is defined here as differences in activity across cell lines, rather than a promoter being active in one cell line only.


HT1080
TE671
HEK293FT
HeLa
XKR3
SLC25A18
BCL2L13
PEX26
DGCR2
TSSK2
DGCR14
UFD1L
CDC45L
CLDN5
TBX1
GNB1L
COMT
RANBP1
OTTHUMG00000030620
ZNF74
PCQAP
PIK4CA
UBE2L3
PPM1F
VPREB1
SUHW1
SMARCB1
OTTHUMG00000030257
CRYBB3
SRR1L
HPS4
MN1
OTTHUMG00000030143
RR22_HUMAN
AP1B1
NEFH
NIPSNAP1
ZMAT5
HORMAD2
LIMK2
DEPDC5
HSPC117
OTTHUMG00000058273
FBXO7
HMG2L1
TOM1
MYH9
NCF4

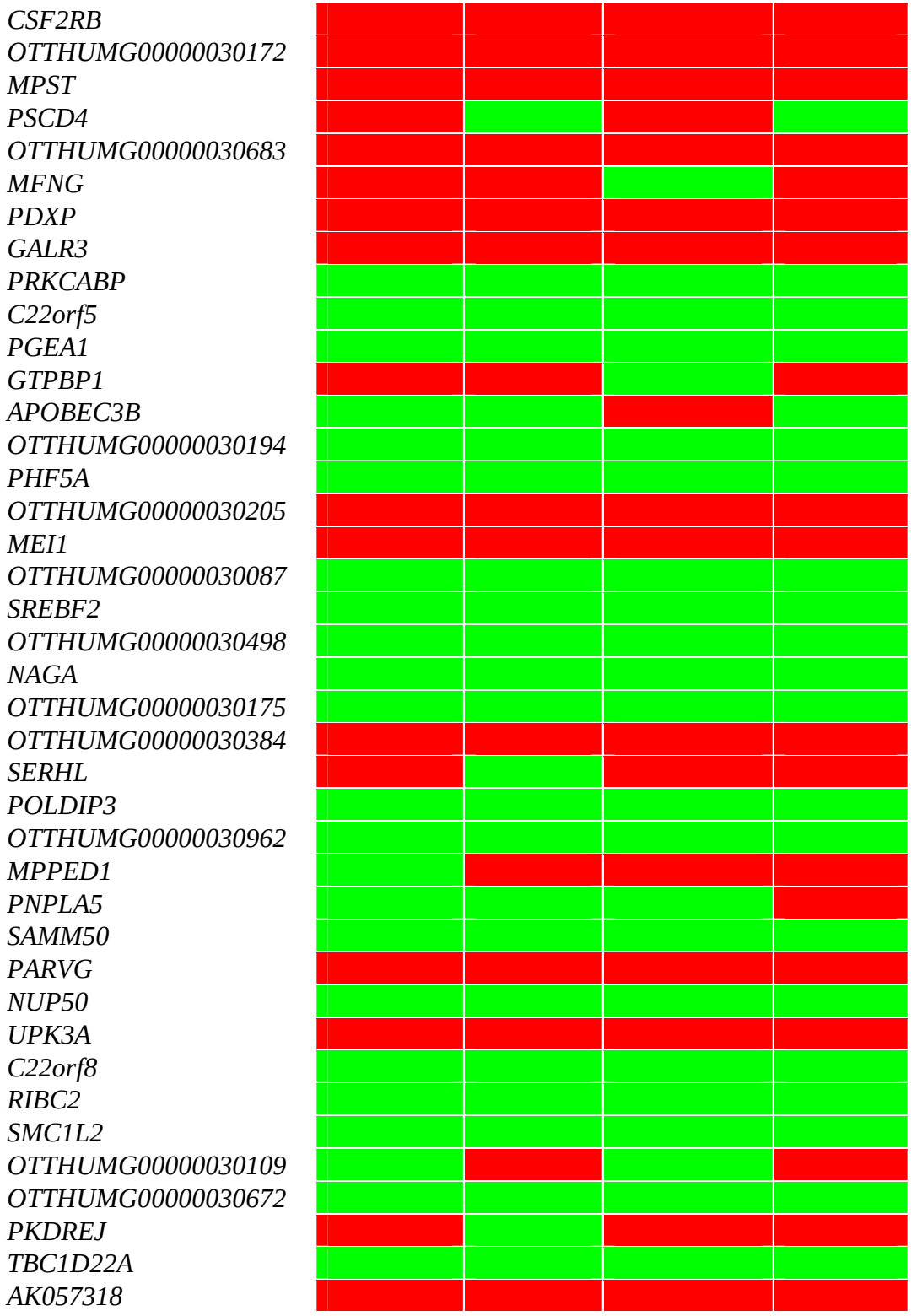


Table 9. Promoters active in each cell line. A promoter was defined as active if at least one haplotype gave a signal at least 7 times that of the promoter-less control plasmid. Promoters in green passed this threshold, while those in red were not active. Promoters are listed in the order of their occurrence along chromosome 22 from centromeric to telomeric ends of the q arm.

4.2.11 Comparison of promoter activities to transcription start site profile and annotation accuracy

Experimental methods to locate and experimentally confirm TSSs by exploiting the 5' cap have recently been developed and applied to mammalian genomes at high-throughput (see section 1.3.2). In particular, CAGE has been used to scan the mouse and human genomes for TSSs (Shiraki et al. 2003; Carninci et al. 2006). This allows the comparison of previously annotated TSSs with experimentally derived TSSs, and a subsequent assessment of the start site annotation.

The CAGE tag data for those genes with cloned promoters were downloaded from the online CAGE data repository run by the FANTOM group (Carninci et al. 2006; Kawaji et al. 2006). 64 of the 84 cloned promoters had their TSSs covered by at least one CAGE tag cluster (a group of overlapping CAGE tags). The TSS from each tag cluster was taken as the position of the highest peak in the distribution of tags in the tag cluster. The relative distance between the TSS according to the CAGE data and the annotated TSSs were plotted against the fraction of promoters with that difference (Figure 26). This showed that the latest annotation of TSSs on chromosome 22 is in general fairly accurate. 61% of annotated TSSs were within 40 base pairs of the experimentally derived position, and 75% were within 60 base pairs. The majority of experimentally verified TSSs seemed to be a few tens of bases downstream of the annotated TSS. If the distribution of active and inactive promoters was analysed separately, 14.5% of inactive promoters were found to have functional TSSs over 100 bases downstream of the annotated TSS, compared to 5.7% of active promoters. In 5 promoters, the CAGE-verified TSS was far enough downstream of the annotated one that it was in fact 3' of the cloned fragment, indicating that the real TSS was not cloned. These promoters might be expected not to function *in vitro*, particularly if they are ones that rely on motifs such as the initiator or DPE (see section 1.1.1). Interestingly however, two of these promoters are active in all cell lines, and a third is active in two out of four. Their CAGE-verified TSS was only between 15 and 25 base pairs downstream of the end of the cloned fragment. Of the remaining two promoters, one was inactive across all cell lines, and one was only active in one. These two had CAGE-verified TSSs 101 and 90 base pairs downstream of the end of the fragment

respectively, meaning that none of the core promoter elements would have been cloned. There were no instances where the CAGE-verified TSS was 5' of the start of the cloned fragment.

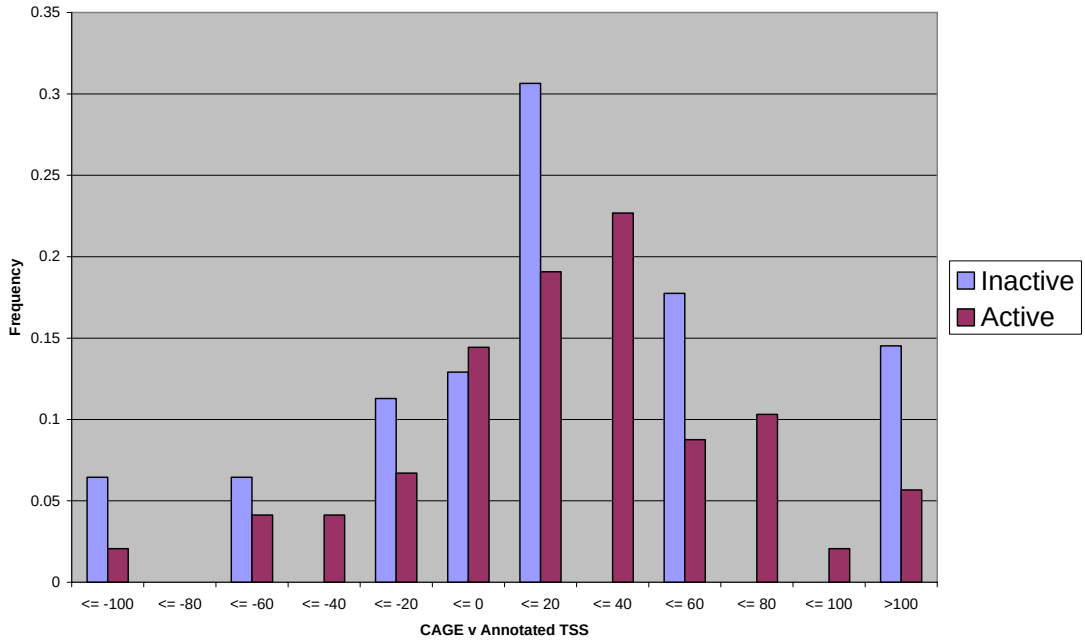


Figure 26. Correlation between accuracy of TSS annotation and cloned promoter activity. The x axis is the position of the CAGE-verified TSS relative to the annotated TSS.

In addition to a simple determination of the start site, CAGE also enables the architecture of the TSS to be examined. Carninci et al found that TSSs can be classified according to the stringency of the start site, with some genes having very tightly defined start sites, and others with much broader start sites where individual transcripts can start from anywhere within a window, which could sometimes span 100 base pairs. It could be hypothesised that if a promoter has a very tightly-defined start site, then any sequence differences that disrupt transcription from that site might have a more dramatic effect than in a promoter with a broader start site. In the latter case, the breadth of the TSS may enable it to tolerate sequence changes that disrupt transcription from a particular part of it. To test this idea, the promoters were classified by whether they had broadly or tightly defined start sites, according to whether at least 50% of the CAGE tags in the cluster fell within a 5 base pair span (Carninci et al. 2006). This classification was carried out by Boris Lenhard at the Bergen Center for Computational Science, an author on the CAGE paper.

Dr. Lenhard's advice was that such classifications were only reliable if at least 100 individual CAGE tags were available for the TSS being assessed. With this restriction, only 4 promoters could be reliably designated as having a tight start site and 9 as having a broad start site. Of these 1 and 3 promoters respectively were not active in the luciferase assays and were discarded. For each of the remaining promoters, the activity difference between the highest and lowest activity haplotypes was calculated, and the average maximum activity difference was compared for tight and broad start site categories. Although there was a higher maximum activity difference in tightly defined promoters (3.9x) relative to broad promoters (3.4x) this difference was not significant (p-value = 0.71, Mann-Whitney test). This is perhaps not surprising given the very small numbers of promoters in each category.

It was also noted that genes with broad TSSs tended to be correlated with CpG island-containing promoters (Carninci et al. 2006). Thus if there was a correlation between TSS definition and the impact of promoter SNPs, this might be detectable by a comparison of CpG island- and non-CpG island-containing promoters. In this case, the mean maximum activity differences in each category were 10.8x and 14.1x respectively. Again, the difference was not significant (p-value = 0.11, Mann-Whitney test).

4.2.12 Analysis and visualization of haplotype differences

Before analysing specific activity differences between haplotypes for functional effects, it was helpful to check whether there was a general trend for activity differences to increase with sequence divergence. For every haplotype pair where at least one haplotype was active (i.e. 7x background activity), the number of polymorphisms where the two haplotypes differed was counted. This was assigned as the divergence score. The absolute difference between the promoter activities was also calculated as the ratio of the more active haplotype to the less active one. Superficially, the plot of these results suggests that in fact the reverse is true; that more diverged haplotype pairs were less likely to have different promoter activities than less diverged pairs (Figure 27). However, closer examination showed that this is likely due to a sampling difference rather than a real trend. For every increase of one mutation in the first 4 divergence levels, the number of haplotypes in that category

decreased by 200. The number of points at divergence levels 5-8 was much lower still. The mean and median promoter activity ratios were flat across the divergence scores. This suggests that there is no correlation between the amount of sequence divergence between promoters and the magnitude of the activity difference between them, and that the context of particular promoter polymorphisms is more important than simple promoter sequence divergence.

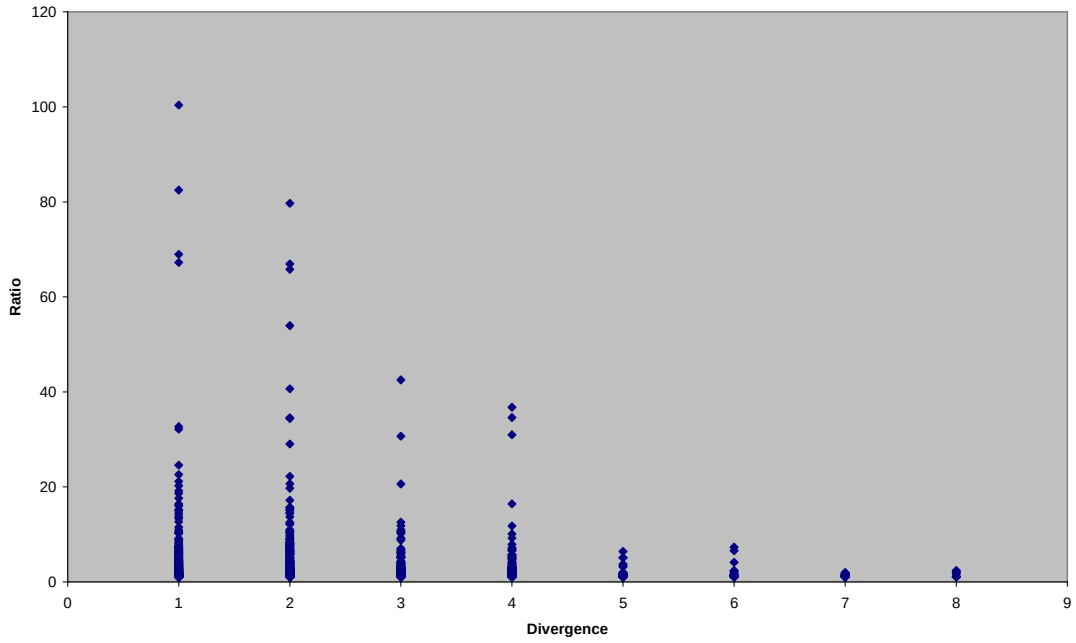


Figure 27. The effect of sequence divergence between promoter haplotypes on the degree of difference in promoter activity. Each point is the ratio of the activity levels of a pair of haplotypes from the same promoter, and all possible haplotype pairs where at least one haplotype is active are plotted. Each biological replicate is plotted separately.

Reporter assay results are normally visualized by plotting a simple bar chart of the mean of each tested construct, and analysed using simple statistical tests either against all possible combinations of constructs, or between ones where a difference is detectable on the chart. While this is the most intuitive representation and works well for studies of single promoters or small numbers of haplotypes, it is problematic when dealing with larger datasets. Where there is a relatively large number of haplotypes per promoter and many promoters to analyse at once, examining bar charts by eye is not an efficient method as it does not make it clear what the finer relationships are between the haplotypes over and above a simple rank of activity level. For the dataset generated in this project, two broad analysis paths were followed.

The first analysis involved improving the visualisation of the data and the integration of biological replicates into one figure. It was evident from manual inspection of simple bar charts of the data that, while the patterns of variation between haplotypes were often reproducible, the absolute magnitude of expression was not (Figure 28). When plotting the data from replicates alongside each other in a chart, it was often not clear how reproducible the variation patterns were, or even which differences were conserved between cell lines. In order to integrate the replicates and better represent variation across cell type, the data were plotted as the Z score. For each haplotype, this was calculated as the difference of the haplotype's activity from the median of the activities of all haplotypes in the promoter, divided by the standard deviations of the activities in the promoter. The Z score for each biological replicate was calculated, and the median between them plotted. An example is shown in Figure 29a. It must be stressed that for the purposes of this project, Z scores were calculated and plotted purely to aid visualisation of the results, and were not used for statistical calculations.

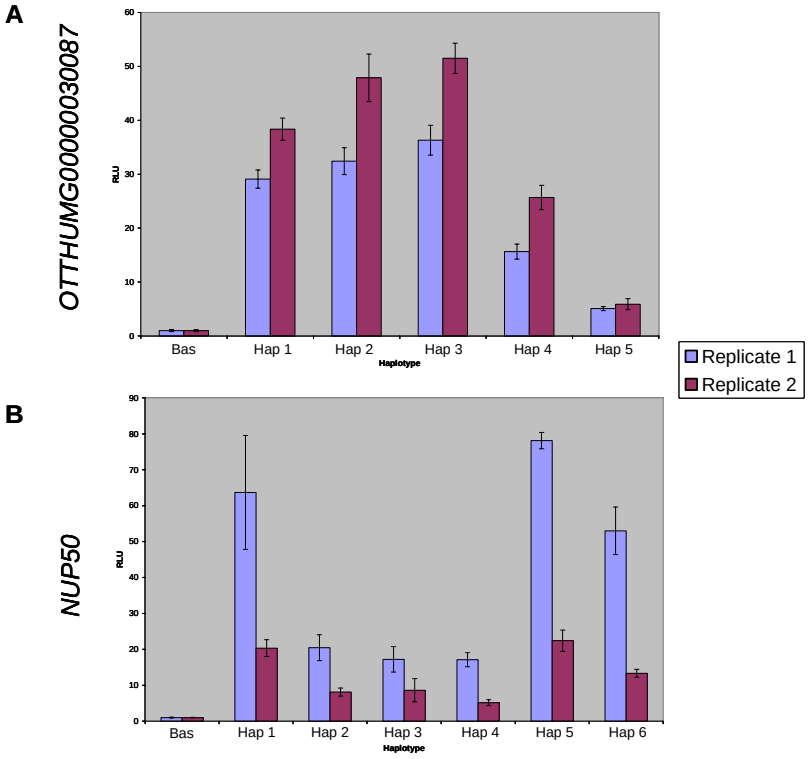


Figure 28. Conservation of promoter activity patterns but not magnitude of luciferase expression in 2 promoters. A) *OTTHUMG00000030087* luciferase results in TE671. B) *NUP50* luciferase results in HeLa. Error bars represent the standard deviation of 4 technical replicates. Promoter activities are plotted in relative light units (RLU), which is the fold increase of the firefly luciferase / *renilla* luciferase ration in a haplotype construct over a promoterless vector (Bas).

While bar charts and Z score plots are useful for seeing the overall picture of functional variation within a promoter, it is difficult to correlate variations with the underlying sequence differences by eye. Some statistical basis is needed for differentiating haplotypic functional variation that is significant from that which is not. Previous experiments, both large scale and small scale, have historically relied on some variant of the t-test to calculate the statistical significance. This can be problematic when doing large numbers of tests, as the number of false positives will start to rise unless corrected for multiple testing. Here, a more conservative two-stage process was used that minimised the number of tests carried out and accounted for multiple testing within the methodology. The first step was to determine whether there was a significant difference between the means of the luciferase expression driven by the different haplotypes within each promoter by carrying out a one-way analysis of variance (ANOVA) test. This will identify variation between haplotypes without giving any information about which haplotypes are different from which others. The ANOVA was carried out for each biological replicate set independently. If a promoter had a p-value below 0.05 in both biological replicates, it was tentatively deemed to be functionally polymorphic.

In order to determine which haplotypes in a functionally variable promoter, as defined by ANOVA, contained the functional alleles, post-hoc statistics were carried out for each possible haplotype pair. This was done using Tukey's Honestly Significantly Different test (Tukey's HSD). This is a relatively conservative post-hoc test that is based on the student's t-test, but incorporates the ANOVA results. Only datasets (in this case promoters) that have significantly different means by ANOVA are subjected to the pairwise comparisons. The critical value for significance in each case is influenced by the amount of variance in the results and the number of means being compared. With Tukey's HSD the experimentwise error rate (i.e. the probability of at least one false positive) is kept at the significance threshold specified (for example the standard value of 0.05). This is a significant advantage in a situation where many non-independent tests are being carried out simultaneously, and which would normally need to be corrected to compensate for an increase in the experimentwise error rate. This comes at a cost of decreased power to detect true positives.

Tukey's HSD was performed on the promoter results using the R statistical language, with the help of Juanma Vaquerizas at the European Bioinformatics Institute. Each biological replicate was treated as a separate experiment. A perl script was then written to integrate the Tukey results into a single visualization of the significance of each haplotype pair. A pair of haplotypes was considered to have significantly different activities if it fulfilled the following criteria:

- The promoter must have significant variance between haplotypes overall in both biological replicates (this is implicit in the Tukey test)
- The activity of the two haplotypes must be significantly different by Tukey's HSD in both biological replicates
- The direction of the difference must be the same in both biological replicates
- At least one of the two haplotypes must have an activity greater than 7x background in at least one biological replicate

The results were plotted as a matrix of all possible comparisons, with each cell coloured according to whether the comparison passes the criteria above and the direction of the difference. An example is shown in (Figure 29b). The Z score plots and matrices for all promoters active in at least one cell line are included in appendix E.

Among the 293 haplotypes in 84 promoters, there were 507 possible haplotype pairs within promoters. Of these, the number of pairs with statistically significant and reproducible differences was 65 (12.8%) in HT1080, 116 (22.9%) in TE671, 102 (20.1%) in HEK293FT and 98 (19.3%) in HeLa.

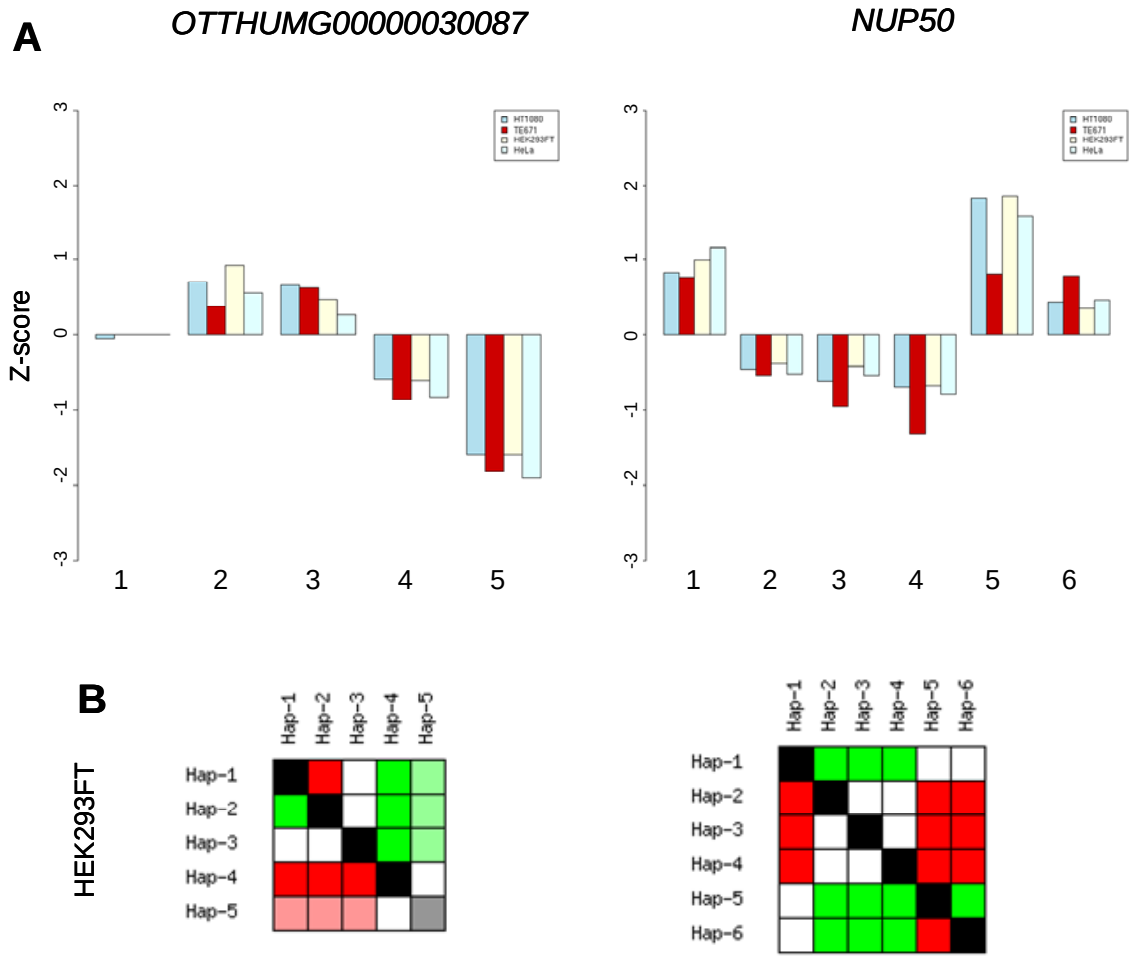


Figure 29. Visualisation of the luciferase reporter results for the *OTTHUMG00000030087* and *NUP50* promoters shown in Figure 28. A) Plot of the Z scores for each of the 4 cell lines. In this visualisation, the Z score represents deviation from the median, and the score itself is the median of the values for each biological replicate. These plots are made purely for visualisation purposes, and all statistics were calculated using the normalised experimental data. B) Matrix showing the significant differences between haplotypes in TE671 cells. Green squares mean that the haplotype represented by that row has significantly higher activity than the one represented by the column, according to Tukey's HSD. Red squares show where the haplotype in that row has significantly lower activity than the one in the column. Pale red and pale green squares carry the same meaning, but also designate that one of the two haplotypes being compared does not meet the 7x activity threshold. White squares designate no significant difference between the haplotypes. On the diagonal, black squares indicate that the haplotype is above the 7x background threshold, and therefore active, whereas grey squares indicate that the haplotype has under 7x background activity. Data for all tested promoters with at least one active haplotype are shown in appendix E.

4.2.13 Analysis of individual functional SNPs

The best way to assess the functional significance of each promoter polymorphism is to have a pair of haplotypes where that polymorphism is the only difference. A perl script was created that would identify the haplotype pair with the least sequence divergence as well as the two alleles of each polymorphism. In total, 152 of the 228 polymorphisms were isolated in a haplotype pair with no other differences. For 51 polymorphisms the closest pair of haplotypes contained only one other difference, 19 contained two and 6 contained six (all belonging to the same promoter). A promoter polymorphism was deemed to be functional if it was isolated in a haplotype pair, and if that pair demonstrated reproducible and significant differences in activity (as defined by a p-value below 0.05 by Tukey's HSD and a consistent direction of change). Where polymorphisms could only be isolated to haplotypes with one other difference, both polymorphisms were designated functional in the absence of further resolving power. 65 polymorphisms were in haplotype pairs that passed these criteria, and therefore reproducibly affected promoter activity (Table 10). Of these, 51 polymorphisms were isolated within a haplotype pair, and were thus confirmed as causative variants. 12 polymorphisms were isolated to 6 pairs of haplotypes, but could not be separated any further, and it was thus unclear whether both were functional, or one was more important than the other. 2 SNPs were isolated to a haplotype pair with three differences, but the third was itself tested as a unique difference in a different haplotype pair and found not to be causative. 13/65 (20%) functional polymorphisms were unidentified in the original promoter re-sequencing, not including indels that were undetectable by re-sequencing and one additional unconfirmed SNP that matches a dbSNP entry. This is not significantly different to equivalent 50/228 (21.9%) unconfirmed SNPs overall ($p = 0.71$, χ^2), suggesting that the unconfirmed and confirmed SNPs share similar distributions across the promoters. 37 (57.8%) polymorphisms in total match a dbSNP entry, the same proportion as in the cloned set overall.

While the majority (80%) of the polymorphisms had statistically significant effects in more that one cell line, only 40% were functional in all 4 cell lines. It would be difficult to characterise 60% of polymorphisms as having cell-specific effects, because it is not always clear whether these are biologically cell-specific as opposed

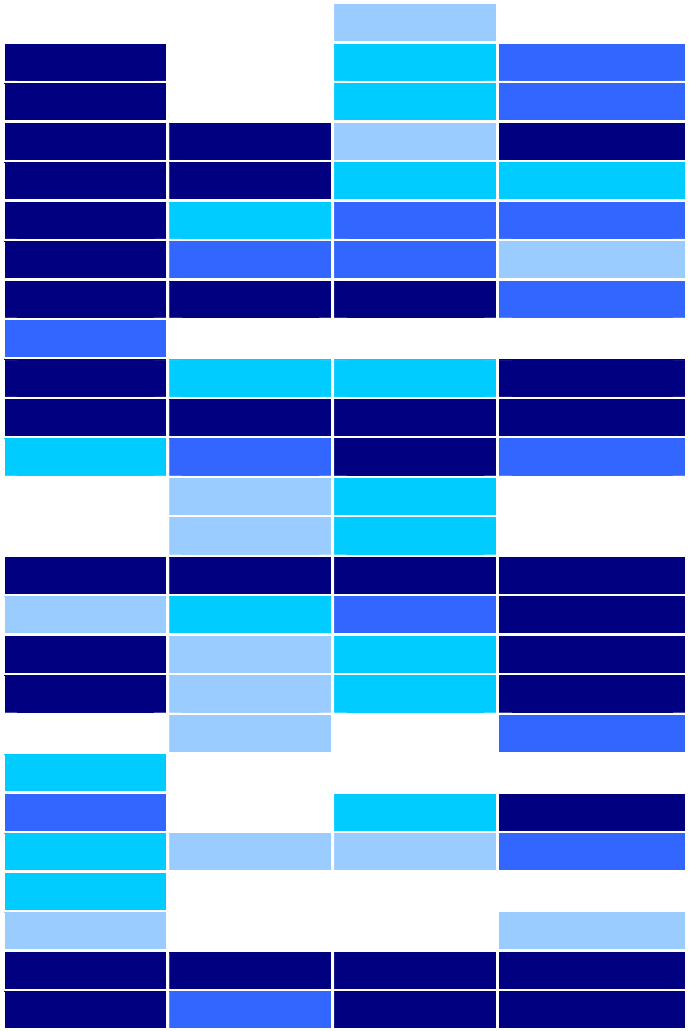
to a lack of statistical significance due to variability between technical replicates. Interestingly, these results using 4 cell lines have not revealed more cell type-specific variation than previous studies using 2 of the same cell lines used here (Buckland et al. 2005). It is also striking that where a polymorphism is functional in more than one cell line, that difference is always in the same direction. There are no examples in this data set of an allele upregulating expression in one cell line and downregulating it in another.

This set of 65 polymorphisms whose effects have either been isolated or near-isolated accounts for the vast majority of sequence-dependent functional variation in this study, with almost every functionally different haplotype pair containing at least one of them. In most cases, variation in adjacent SNPs did not affect the activity difference, at least at the qualitative level. True quantitative assessment of this was not possible, because the magnitude of the expression difference was inconsistently reproduced across biological replicates.

A strong bias for functional polymorphisms to be located within 200 base pairs upstream of the TSS has been previously reported (Buckland et al. 2005). This bias was not reproduced in this study, with no obvious trend in the location of functional polymorphisms relative to general polymorphisms visible (Figure 30). This may be due partly to the different criteria for accepting functional SNPs in the two projects (see section 4.3).

Promoter	SNP	Alleles	dbSNP	Divergence	Comparison	Low	High	TE671	HT1080	HeLa	HEK293FT
DGCR2	-467	C/T	rs17526612	1	3-2	C	T				
DGCR2	-13	C/T	rs17526619	1	2-1	T	C				
DGCR14	-408	[A]n		1	8-1	12	8				
DGCR14	-212	C/T	rs1936951	3	6-1	T	C				
DGCR14	-207	T/A	rs1936950	3	6-1	A	T				
DGCR14	-152	C/T	rs737923	1	7-6	C	T				
CDC45L	-124	C/G	rs4141528	1	2-1	G	C				
GNB1L	-288	C/T	rs28451568	1	2-1	C	T				
COMT	-268	C/T	rs13306278	1	5-3	C	T				
RANBP1	-66	G/T	rs2286929	1	5-4	G	T				
OTTHUMG00000030620	-324	G/A		1	3-2	G	A				
UBE2L3	-479	T/-	rs9623962	1	1-2	-	T				
SUHW1	-65	A/T	rs4822092	1	2-1	T	A				
OTTHUMG00000030143	-119	G/C		1	2-1	G	C				
NIPSNAP1	-278	T/G		2	2-1	G	T				
NIPSNAP1	-254	A/G		2	2-1	G	A				
ZMAT5	-297	C/T	rs17526577	1	3-2	T	C				
ZMAT5	-95	C/A		1	3-1	A	C				
DEPDC5	-199	G/C		1	2-1	C	G				
HSPC117	-297	T/-		2	2-1	T	-				
HSPC117	-115	C/T	rs17555307	2	2-1	T	C				
FBXO7	-350	C/-		1	3-1	C	-				
TOM1	-302	C/T		1	4-2	T	C				
MYH9	-115	C/-	rs17526626	1	4-3	C	-				
PSCD4	-98	[GTTT]n		1	3-2	6	5				
PRKCABP	-64	G/A	rs11089858	1	2-1	A	G				

Promoter	SNP	Alleles	dbSNP	Divergence	Comparison	Low	High	TE671	HT1080	HeLa	HEK293FT
PGEA1	-524	C/T		1	2-1	C	T				
GTPBP1	-349	C/G	rs2267393	2	2-1	C	G				
GTPBP1	-335	C/T	rs2267394	2	2-1	C	T				
APOBEC3B	+30	T/C		1	3-1	C	T				
OTTHUMG00000030194	-426	G/T		1	4-3	T	G				
OTTHUMG00000030194	-229	G/A		1	5-4	G	A				
PHF5A	-525	C/T		1	3-1	C	T				
PHF5A	-142	G/A		1	2-1	A	G				
OTTHUMG00000030087	-300	C/T		1	3-1	T	C				
OTTHUMG00000030087	-144	G/A	rs738248	1	4-3	A	G				
OTTHUMG00000030087	+73	C/G	rs139562	1	5-3	C	G				
OTTHUMG00000030498	-158	C/G	rs4822079	1	2-1	G	C				
NAGA	-136	A/T	rs2859438	2	2-1	A	T				
NAGA	-106	G/A	rs133377	2	2-1	A	G				
OTTHUMG00000030175	-479	C/T		1	5-3	C	T				
OTTHUMG00000030175	-126	G/A	rs8135801	1	2-1	A	G				
SERHL	-450	G/A		2	2-1	A	G				
SERHL	-356	G/C		2	2-1	G	C				
POLDIP3	-438	G/A	rs137115	1	2-1	G	A				
POLDIP3	-281	G/A	rs137114	1	3-2	G	A				
OTTHUMG00000030962	-347	C/A		1	4-1	A	C				
OTTHUMG00000030962	-249	C/T	rs5759182	1	2-1	C	T				
PNPLA5	-418	C/G	rs11913819	1	2-1	G	C				
SAMM50	-21	C/A		1	3-1	C	A				
NUP50	-514	C/A		1	5-4	A	C				
NUP50	-153	G/C	rs132847	1	3-1	C	G				



Promoter	SNP	Alleles	dbSNP	Divergence	Comparison	Low	High	TE671	HT1080	HeLa	HEK293FT
NUP50	-43	G/T	rs3788634	1	2-1	T	G				
C22orf8	-431	A/T	rs226504	2	2-1	T	A				
C22orf8	-110	GGGCG/ -----		2	2-1	-----	CCCGC				
RIBC2	-388	G/A		1	5-4	G	A				
RIBC2	+41	C/A	rs2272804	1	5-2	A	C				
SMC1L2	-268	G/A		1	5-1	G	A				
SMC1L2	-200	C/T	rs2272805	1	6-5	C	T				
SMC1L2	-126	G/T	rs2272804	1	5-2	G	T				
OTTHUMG00000030109	-335	C/T	rs9615411	1	4-3	T	C				
OTTHUMG00000030109	-17	C/T		1	2-1	C	T				
OTTHUMG00000030109	+47	C/T	rs3747243	1	4-2	C	T				
OTTHUMG00000030672	-421	G/A	rs6008320	1	3-1	G	A				
TBC1D22A	-91	C/T	rs2295441	1	3-2	T	C				

Table 10. Functional promoter polymorphisms discovered by luciferase assays of cloned haplotypes. For each polymorphism, the haplotype pair with the lowest sequence divergence is listed (Comparison), along with the divergence itself. The divergence is the number of polymorphisms where the two haplotypes in the pair differ (e.g. a haplotype pair that differed by a single 5 base pair indel would have a divergence of 1). Low and high alleles refer to the genotype at that polymorphism in the haplotypes that had lower and higher activities respectively. For each cell line, the less significant of the two p-values calculated from the two biological replicates by Tukey's HSD is categorised into a significance level as per the blue shading in the legend above. Where no shading is present, that comparison was not reproducibly significant in that cell line.

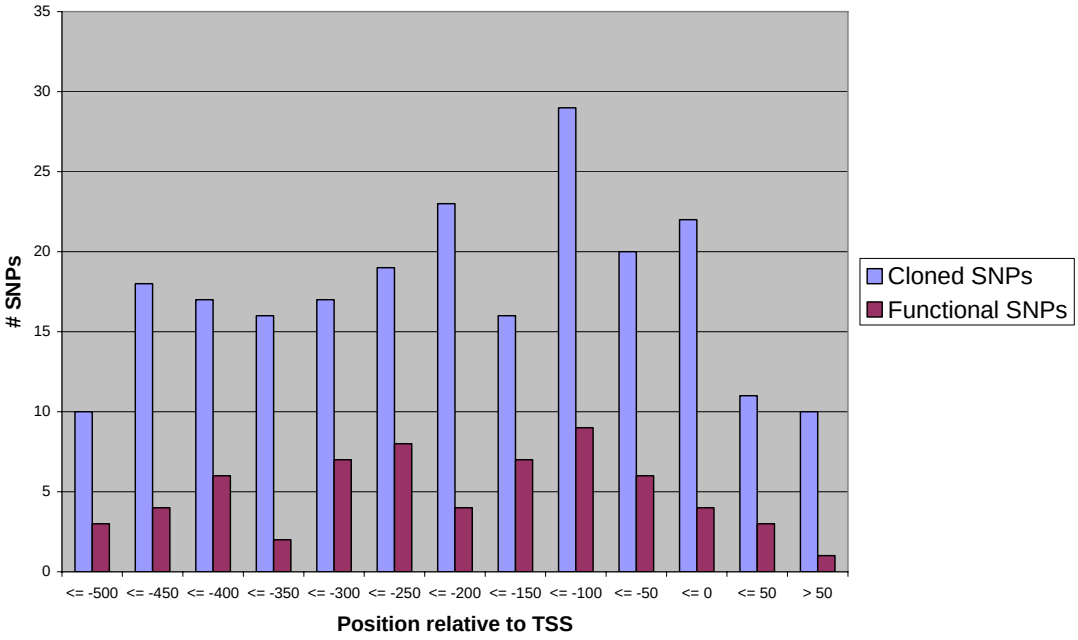


Figure 30. Distribution of cloned promoter polymorphisms and functional promoter polymorphisms as a function of distance from the TSS. No clear differences are visible between the distributions of functional relative to cloned polymorphisms overall. There is a marked drop in the number of both classes of SNPs 3' of the start site, which is probably a combination of reduced SNP ascertainment near the ends of PCR amplicons and increased selective restrictions within the gene itself.

4.2.14 Synergistic effects between functional SNPs

While at least one of the 65 isolated promoter SNPs is variable in most haplotype pairs, there were 14 haplotype pairs that have different activities and differed at 2 or more sites, but where these polymorphisms did not cause a difference in isolation (or in their closest available haplotype pair). These were distributed across 6 promoters. One of these, the promoter for the VeGA gene *OTTHUMG00000030257*, contained the hypervariable microsatellite region, which was not identical in any pair of haplotypes. It was therefore difficult to determine whether differences are caused by this region or are effects of other polymorphisms in the promoter. The number of differences between the members of each pair ranged from 2 to 7. The obvious explanation for these differences is a synergistic effect of these SNPs on promoter activity, where the right combination of alleles is required before a change in activity is observed. It is also possible that a subset of the differences in these promoter pairs is functional, and the others have no effect. Where these haplotypes differ by more than two polymorphisms, it is not clear whether they are all working synergistically or

only a subset of them. In one case, the *RIBC2* promoter, synergy between a pair of SNPs is clearly deducible as there are only three polymorphisms and 5 haplotypes, enabling more detailed dissection of the functional effects. The presence of an A at position +41 and a G at position -388 cause a significant downregulation of promoter activity (Figure 31). Individually neither of these SNPs has an effect on promoter activity, as evidenced by the lack of statistical significance in the comparisons of haplotypes 1 and 2, and haplotypes 1 and 4. This demonstrates that synergistic effects between multiple SNPs can be a factor in causing sequence-dependent promoter activity variation.

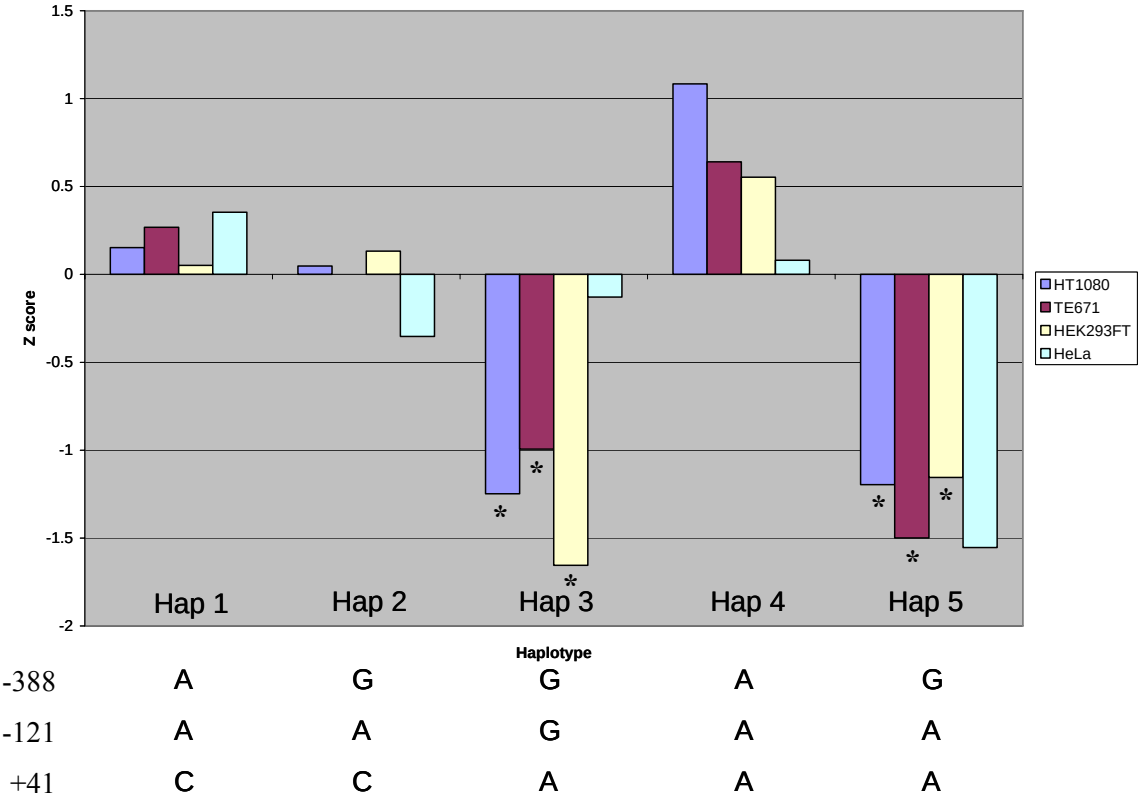


Figure 31. Z score plot of haplotype activities in the *RIBC2* promoter. Haplotypes 3 and 5, both containing a G at position -388 and an A at position +41 are significantly lower than other haplotypes in HT1080, TE671 and HEK293FT (marked by asterisks). There were no reproducibly significant differences between any haplotypes in HeLa cells, despte the dip in the Z score shown here.

4.2.15 Context analysis of functional SNPs

The locations of the 65 functional SNPs were analysed for their presence in putative functional elements using the same criteria as in chapter 3. 41 (63%) were present in a

motif of any kind, compared to 115 (50.4%) for the whole set of 228 cloned SNPs. This equates to an enrichment of only 1.25x, suggesting that the currently known regulatory motifs are not good predictors of function in an *in vitro* system at least in this panel of 4 cell lines. The single motif class most enriched around functional SNPs compared to non-functional ones is cisRED, where a 1.87x enrichment was observed. The poorest motif class was the TFBS motifs from Transfac.

	Functional SNPs (65)	Cloned SNPs (228)	Enrichment
phastcons regions	8	21	1.34
cisRED motifs	8	15	1.87
TFBS (Tranfac)	3	12	0.88
TFBS (Jaspar)	23	68	1.19
Conserved TFBS	0	0	N/A
Quadruplex sites	2	4	1.75
SNPs in putative regulatory regions	34	97	1.23

Table 11. Enrichment of functional SNPs vs promoter SNPs in putative regulatory motifs.

Some previous work has attempted to use a combination of evolutionary conservation and the presence of TFBS to predict functional SNPs *a priori* (Belanger et al. 2005; Mottagui-Tabar et al. 2005). This strategy was also tested by calculating the number of cloned and functional SNPs present in a TRANSFAC or JASPAR binding site that was itself within a conserved region. Conservation was represented either by a phastcons region or the presence of a cisRED motif (although the latter is not strictly speaking a measure of conservation, the motifs are discovered using methods heavily reliant on comparative genomics). This revealed that 9 cloned polymorphisms were present in such locations, and that 6 of these were functional. This corresponded to an enrichment of 2.35x for functional polymorphisms within these regions, an improvement on any of the putative elements alone but still not a large enrichment that would be useful for prediction.

The 5x regulatory potential scores (Kolbe et al. 2004) of the functional polymorphisms were also compared to the overall scores for the cloned polymorphism set. There was an increase in the proportion of polymorphisms with a score of 0.01 or greater in the functional set (57% vs. 47%) but this was not statistically significant ($p = 0.074$, χ^2). This is the score associated with conservation patterns present in known regulatory elements. The mean scores for functional and promoter SNPs were 0.063 and 0.06 respectively, also not significantly different ($p = 0.42$, t-test). Apart from this overall skew towards higher scores the profile of score frequencies is not markedly different, and does not present features that could clearly be used as predictors of *in vitro* function for particular polymorphisms (Figure 32).

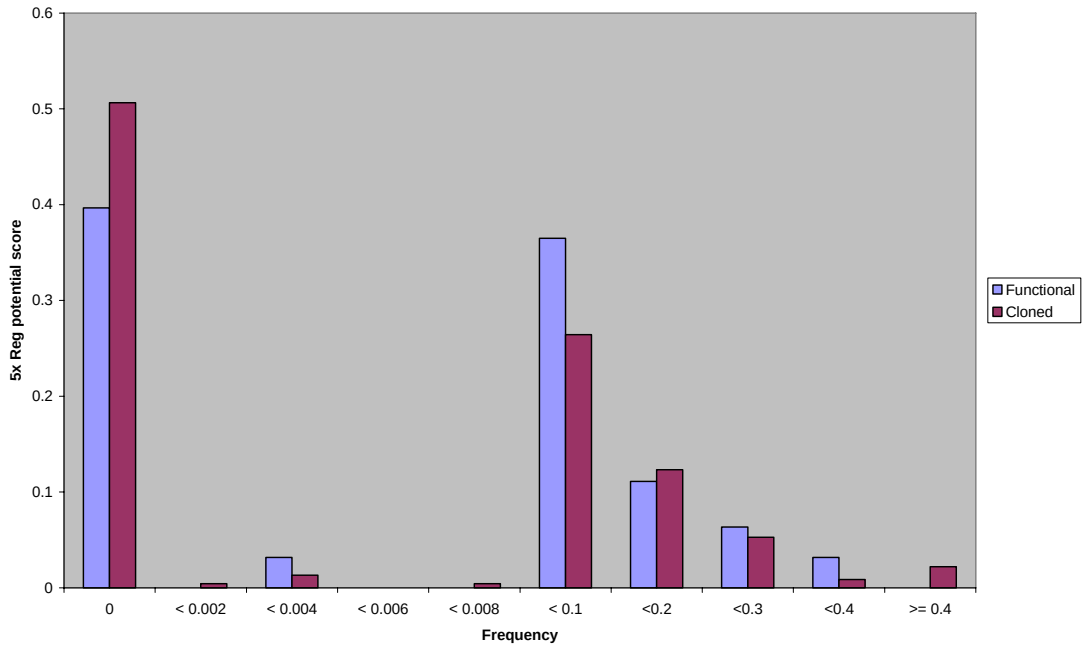


Figure 32. Frequencies of regulatory potential score for functional polymorphisms and promoter polymorphisms overall.

4.2.16 Evolutionary analysis of functional polymorphisms

A simple list of functional polymorphisms does not reveal the direction of each mutation, and thus the direction of the change in promoter activity (as determined by the luciferase experiments) that resulted from that mutation. This is of interest

because recent theoretical work has proposed a neutral model of transcriptome evolution where changes that lead to a decrease in gene expression (downregulatory changes) outnumber those causing an increase (upregulatory changes) (Khaitovich, Paabo, and Weiss 2005). Upregulatory changes, when they do occur, were predicted to cause a larger magnitude change on average than downregulatory changes (Khaitovich, Paabo, and Weiss 2005).

In order to determine which allele at each functional SNP is ancestral, the chimp and macaque genomes were used as outgroups to root the SNPs, assuming that the allele present in chimp is the same as the ancestral allele in human (see chapter 3). GALAXY 2.1 was used to extract the chimp or macaque allele from precomputed alignments of human-chimp and human to macaque (Giardine et al. 2005). Where a SNP was not covered by the chimp alignment, the macaque alignment was used. 57 functional polymorphisms in total were covered by at least one of the two primate genomes. One of these was a poly-A microsatellite, and was ignored due to the extreme variability of these repeats (making it difficult to say whether the primate alleles are themselves hypervariable and thus not suitable as a root). 28 upregulations and 27 downregulations were discovered in this study, showing no evidence for this bias ($p=0.89$ by χ^2 test). If the SNPs not present in the re-sequencing (and whose veracity is therefore in question) are removed from consideration, the figures are 16 upregulations and 21 downregulations ($p=0.41$ by χ^2 test).

A plot of the allele frequencies for the low-activity and high-activity alleles reveals a skewed distribution, with high-activity alleles more frequently having high allele frequencies than low-activity alleles (Figure 33). This can be caused by a combination of two factors; either a mutation causing a downregulatory change that fails to spread in the population or a mutation causing an upregulatory change that expands in the population. There is no direct information here on the mechanism of this potential expansion, whether by selection, genetic drift, or founder effects.

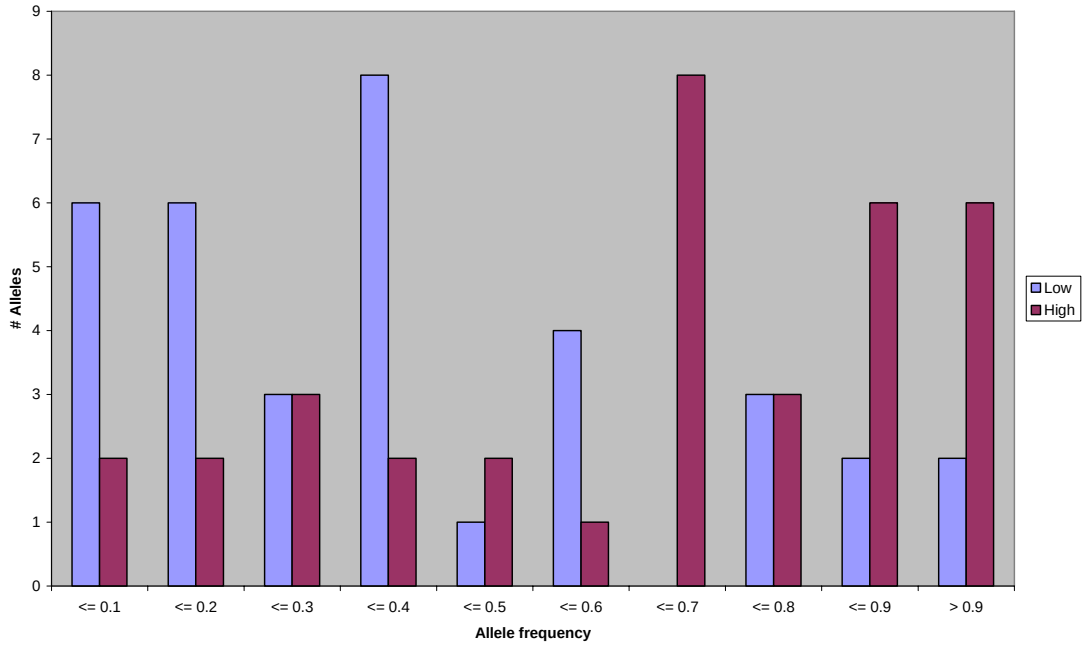


Figure 33. Allele frequencies of the low- and high-activity alleles in the functional SNPs for which frequency information was available

If the frequencies of only the derived alleles are plotted, a striking peak is visible for high frequency alleles that cause an upregulation of promoter activity (Figure 34). This bias towards high frequency appears quite extreme, with 60% of high-activity derived alleles having a frequency greater than 0.8, and 30% over 0.9. This indicates that, in this dataset, mutations causing an increase in promoter activity have expanded considerably in the population. Whether this is due to selection or other factors is still not clear from this data alone, but the number of very high frequency alleles suggests that selection may have been a factor in the population history of at least some of the SNPs.

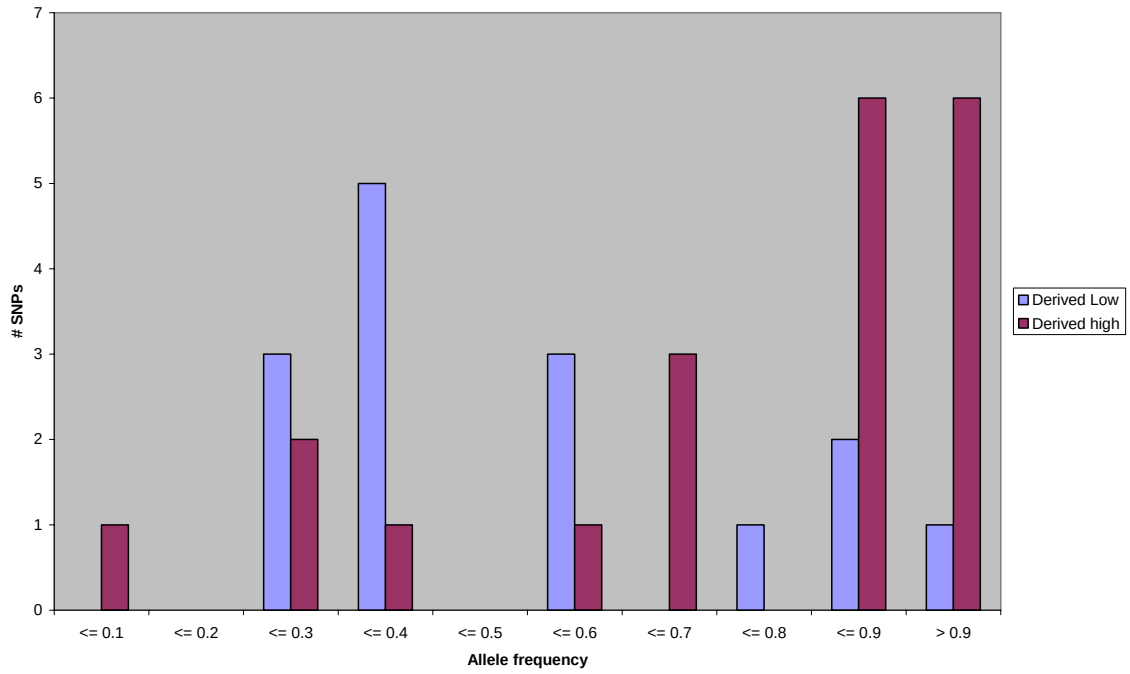


Figure 34. Allele frequencies of low- and high-activity derived alleles. The distribution of derived high-expression alleles (upregulatory mutations) is skewed towards high frequencies relative to derived low-experssion alleles (downregulatory mutations).

4.3 Conclusion

A novel strategy for rapidly cloning polymorphisms from promoters in a massively multi-parallel manner has been developed, making use of the Gateway cloning technology available from Invitrogen. It has been applied to the creation of a substantial library of promoter haplotypes, which is a valuable resource for studying promoter regulation *in vitro*. All SNPs will be submitted to dbSNP via the ExoSeq pipeline, and the haplotypes and luciferase results will be made available as supplementary material to a publication.

71.4% of promoters tested were active in at least one cell line. This compares with 66% of putative promoters in the ENCODE regions that have been shown to be active in transient transfection assays in at least one cell line (Cooper et al. 2006). The ENCODE promoters were tested on 16 cell lines, compared to the 4 used here, so it is at first surprising that in fact a greater proportion of promoters was confirmed in this project. However, the threshold used to determine positive activity was very different, and direct comparison may not be straightforward. Cooper et al used a threshold of 3 standard deviations above the activity of a combination of 102 cloned negative control fragments, whereas the data presented in this thesis used an arbitrary activity threshold. The ENCODE promoter set also contained a large number of putative alternative promoters, which had lower rates of confirmation relative to those predicted on the basis of the most 5' possible site, which would be a better comparison with the data presented here. Cooper et al report that the proportion of active promoters based on the longest possible gene was higher than the overall proportion, although they do not state what the exact number is. In an earlier study by the same group, 90% of promoters predicted on the basis of longest available cDNAs in the mammalian gene collection (MGC) were functional (Trinklein et al. 2003). This was done using only 4 cell lines, including HeLa and HT1080 as well as HEK293 (related to but predating HEK293FT). This result is striking for a different reason; that the rate of promoter confirmation is so much higher using essentially 3 of the 4 cell lines used here, as well as only a quarter of the cell lines in the subsequent ENCODE study that showed lower rates of confirmation. However, the authors suggest that the data was biased towards highly expressed genes, as the promoters were predicted using an early version of the MGC collection. Buckland and colleagues carried out a

large-scale experimental survey of promoter variation in luciferase assays using HEK293 and TE671 cells, making this data the most useful for comparison with the work carried out in this thesis. They reported 63% and 87% of cloned promoters were active according to fold activity cutoffs of 10x and 2x background respectively. This seems to correlate well with the results obtained with a 7x cutoff, with 71% closer to the 10x value reported by Buckland et al. It also indicates that the use of two further cell lines has not added much capacity to detect promoters, perhaps because HeLa and HT1080 are not sufficiently different to HEK293FT and TE671 in terms of their expression profiles.

The extent of cell specific promoter activity was low, with only 12 promoters (14.3%) being differentially active across the 4 cell lines. This matches very well with studies of a larger promoter set carried out across a cell line panel of equal size, where 15% of promoters were found to be differentially active (Trinklein et al. 2003). Direct comparison was not possible with the Buckland data, as they do not report on overall promoter activity levels and confine their analysis to the promoters with functional SNPs. However, a crude analysis of the supplementary information accompanying the Buckland paper revealed that 185/664 haplotype clones (27.9%) were differentially active using a cutoff of 7x background. While this is a much higher rate than the one reported in this study, even using the same cell lines, it may be due to selection bias in the promoters. For example, part of the Buckland dataset consisted specifically of brain-expressed genes, which would bias the promoter set to promoters active in TE671 (a medulloblastoma line) but not in HT1080 (a fibrosarcoma line).

The presence of extensive sequence-dependent variation in promoter activity has been clearly demonstrated. This in itself is not a novel finding. Although estimates of both the proportion of functional SNPs and the number of promoters harbouring them varied, previous studies have demonstrated that a significant fraction of genes contain putative functional variation in their promoters (Rockman and Wray 2002; Buckland et al. 2005). For the purposes of comparison with previous work, functional but unconfirmed SNPs will be ignored, and only confirmed functional SNPs and indels will be considered. Using this criterion, 35 promoters both demonstrated sequence dependent promoter activity variation by ANOVA, and had at least one pair of haplotypes that were significantly different using Tukey's HSD and the criteria

described above. This is 41.7% of all promoters tested, including those that were not active. This is considerably higher than the equivalent figure of 22% found by Buckland et al (Buckland et al. 2005). Several factors may have been behind this much higher rate of functional promoter polymorphism discovery. This study tested the haplotype library against 4 different cell lines, whereas Buckland et al used only two. This is bound to increase the amount of functional variation discovered, as the context dependence of promoter function means that only a subset of functional variation is likely to be discovered in a single cell line. The chromosome 22 promoters cloned had an average of 2.9 haplotypes per promoter, compared with 2.7 for the Buckland set. This is despite the fact that the degree of polymorphism in the chromosome 22 set being 2.2 polymorphisms per promoter compared to 2.6 in the Buckland set. The difference in the number of haplotypes is probably due to the difference in the panel of individuals used for SNP detection. Buckland et al used a panel of 16 ethnically diverse individuals, while I used a larger panel of 48 individuals, but from a single Caucasian population. The admixture-like effect of using an ethnically diverse panel means that the number of haplotypes will be relatively small compared to the number of SNPs (Pritchard and Przeworski 2001). In the larger single population, the SNPs will have been segregating together for longer, and recombination will have had time to shuffle them into a larger number of haplotypes. In addition, the use of a larger panel means that there was a more extensive sampling of the haplotypes available. This allowed a higher number of possible allelic combinations to be tested in the chromosome 22 set, and a higher degree of resolution was thus achieved in the assignment of functional information to individual polymorphisms.

At the SNP level, 65/228 (28.5%) in total were involved in a functional haplotype difference, with the majority having been isolated within an otherwise homogeneous haplotype pair. If unconfirmed SNPs are removed, this becomes 52/178 (29.2%). Buckland et al reported 40 isolated functional polymorphisms out of 648 cloned, or 6.2%. If only isolated and confirmed polymorphisms from the chromosome 22 set were counted the equivalent figure is 39 / 178 (21.9%), approximately 3.5 times higher than would be expected from the Buckland data. However, previous publications containing subsets of the Buckland dataset have sometimes reported higher figures such as 18% (Buckland et al. 2004a) and 22% (Buckland et al. 2004b),

relatively closely aligned with the data produced in this project. The reason for the low rate of functional SNPs in the overall Buckland dataset is not clear, as their criteria for accepting a SNP as functionally significant have remained consistent across their published work. Although their choice of cell line has not always remained consistent, it has only varied by the replacement of TE671 for JEG-3 in one paper (Buckland et al. 2004a) and remained unchanged in the other (Buckland et al. 2004b). It would be surprising if the elimination of JEG-3 from a subset of the final published data could account for such a marked loss of functional SNPs, particularly given the relatively similar behaviours of the cell lines observed here. It may instead be a consequence of the way they selected the promoters to be tested in their dataset. This was done in a very heterogeneous manner, combining genes of clinical interest (e.g. genes involved in schizophrenia, expressed in brain), genes in defined functional classes (e.g. glutamate receptors and glutathione-S-transferases), genes clustered positionally (e.g the DiGeorge region and chromosome 21) as well as "a random selection of genes found using 'promoter' as a search term in 'Entrez'" (Buckland et al. 2005). It is possible that the role of the promoter, and hence importance of promoter polymorphism, varies depending on gene class, and that combining genes selected by function with genes selected on other criteria would bias results. Even though overall promoter polymorphism uncovered in this project was not correlated with any functional gene class, this did not test whether functional polymorphism could have varying levels of importance depending on the regulatory regime of certain gene classes. The selection of genes in this project was only as unbiased as the gene complement on chromosome 22, but analysis of the GO terms of chromosome 22 genes and 5 lists of random genes from the genome showed no detectable bias either for or against any gene class (data not shown). This data may thus present a truer picture of the role of promoter polymorphism in affecting promoter activity.

The most striking result reported by Buckland et al was a strong bias towards the transcription start site in the location of functional SNPs. They found that over 50% of functional SNPs could be found within 100 bases of the TSS (Buckland et al. 2005). This result was not reproduced here; there was no discernable bias in the location of functional SNPs. It was surprising for such a strong result to emerge from one study and not from another. One possible explanation is that Buckland et al placed a magnitude threshold for what they accepted as a functional SNP, requiring that it

cause a difference in activity of at least 1.5x in 3 biological replicates. No magnitude threshold was used for the chromosome 22 data, with the requirements being statistically reproducible changes by Tukey's HSD in the same direction in 2 biological replicates. This may have biased the Buckland dataset towards SNPs with more drastic effects on promoter activity relative to the chromosome 22 set. It is not unreasonable to propose that SNPs very near to the core promoter, and thus potentially disrupting the binding of the Pol II holoenzyme or pre-initiation complex, may be more likely to have large effects than SNPs in a more distal TFBS. Also, Buckland et al only repeated the experiments for SNPs that passed the magnitude threshold on the first attempt. The chromosome 22 data suggest that the magnitude of an expression difference is not as well reproduced as the pattern of promoter activity across haplotypes. While this does not necessarily hold true for the Buckland data, as their reporter system was very different to the dual-luciferase system used here, it does suggest that they were missing significant numbers of SNPs that showed a smaller statistically significant difference but which was not replicated. As the numbers reported in the paper are for the initial biological replicate only, it is not possible to test whether this is the case.

The presence of synergistic effects between promoter SNPs was also demonstrated, although the extent and importance of this phenomenon is not clear. In only one case (the RIBC2 promoter) was it possible to demonstrate conclusively that a pair of SNPs were both required to produce a change in promoter activity, and that each SNP on its own had no discernable effect. The fact that so much of the variation observed between haplotype pairs can be accounted for by one or more of the isolated functional SNPs suggests that the effects of functional SNPs may be more often additive than synergistic i.e. that individual functional SNPs usually exert their own unique effect irrespective of genotype of flanking SNPs. In any given case, this may be either because the TFs involved exert additive effects on transcription initiation, but are not necessarily fatal when removed, or because the SNPs cause changes in the conformation of promoter DNA whose functional effects are additive (see section 6.1).

There was no correlation between the amount of sequence divergence between a pair of haplotypes and the difference in promoter activity between them. Although a trend

can be seen by looking at a scatterplot of divergence against activity ratio for each possible haplotype pair, this was not significant. It is most likely an artefact of sampling bias due to the number of haplotype pairs available decreasing with increasing divergence. While a positive association between increased haplotype divergence and activity difference might be naively expected, to the knowledge of this author it has never been demonstrated. A similar lack of concordance between absolute promoter sequence divergence and transcription (and hence, presumably, promoter activity) has previously been reported in *Drosophila* (Brown and Feder 2005). This suggests that promoter SNP functionality is a highly context-dependent property, and that closely related promoters with mutations in key regions are more likely to have different expression levels than highly diverged promoters with mutations in functionally redundant bases. If this is the case, the data from this project suggest that the prediction of such key regions, the majority of which would presumably be binding sites, is still a difficult problem. None of the regulatory elements whose co-localisation with promoter polymorphism was examined showed any significant enrichment for functional SNPs. This suggests that the current knowledge of *cis*-regulatory elements may be insufficient to confer predictive power, at least on the scale of the *in vitro* studies carried out to date. Functional elements that relied on conservation as an important component, in this case 5x regulatory potential score (Kolbe et al. 2004), cisRED (Robertson et al. 2006) and phastcons (Siepel et al. 2005) seemed to outperform TFBS weight matrices alone. The only putative element that was structural rather than relying on binding was the quadruplex-forming sequence. Although enrichment was high relative to the two TFBS classes, the numbers were miniscule, with only 4 cloned SNPs present, 2 of which were functional. It is thus difficult to draw any conclusions about this motif type, and more targeted methods may be required to investigate its correlation with functional SNPs. Combining conservation and the presence of a putative binding site improved specificity of functional SNP prediction by between 25% and 76% (with 67% of polymorphisms proving functional compared to 53% of those in cisRED alone and 38% in phastcons alone). However, this method would only detect 9.2% of functional polymorphisms, a significant drop in sensitivity.

The unconfirmed SNPs that emerged from the haplotype cloning were neither over- nor under-represented in the functional polymorphisms, with 28% of functional SNPs

being unconfirmed compared to 25% overall. Likewise, confirmed and unconfirmed SNPs were just as likely to be functional (30% and 32% respectively). This may be evidence that many of the unconfirmed SNPs might indeed be real, as if they were errors and thus randomly distributed along the promoter, one could speculate that they would have a different representation in the functional SNP set compared to the non-functional set.

5 Microarray analysis of the transcription factor complement of transformed cell lines

5.1 Introduction

In Chapter 4, evidence was obtained of extensive sequence dependent promoter activity variation. This agrees with previous studies indicating that promoter sequence influences promoter activity *in vitro*, although the degree of that influence was found to be greater in this study. While some promoter sequence polymorphisms have a full trail of evidence linking them to *in vivo* gene expression variation (Rockman and Wray 2002; Knight 2005), it is still difficult to predict the effect of particular promoter changes in the native genomic context. Despite the association of a number of promoter polymorphisms with *in vivo* effects (Knight 2005), in the majority of cases the covariance of *in vitro* and *in vivo* expression has not been demonstrated conclusively. Indeed, the extent to which the activity of a promoter *in vitro* is indicative of the amount of gene expression level *in vivo* is still unclear. This is in part due to the number of other factors besides promoter strength that influence the quantity of mRNA produced, including chromatin state, TF background and upstream *cis*-regulatory elements (see section 1). However, most reporter studies of promoter polymorphisms that have gone on to test corresponding function *in vivo* have done this in a different system (e.g. lymphoblastoid cell or primary tissue RNA) to the one in which the reporter assays were carried out. This is probably for two main reasons; the majority of polymorphisms studied are natural and thus not present in transformed cell lines, and studies in primary human tissue carry more clinical interest. In contrast, studies of allele-specific expression using transcribed markers are usually carried out in primary tissues or lymphoblastoid cell lines, but subsequent *in vitro* reporter assays are often only carried out in specific cases. Rarely has there been any attempt to assess the TF complement of the cells in which the experiments, whether *in vitro* or *in vivo*, have been done. This could prove an important source of information for explaining the mechanistic basis of promoter SNPs. For example, a SNP in a putative TFBS is less likely to function by disrupting binding at that site if the TF that is supposed to bind there is not in fact expressed at all.

Methods for assaying the binding of proteins to DNA are not new, with EMSA being a well-established assay and ChIP-chip now becoming one of the most important genomics-scale techniques for looking at protein-DNA interactions. While EMSA is useful for detecting the binding of any TF to a target sequence, it requires a candidate

sequence for use as a probe. Where candidate binding sites are known, these probes can be short oligonucleotides that allow an experiment to identify any TFs binding to that site. Often, binding sites are not known with any confidence, and in this case larger probes are sometimes used (e.g. several hundred bases of a putative promoter). In this case, TF binding can still be assayed but the precise locations of the binding site is not possible. In contrast, ChIP-chip can be used to discover binding sites without prior knowledge of their locations, and can be applied genome-wide depending on the design of the array used. These can range from whole genome arrays to small custom-made arrays. The major limitation of ChIP-chip is the availability of a suitable antibody to the TF of interest. Such antibodies are still relatively few, and as such only a small number of factors can be readily analysed in this way. The chromatin immunoprecipitation stage of this technique requires large amounts of material and is time- and labour-intensive to perform. So while ChIP-chip is a high-throughput technique in terms of the DNA-level data produced, it is low-throughput in terms of the number of TFs that can be put through it, as well as being difficult to achieve true binding site-level resultion. With upwards of two thousand known and putative TFs in the genome, a complete picture of the TF binding landscape in a cell is unfeasible outside of a large consortium.

Despite this, knowledge of the TFs that are present in the cells in which the promoter assays were carried out can still be valuable. Where functional promoter SNPs are found in putative binding sites, the presence of that TF can be confirmed in that cell line. While this would not confirm that the binding site is biologically functional, the absence of the TF would rule it out. If the functional SNP in question was only functional in a subset of the cell lines, the presence or absence of the TF could explain this behaviour. In this chapter, the whole genome expression profiles of the four cell lines used for the promoter assays was investigated. This was done using the Affymetrix U133 Plus 2.0 oligonucleotide array, which contains 54,120 probe sets targeting the majority of known genes in the human genome. This is a rapid way to characterise the 4 cell lines in a lot of detail. The expression profiles were used to explore several fundamental questions. Firstly, if the promoter of a protein coding gene is found to be active in a certain cell using a reporter assay, does this predict whether that gene is in fact expressed in the same cell *in vivo*? This is essentially a test for the effect of taking a promoter out of its genomic context, and should produce

an interesting overview of the relative importance of the TF complement versus upstream regulators and chromatin. A related question is whether variation of promoter activities between cell lines is reflected in the differences in TF complement? This may reveal a general trend for the importance of control by the production of TFs compared to other forms of control not detectable in an expression array (such as phosphorylation of TFs). If the former is the main component of control in the set of genes under study, one might predict that comparison of TF expression would show similar relationships as comparison of the promoter activities.

Secondly, is the level of promoter activity as defined by reporter assays predictive of the *in vivo* expression level? The answer to this question is likely to vary depending on gene type. Since the sequence of the promoter is fixed, it is not able to dynamically regulate the expression level of a gene. One might predict that the expression level of housekeeping genes might be governed mainly by their promoters, whereas other genes under dynamic regulation might have their expression level governed by upstream elements under the control of post-translationally modified TFs, or by epigenetic control such as chromatin modification.

In the last chapter, no enrichment of functional SNPs in known TF binding sites (TFBSs) was detected. This is either because they caused a functional difference by some other method (e.g. a change in DNA flexibility) or they are in a binding site that is not currently known. The latter explanation is not unlikely, given that many of the binding sites in TRANSFAC and similar databases are based on the study of a relatively small number of natural binding sites, and that the activity of binding sites may be cell-type specific and only active under certain conditions. It has been proposed that the sum total of unknown binding sites is likely to consist of a larger number of rare sites rather than a smaller number of common ones (Buckland 2006). If that is the case, it is possible that more success will be had in finding an explanation for the functional SNPs discovered in this project if motifs important to the regulation of the genes in these particular cell lines are discovered *de novo* and investigated. The whole genome expression data for the cells will be used to try and discover regulatory motifs. This will be done by comparing the expression profile of each of the genes whose promoters were cloned with the profile of the other genes on the array across all 4 cell lines. For each cloned promoter gene, a list of other genes whose expression

profiles are closely correlated will be constructed. The promoters from these genes will then be recovered from the genome and subjected to a motif discovery algorithm. In theory, this should discover motifs important in the cell-specific expression differences of these genes. These motifs would then be checked to see if they are enriched for the presence of functional promoter SNPs discovered in the previous chapter. This method has been successfully applied in yeast (Roth et al. 1998; Spellman et al. 1998), although application in higher eukaryotes is sometimes more problematic due to the potential dispersion of regulatory elements at large distances from the TSS.

The aim of the work described in this chapter is essentially to gain some information on the relevance of proximal promoter strength, as defined by the reporter assays carried out in the last chapter, to *in vivo* expression of a gene from the same promoter but in the context of upstream regulatory inputs in addition to TF complement.

5.2 Results

5.2.1 Preparation and hybridisation of RNA samples from cell lines

In order to analyse the whole genome expression profiles of the cell lines used for the promoter assays, and be able to mine them for information on TF background and *in vivo* expression of genes downstream of cloned promoters, suitable RNA samples needed to be extracted from the cells. Ideally, the RNA to be used for the whole genome array experiments would be prepared from the same batch of cells as that used for the transfection experiments in chapter 4. This would minimise any biological differences between the cells in which the promoter constructs were transfected and the cells whose expression profiles were assessed. For logistical reasons, this was not possible, and RNA was prepared from different batches of cells at the same passage number. The cells from which RNA was prepared were grown to between passages 3 and 6 after thawing from liquid N_2, the same stage as those used for transfection experiments. After harvesting, RNA was prepared using the commercially-available RNeasy mini kit (QIAgen) recommended by Affymetrix for preparations that are compatible with the expression array platform. 3 different batches of each cell line were grown in separate flasks prior to RNA preparation. The corresponding 3 biological replicate RNA preparations were produced from independent cultures thawed from frozen stock on different days. RNA was prepared by following the recommended protocol from QIagen, and the purity of the samples was confirmed by OD_{260}.

The gene expression profiles of the cell lines were interrogated by hybridising the RNA to the Affymetrix U133 Plus 2.0 arrays. The prepared RNA samples were converted to cDNA by reverse-transcription, and then to biotin-labelled cRNA following all recommended protocols. This was then fragmented prior to hybridisation on the arrays. Each labelled cRNA sample was hybridised overnight on a separate array. Signal was developed by applying the fluorescent dye phycoerythrin linked to streptavidin (in order to bind the biotin in the hybridised cRNA). The signal was then amplified by applying biotin-coated anti-streptavidin antibody followed by further streptavidin-phycoerythrin.

5.2.2 Normalisation of expression data

The raw data from the U133 Plus 2.0 arrays consists of a fluorescence intensity value for each of the 50,000+ probes on the array. This alone is not informative, and must be transformed into a data set that gives one expression value per transcript per array, and these values should be comparable across arrays. Two main normalisation axes are involved in this transformation; the integration of data from individual probes into a single value for a probe set (and hence a transcript) and normalisation of these integrated intensity values across multiple arrays and/or experimental conditions, such that arrays are directly comparable. A wide variety of statistical methods have been developed to achieve this, each based on different assumptions and exploiting different properties on the arrays (Shedden et al. 2005). The choice of normalisation method is important, as this can have an effect at least as great as experimental or biological variation across arrays (Hoffmann, Seidl, and Dugas 2002).

The method used here is GC-content Robust Multi-array Analysis, or GCRMA (Wu et al. 2004). It was chosen because it is one of the best-performing methods currently available for normalising Affymetrix data (Irizarry, Wu, and Jaffee 2006). It performs significantly better than the mas5.0 algorithm provided by Affymetrix with the array platform (Harr and Schlotterer 2006). Full details of the method are available from (Wu et al. 2004). Briefly, there are three steps to the procedure; background correction, normalisation across arrays and combination of individual probe data to produce probe set-level values. Background correction is carried out using a linear model, and accounts for the sequence composition of individual probes. Crucially, it does not make use of the perfect-match and mismatched probe pairs that the Affymetrix proprietary method relies on. The intensity levels between arrays are then normalised using a quantile normalisation procedure. This normalises the peaks and widths of the distributions of the intensities in each array, rather than using a simple normalisation factor. Finally, the data from multiple probes are combined to produce a single value per probe set using a method called median polish (Wu et al. 2004).

5.2.3 Quality control of scanned arrays

The first step in the analysis of array data was to assess the quality of the arrays themselves. This included the quality of the samples and of the hybridisation

procedure. Appreciable differences in either of these factors could preclude the comparison of arrays. The data used to assess the quality and comparability of the arrays was put through the background correction and quantile normalisation steps of the GCRMA method, but was then analysed at individual probe level rather than probe set level. These analyses were carried out in collaboration with Juanma Vaquerizas at the European Bioinformatics Institute.

The OD_{260} characteristics of the original and fragmented samples give information on the presence of contaminants, but not on the integrity of the RNA itself. RNA is prone to degradation during preparation, manipulation and storage, particularly if samples are contaminated with RNAses from the laboratory environment. RNA integrity can be assessed pre-hybridisation using a bioanalyzer, but this device was not available. The degree of degradation was therefore assessed post-hybridisation by examining the mean intensities of the individual probes in each probe set on the array as a function of their location along the length of the transcript. The reverse transcription reaction that generates the cDNA during sample preparation is primed with an oligo-dT primer from the 3' end of the transcript. It would therefore be expected that the 3'-most probes would on average have the highest relative intensities, and that the intensity would decay towards the 5' end as a function of the degree of RNA degradation. This was the case of 11 of the 12 arrays analysed (Figure 35a). The first replicate of HEK293FT showed a far greater degree of degradation, as evidenced by a flat intensity profile across probes.

The arrays were also tested for hybridisation anomalies by comparing the distributions of the logarithms of the intensities. A well-hybridised array should have a smooth, tight profile with a single peak. Bimodal or multi-modal distributions are indicative of non-uniform hybridisation on the arrays, and can preclude cross-array comparison. All arrays hybridised showed the expected histogram shape. However, the peak for the first replicate of HEK293FT was shifted noticeably to the right compared to the other arrays, which were all tightly clustered (Figure 35b). This shows that the array for HEK293FT replicate 1 is brighter than the other arrays. This would be caused by a variety of factors including too much RNA loading on the array or a difference in the labelling efficiency of the sample, although in this case it may

be related to the evidence of poor sample quality seen in the degradation plot (Figure 35a).

The relative log expression (RLE) of each array was then analysed. This is a measure of the intensity distribution relative to the median peak of all the arrays in the experiment. RLE was visualised with a box plot showing the median and interquantile range (the range of intensities between the 25^{th} and 75^{th} percentile) of each array (Figure 35c). Again, the first HEK293FT replicate was anomalous, showing an intensity distribution that was biased relative to the median. All other arrays had similar distributions, as evidenced by the closeness of the medians to 0 and the small inter-quantile ranges.

Finally, the normalised unscaled standard error (NUSE) for each array was plotted in a similar box-plot. NUSE is a measure of the standard error during the background correction process (Figure 35d). HEK293FT replicate 1 had a higher error associated with background correction, suggesting that the signal-noise ratio is lower than the other arrays. It also had a higher degree of variation associated with that error, as evidenced by the larger interquantile range.

Following these quality assessments, it was decided that the first replicate of the HEK293FT cell line would not be used in the analysis. This is because of evidence that the RNA sample used suffered degradation as well as marked differences in the distribution of signal intensities and NUSE that suggest this array is not directly comparable to the others in this set. Including this array could result in spurious gene expression changes being detected that are caused by these non-biological factors.

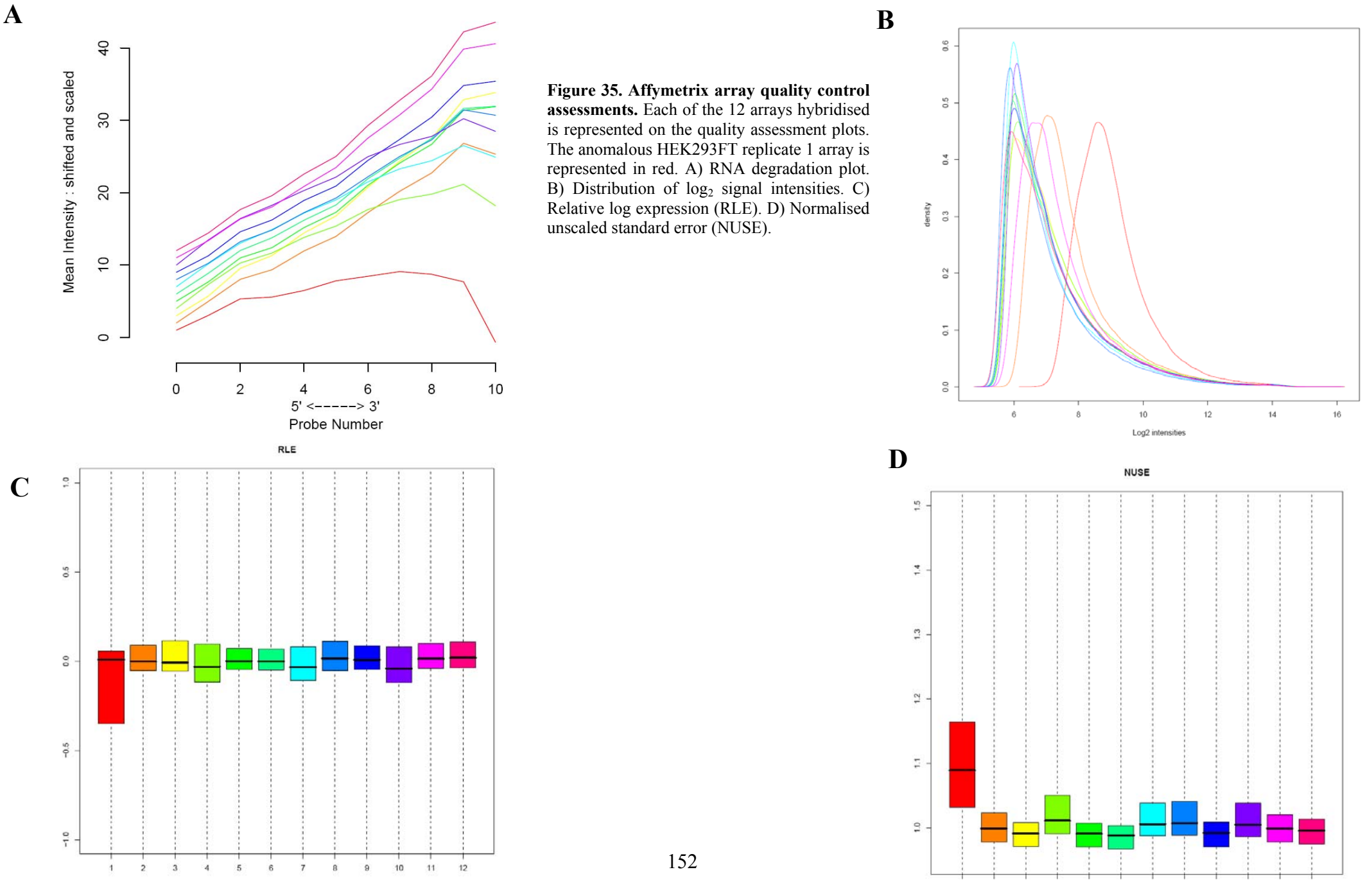


Figure 35. Affymetrix array quality control assessments. Each of the 12 arrays hybridised is represented on the quality assessment plots. The anomalous HEK293FT replicate 1 array is represented in red. A) RNA degradation plot. B) Distribution of $\log_2$ signal intensities. C) Relative log expression (RLE). D) Normalised unscaled standard error (NUSE).

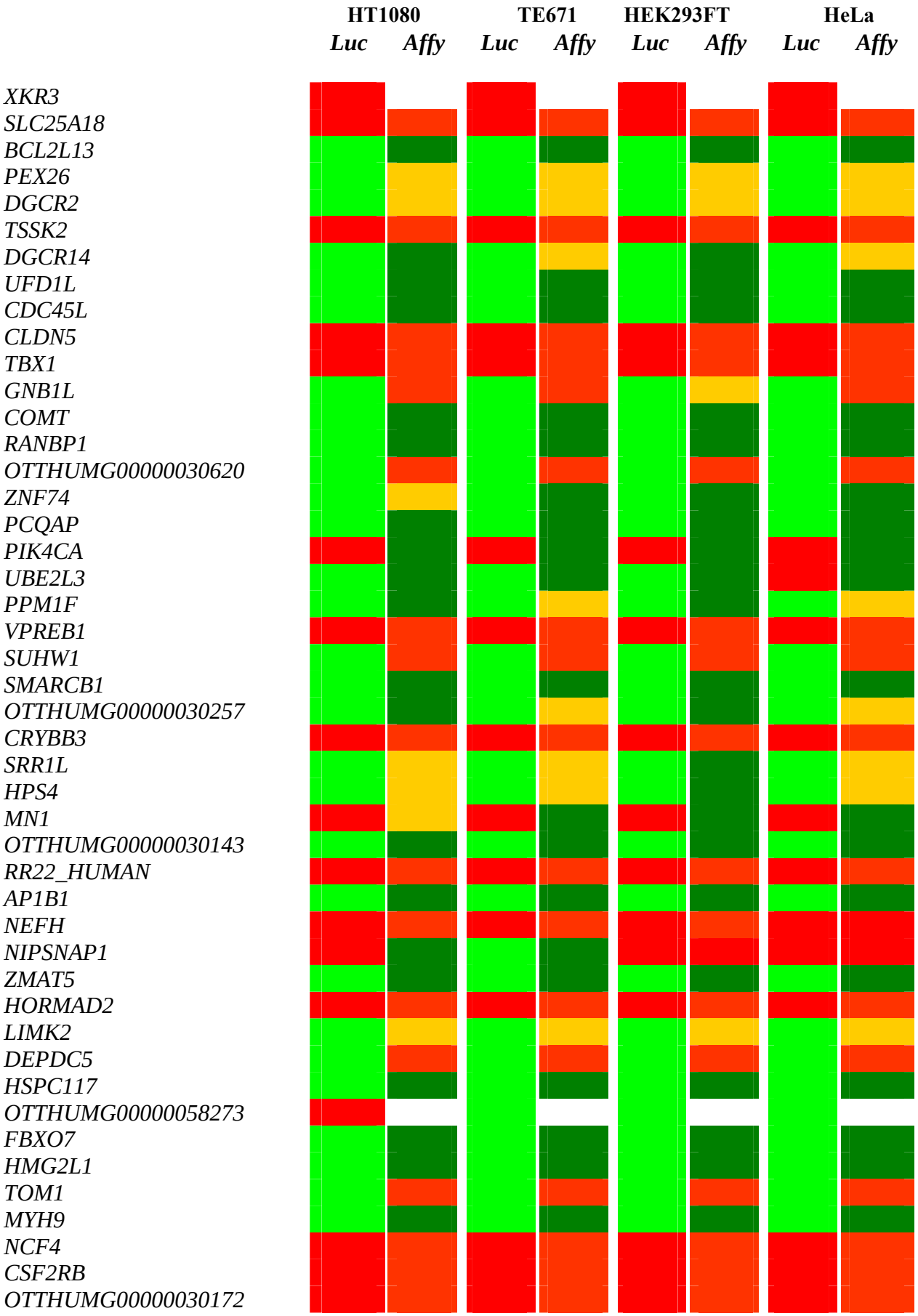
5.2.4 Comparison of endogenous gene expression with cloned promoter activity

As the U133 Plus 2.0 arrays cover the whole genome, the majority of the genes whose promoters have been analysed in reporter assays are likely to be represented on the array. The expression level of the genes could therefore be compared to the activity of the promoters in the same cell lines. This would give some information about the degree to which *in vitro* promoter activity is predictive of *in vivo* gene expression. The probe sets associated with each gene for which a promoter had been cloned were identified using Ensembl BioMart. At least one probe set was identified for 77 of the 84 genes.

In the last chapter, a 7x threshold over background activity was used to determine whether a promoter was active or not. In order to make a comparison with *in vivo* gene expression, a similar yes/no expression call was required for the array data. The most common method has been the proprietary mas-P/A method developed by Affymetrix. This subtracts the mismatch probe signal from each corresponding perfect match probe, and then uses statistics based on the t-test to determine whether the transcript represented by that probe set is present or absent. In practice, these calls are highly unreliable as the mismatch probe signals are often above the true background level. A second method called PANP was used in this study (Warren et al. 2006). Instead of the mismatch probes, this method exploits a group of probes that has been identified by Affymetrix as being designed from transcripts that were incorrectly annotated on the reverse strand to the one from which they are really transcribed. As such, they are antisense to any known transcripts and should in theory give a true representation of background signal. The GCRMA-normalised expression from the 11 arrays that passed the quality control steps were subjected to the PANP algorithm. This returned a single call per probe set per array that designated that transcript as either present, marginal or absent. These calls were produced by computing a gene expression level above which a probe set could be designated marginal or present at p-values of 0.02 and 0.01 respectively. These thresholds were specific to each array. Where a gene was represented by multiple probe sets, a single call was ascertained by applying the thresholds to the median of all probe sets. The

calls from the replicate arrays for each cell line were combined by simply accepting the call that was most frequent in the set of arrays.

The expression status of each gene was compared to the activity of the equivalent promoter in the luciferase reporter assays. The two data sources were deemed to match if the promoter was inactive and the gene was called absent, or the promoter was active and the gene was called present. Marginal calls were deemed to be compatible with both active and inactive promoters, and were thus called as matches regardless of promoter state. Using these criteria, 240/308 (78%) of the gene expression calls matched the activity designation of the respective promoters (Table 12). Of the 68 that did not match, 44 were instances of active promoters whose genes were called absent in the arrays, and 24 were of inactive promoters whose genes were called as present.


HT1080
TE671
HEK293FT
HeLa
Luc
Affy
Luc
Affy
Luc
Affy
Luc
Affy
XKR3
SLC25A18
BCL2L13
PEX26
DGCR2
TSSK2
DGCR14
UFD1L
CDC45L
CLDN5
TBX1
GNB1L
COMT
RANBP1
OTTHUMG00000030620
ZNF74
PCQAP
PIK4CA
UBE2L3
PPM1F
VPREB1
SUHW1
SMARCB1
OTTHUMG00000030257
CRYBB3
SRR1L
HPS4
MN1
OTTHUMG00000030143
RR22_HUMAN
AP1B1
NEFH
NIPSNAP1
ZMAT5
HORMAD2
LIMK2
DEPDC5
HSPC117
OTTHUMG00000058273
FBXO7
HMG2L1
TOM1
MYH9
NCF4
CSF2RB
OTTHUMG00000030172

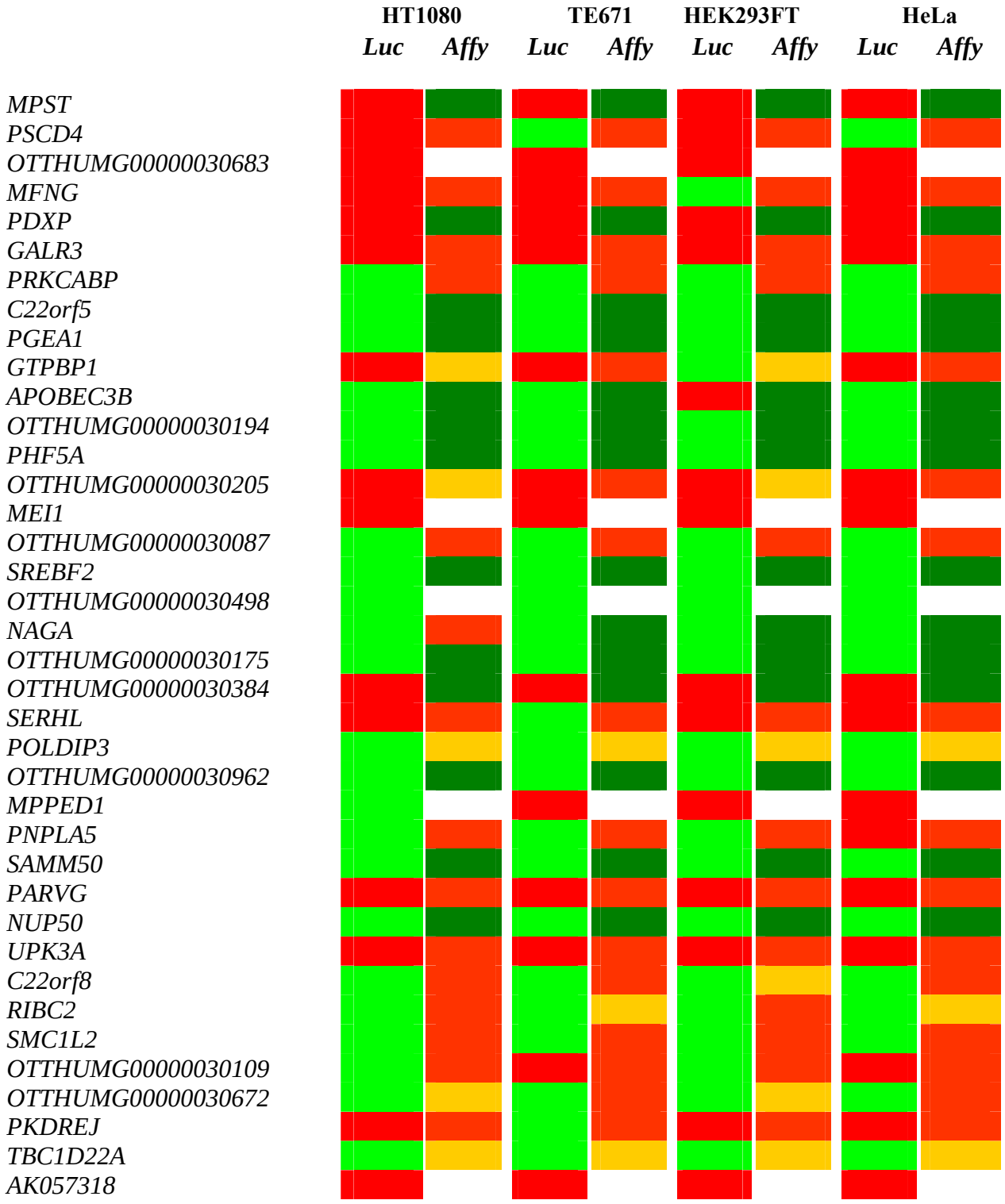


Table 12. Concordance of promoter activity and gene expression for tested promoters in 4 cell lines. Active and inactive promoters in each cell line are designated by green and red shading respectively. The consensus gene expression call is shown next to the promoter activity information in a slightly different colour scheme (P = dark green, M = yellow, A = pale red). Where a gene had no probes on the array, no shading is shown. Promoters are listed in the order of their occurrence along chromosome 22 from centromeric to telomeric ends of the q arm.

The effect of changing the 7x promoter activity threshold on the correlation with gene expression was examined. The numbers of matching calls between the two data sources was counted for activity thresholds between 5x and 9x, and the mismatches further classified into active promoters called absent and inactive promoters called present. While there is some fluctuation in the number of mismatches, changing the activity threshold does not seem to affect this in a linear way (Figure 36). As would be expected, there is a small but observable increase in the number of present/inactive mismatches and a corresponding decrease in the absent/active mismatches as the activity threshold is raised. These changes are small, with only 7 mismatches difference between the highest and smallest number in both categories, just 2.2% of the total number of gene/promoter pairs. This suggests that the mismatches are caused by a disregulation of the cloned promoters as a result of being taken out of their *in vivo* environment, rather than an artefact of the placement of the activity threshold.

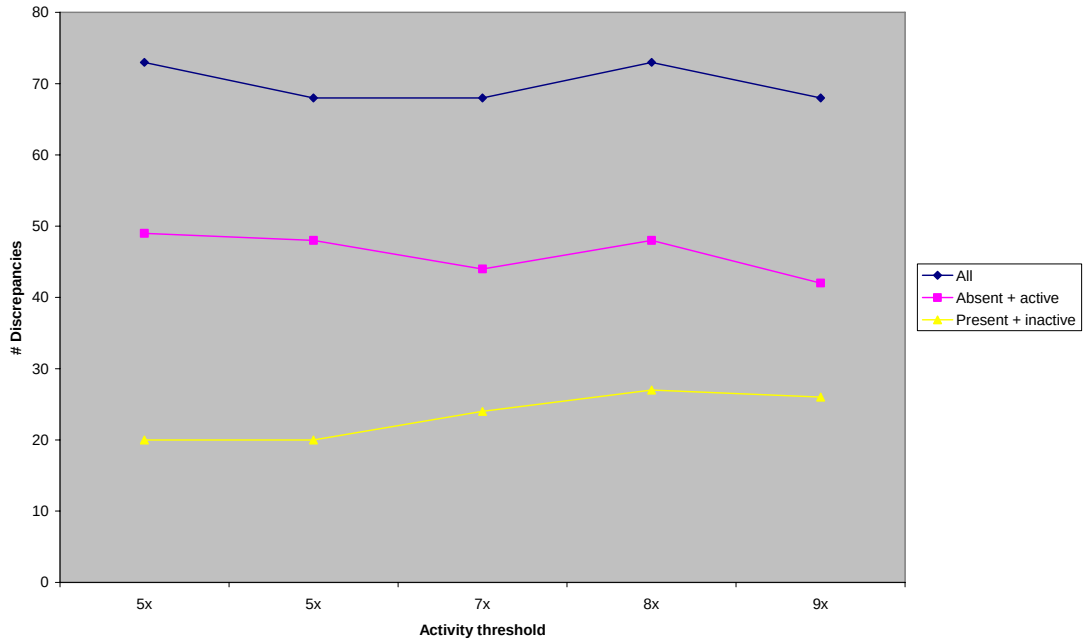


Figure 36. Relationship of promoter activity threshold to the number and type of mismatched calls between the luciferase and expression data. The number of instances where the presence, marginal or absent calls matched what would be expected from the promoter activity (y axis) was examined as a function of the promoter activity threshold (x axis)

In vivo expression of the genes was also compared to the level of promoter activity rather than a binary active/inactive call. The luciferase value for the highest activity haplotype in each cell line was plotted against the median expression level of all

probe sets in the arrays for the same cell line (Figure 37a). In theory, given that promoter strength is positively correlated with gene expression, one would expect a linear relationship to be visible on the plot. Such a relationship is not immediately apparent, with a wide range of promoter activities being found at all gene expression levels. However, there is a higher frequency of low luciferase values the lower the expression level of the gene. This is visible as a distinct peak at the low end of the distribution of luciferase activities for genes called as absent, with much smaller peaks for the marginal and present genes (Figure 37b). The median promoter activity for absent genes is 2.96, well below the 7x activity threshold. In contrast, expressed genes had a median promoter activity of 30.8 (Figure 37c). This is difference is highly significant ($p < 2.2 \times 10^{-16}$ by Mann-Whitney test). Interestingly, the equivalent value for genes called as marginal in the arrays was 60.2, twice as high as the value for present genes (Figure 37). The difference between the present and marginal promoter activities is also significant ($p = 3.76 \times 10^{-6}$ by Mann-Whitney test). Whether this observation is biologically relevant is not immediately clear, as there are relatively few marginal calls compared to present and absent. It can be hypothesised that more of this set of genes are regulated by negative upstream or *trans*-acting regulatory elements *in vivo* than by positive elements. This may also explain the high number of absent genes with active promoters compared to present genes with inactive promoters.

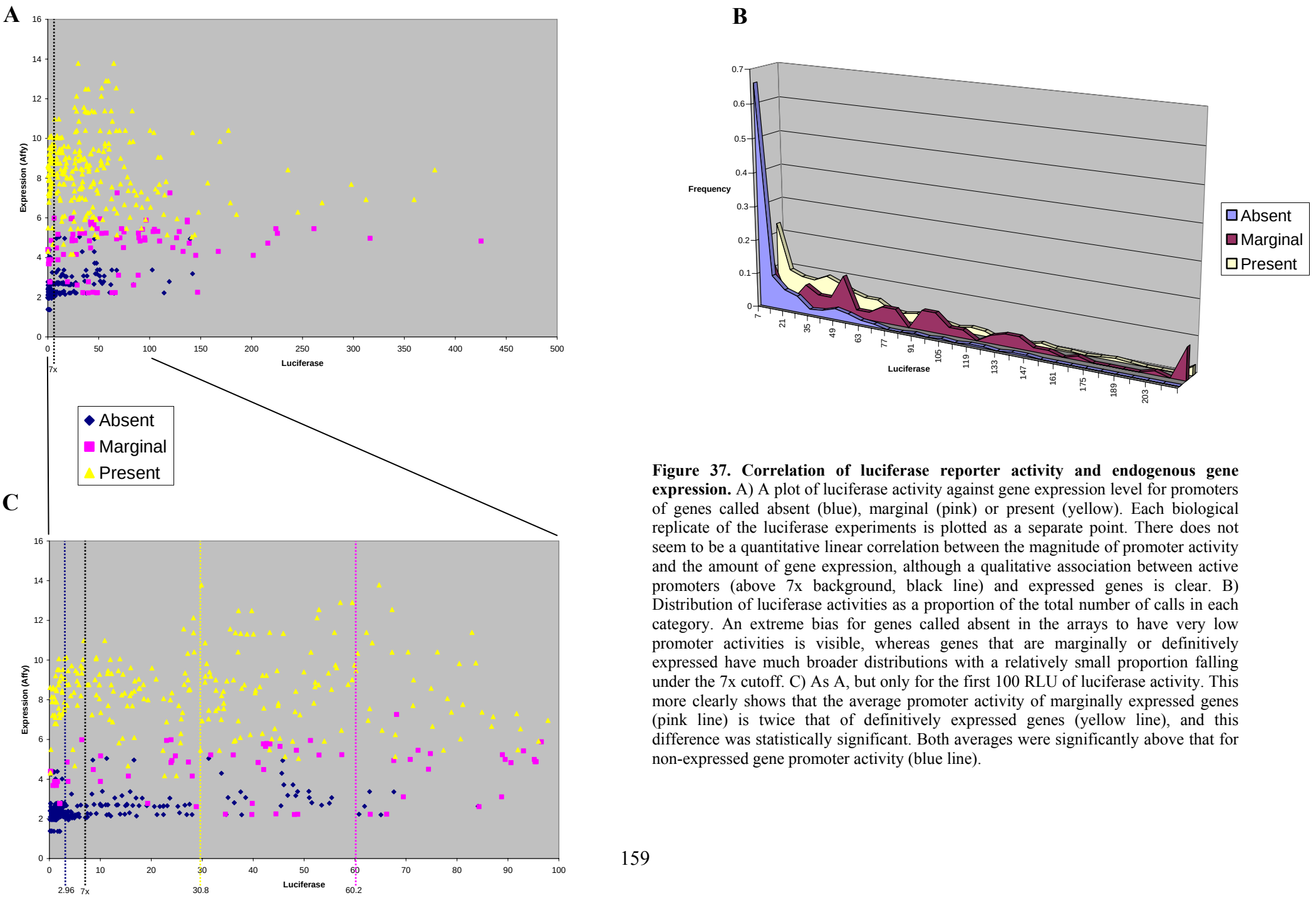


Figure 37. Correlation of luciferase reporter activity and endogenous gene expression. A) A plot of luciferase activity against gene expression level for promoters of genes called absent (blue), marginal (pink) or present (yellow). Each biological replicate of the luciferase experiments is plotted as a separate point. There does not seem to be a quantitative linear correlation between the magnitude of promoter activity and the amount of gene expression, although a qualitative association between active promoters (above 7x background, black line) and expressed genes is clear. B) Distribution of luciferase activities as a proportion of the total number of calls in each category. An extreme bias for genes called absent in the arrays to have very low promoter activities is visible, whereas genes that are marginally or definitively expressed have much broader distributions with a relatively small proportion falling under the 7x cutoff. C) As A, but only for the first 100 RLU of luciferase activity. This more clearly shows that the average promoter activity of marginally expressed genes (pink line) is twice that of definitively expressed genes (yellow line), and this difference was statistically significant. Both averages were significantly above that for non-expressed gene promoter activity (blue line).

5.2.5 Correlation of binding sites at functional SNPs with transcription factor expression

26 of the functional promoter SNPs discovered in chapter 4 were located within a putative TFBS, whether defined by TRANSFAC or JASPAR. While this suggested that they functioned by interfering with the binding of the associated TF, this could not be confirmed without separate experiments such as EMSA or ChIP-chip. The opportunity to do these studies for the 26 SNPs did not arise over the course of this project. However, with whole genome array data for the cell lines available, it was at least possible to determine whether the TFs in question were expressed, and whether differential expression in these factors could in any way account for any cell-type specific functional differences in these SNPs. The first step was to generate presence / marginal / absence calls for all TFs in the genome, and then determine whether they are differentially expressed in the cell lines. The calls were generated with the same PANP algorithm as was used above (Warren et al. 2006). Differential expression was analysed by applying the LIMMA linear modelling algorithm included in the Bioconductor analysis package on the GCRMA-normalised data for the whole genome arrays. This integrated the expression levels from the replicate arrays for each cell line into a single expression measurement, assessed the significance of expression differences for each probe set between pairs of cell lines, and generated a p-value for each probe set following correction for multiple testing using the false discovery rate method (Benjamini and Hochberg 1995).

Four of the functional SNPs were in TFs for which probes on the Affymetrix array could not be located, and they were therefore discarded from this analysis. The remaining 22 SNPs were found in a total of 39 putative binding sites, with 13 SNPs in multiple binding sites. The probe sets that mapped to the genes for the TFs with binding sites around the SNPs were identified using the Ensembl BioMart tool. Any probe sets with a _x_ designation, signifying potential cross-hybridisation to multiple genes were discarded. The exception was the ELK1 TF gene, for which the only two available probe sets carried that designation. Both P/M/A calls (grey vs. white shading) and differential expression (as calculated by the LIMMA algorithm) were plotted together in order to better visualise the behaviour of TFs for which putative binding sites were found (Table 13).

Five SNPs were in binding sites for TFs that were called as absent in all four cell lines, suggesting that for those SNPs, the binding site was not biologically functional. One of these five SNPs was also in another binding site for a factor that was expressed. A sixth SNP was in a binding site defined by a weight matrix for the cEBP TF, of which probe sets for 3 isoforms were present on the array. One of these, cEBPE, was called absent across all cell lines, whereas the other two, cEBPB and cEBPG were both present. For the purposes of this analysis, cEBPB was used as the probe, as it was the only one differentially expressed.

These 21 SNPs were in a total of 28 putative binding sites, with 7 polymorphisms found in binding sites for two different TFs. Overall, there were 8 instances were the TF was expressed at least in all cells in which the polymorphism was functional. This evidence would be consistent with a role for that TF in the mechanism of the polymorphism, although it is not conclusive evidence on its own. For 14 binding sites, the TF was called absent in at least one cell for which a functional effect was observed, apparently ruling out TF binding as the mechanism for the polymorphism. In the final 6 cases, there was a degree of ambiguity due to the presence of multiple probe sets, where one showed consistency and another did not. Nothing could be said about consistency in these cases.

14 SNPs were in binding sites for which the TF was differentially expressed in at least one pair of cell lines (Table 13). This included one SNP that was in two binding sites for which the factors were differentially expressed. Of these, however, only two TFs had an expression profile that could account for the function of the SNP. These were a C/G SNP in the *CDC45L* promoter that was located in a REL binding site, and a C/A SNP in the *RBIC2* promoter that was within a CREB binding site.

Promoter	SNP	Alleles	Motif	Probe Set	HT1080	TE671	HEK293T	HeLa
DGCR14	295	C/T	ZNF42_5-13	40569_at				F
DGCR14	300	T/A	ZNF42_5-13	40569_at				F
DGCR14	300	T/A	Mycn	209756_s_at				F
DGCR14	300	T/A	Mycn	209757_s_at				F
CDC45L	381	C/G	REL	206036_s_at				F
OTTHUMG00000030620	184	G/A	ZNF42_5-13	40569_at		F		
SUHW1	471	A/T	cEBPB	212501_at				F
NIPSNAP1	259	T/G	Mycn	209756_s_at		F	F	F
NIPSNAP1	259	T/G	Mycn	209757_s_at		F	F	F
DEPDC5	305	G/C	ELK1	203617_x_at				F
DEPDC5	305	G/C	ELK1	210376_x_at				F
FBXO7	172	C/-	SP1	214732_at	F	F	F	F
FBXO7	172	C/-	SP1	224754_at	F	F	F	F
FBXO7	172	C/-	REL	206036_s_at	F	F	F	F
PSCD4	419	[GTTT]n	FOXI1	208006_at		F		
PSCD4	419	[GTTT]n	Foxa2	40284_at		F		
PSCD4	419	[GTTT]n	Foxa2	210103_s_at		F		
PGEA1	8	C/T	ELK1	203617_x_at				F
PGEA1	8	C/T	ELK1	210376_x_at				F
GTPBP1	136	C/G	Fos	209189_at		F	F	F
GTPBP1	150	C/T	ELK1	203617_x_at		F	F	F
GTPBP1	150	C/T	Myb	204798_at		F	F	F
GTPBP1	150	C/T	ELK1	210376_x_at		F	F	F
APOBEC3B	521	T/C	RORA	240951_at	F	F	F	F
APOBEC3B	521	T/C	RORA	210479_s_at	F	F	F	F
OTTHUMG00000030087	602	C/G	ELK1	203617_x_at	F	F	F	F
OTTHUMG00000030087	602	C/G	ELK1	210376_x_at	F	F	F	F
OTTHUMG00000030087	602	C/G	Myb	204798_at	F	F	F	F

Promoter	SNP	Alleles	Motif	Probe Set	HT1080	TE671	HEK293T	HeLa
SERHL	45	G/A	SP1	214732_at	F	F	F	F
SERHL	45	G/A	SP1	224754_at	F	F	F	F
POLDIP3	78	G/A	Fos	209189_at	F			
POLDIP3	78	G/A	V$PAX6_01	235795_at	F			
NUP50	371	G/C	MAX	209332_s_at	F	F	F	F
NUP50	371	G/C	USF1	231768_at	F	F	F	F
C22orf8	77	A/T	HAND1-TCF3	220138_at	F	F	F	F
SMC1L2	422	G/T	CREB1	237289_at		F	F	
RIBC2	554	C/A	CREB1	237289_at	F	F	F	
OTTHUMG00000030109	528	C/T	ELK1	203617_x_at		F	F	F
OTTHUMG00000030109	528	C/T	ELK1	210376_x_at		F	F	F

Table 13. Differential expression of transcription factors with binding sites around functional SNPs. Each instance of a SNP within a binding site for which a probe set was available on the Affymetrix array is shown as a separate line. Where a TF is represented by more than one probe set, each one is included as a separate line. Grey shading indicates that the probe set was called absent, whereas unshaded cells are where probe sets were called present. Green shading indicates that the probe set was upregulated in that cell line, whereas red shading indicates downregulation of a probe set that was called present. The latter two designations are based on pairwise comparisons of all cell lines using GCRMA-normalised data processed through the LIMMA linear modelling algorithm. "F" indicates that the SNP was functional in that cell line. Cells lacking an "F" show cell lines where the SNP was not functional. Note that one of the ELK1 probe sets was called absent in TE671 despite no statistically significant differential expression versus any other cell line. This is because in all other cells the call was marginal rather than present, indicating that the difference was small.

5.2.6 Classification of cell lines by promoter activity and gene expression

Comparing the binary active/inactive promoter calls across the cell lines, there is a high degree of agreement across all 4 lines. In order to determine how different the cells are in the way they respond to the cloned promoter library, the correlation coefficient was calculated between all luciferase values for each possible pair of cell lines (Table 14). The median value between the two biological replicates was used for these calculations. This showed that HT1080 was the cell line that was most different from all the others, with correlations between 0.14 and 0.18. HEK293FT was approximately as different from HeLa as from TE671, but the latter two cell lines were more diverged from each other than either was to HEK293FT. The two biological replicate datasets for each cell line were also correlated with each other. In 3 out of 4 cell lines, the two replicates were more closely correlated than the median of the two replicates was to any of the other cell lines. In HeLa cells, the two biological replicates were less well-correlated with each other than to HEK293FT, suggesting that there is more noise in the HeLa data.

HT1080	0.83			
TE671	0.14	0.80		
HEK293FT	0.15	0.68	0.70	
HeLa	0.18	0.49	0.62	0.55
	HT1080	TE671	HEK293FT	HeLa

Table 14. Correlation between promoter activities in the 4 cell lines. Correlations within cell lines were calculated between the two biological replicates. For between-cell line correlation, the medians of the two biological replicates for each haplotype were used.

If the activity of the transfected promoter constructs was purely a function of the TF complement of the transfected cells, one could hypothesise that the overall differences in the behaviour of the cloned promoters will be proportional to the differences in the TFs present in each cell. The differences between cell lines were evaluated globally using the correlations calculated above. In order to compare these with the corresponding differences between the cell lines in terms of the expression of TFs, the cells were classified according to how different the TF expression profiles were from

each other, using the array data to assess TF expression. Genes were identified as TFs according to a manually refined and curated list of the contents of the DBD TF database (Kummerfeld and Teichmann 2006). Overall correlation coefficients between cell line pairs were calculated based on the GCRMA-normalised expression values for the cloned promoter genes and for all TFs separately. The Affymetrix probe sets on the U133 Plus 2.0 array that corresponded to TF genes and to cloned promoter genes were extracted from Ensembl using the BioMart tool. Any probes that cross-hybridised to multiple transcripts (designated by a _x_ code in the probe name) were removed. This analysis showed much smaller distances between the cell lines than suggested by the correlations between the *in vitro* promoter activities (Figure 38). In addition, HT1080 was not significantly more different than any other cell lines, as was found using the promoter activities.

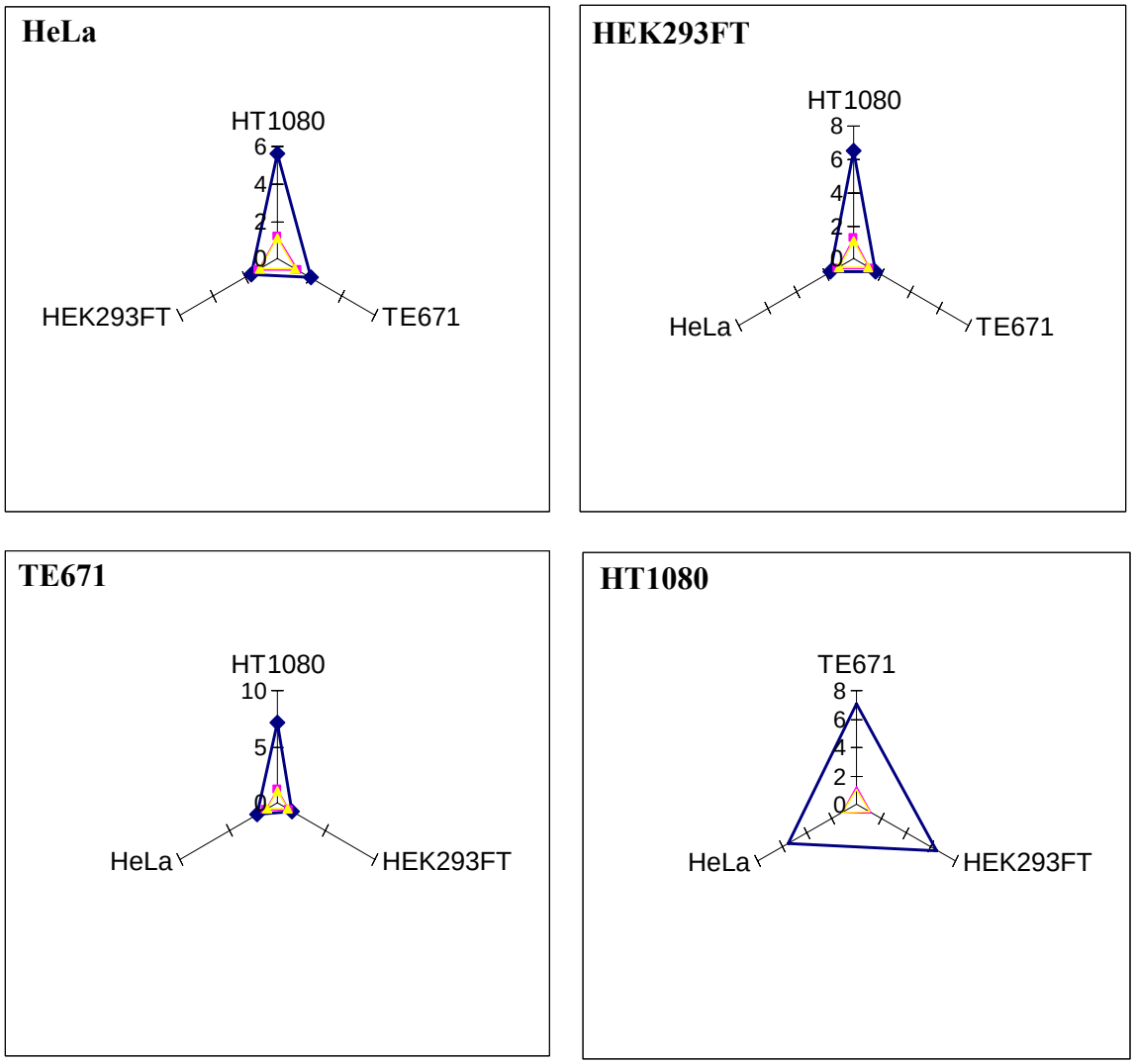


Figure 38. Distances between the 4 cell lines according to the overall activity / expression profiles of cloned promoter constructs (blue), transcription factors (pink) and cloned promoter genes (yellow). Each of the four panels compares one cell line (in bold) to the three others, showing how close it is to each of them. Distances between cell lines are plotted as the reciprocal of the correlation coefficient for each cell line pair for promoter activities (Table 14), endogenous expression of the cloned promoter genes and expression of TFs. The latter two correlations were computed from the GCRMA-normalised microarray data in Bioconductor.

5.2.7 Search for regulatory elements active across the 4 cell lines

It was previously shown in chapter 4 that current models of regulatory elements are poor predictors of functional promoter sequence variation. In terms of TFBSs, one of the reasons for this poor performance may be that many of the motifs in the various TFBS databases are constructed from relatively few sequences tested in a limited range of conditions. It is possible that better results would be obtained by carrying out *de novo* motif prediction for any set of conditions for which regulatory variation is to be predicted. The whole genome expression data can be exploited for this purpose by

identifying genes whose expression profile across the 4 cell lines closely correlates with that of the genes whose promoters were tested. The hypothesis is that if a set of genes have similar expression profiles in a set of multiple conditions, this is because they are reacting in the same manner to the TF complements they are being placed in. Therefore they might share common regulatory elements to which these factors bind. The idea behind this method is relatively well-established, and has been used previously to look for regulatory elements in co-regulated genes in yeast (see section 5.1).

The clustering of the expression data and identification of co-regulated genes was carried out by Robert Andrews and Gregory Lefebvre at the Sanger Institute. The GCRMA-normalised whole genome data was processed through LIMMA to integrate the biological replicates into one value per cell line, and the data was then clustered into a tree using XCluster (Gavin Sherlock). This uses the hierarchical clustering method Average Linkage (Eisen et al. 1998), which builds a single tree of all the genes by calculating the distance between each possible pair of genes, and iteratively joining the closest pair at a node. The concordance between the expression profiles of the cloned promoter genes and each of the remaining genes on the array was assessed independently and assigned a score. This was done by successively partitioning the tree 1000 times using the R statistical package, starting at the root of the tree and moving down. The number of partitions where the cloned promoter gene and the probe set being compared to it segregate together was counted and assigned as the concordance score. The number of partitions before the two are separated on the tree was assigned as the score. This process is explained diagrammatically in Figure 39.

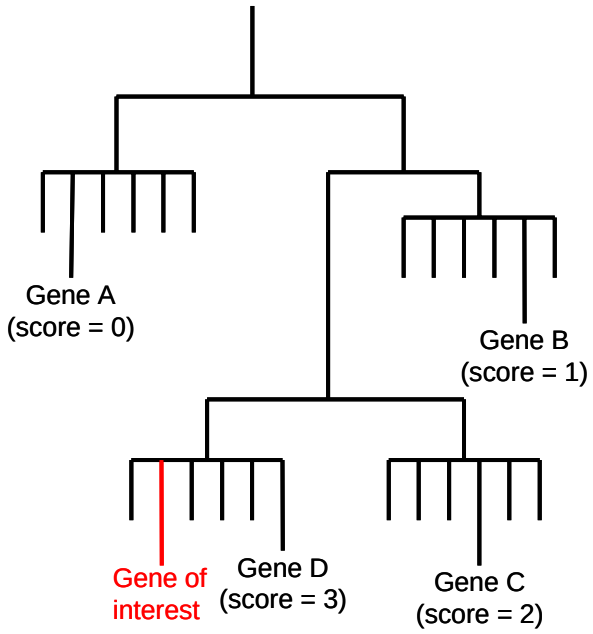


Figure 39. Simplified tree showing the scoring system used to identify co-regulated genes. For a particular gene of interest, the co-expression of each other gene on the array is calculated as the number of partitions on the tree for which the two genes segregate together. 4 genes are highlighted on this tree with the scores they would be assigned in each case. Gene D segregates with the gene of interest through 3 partitions, and is thus given a score of three. Genes C, B and A all separate from the gene of interest earlier, and are assigned scores accordingly.

For each probe set representing one of the cloned promoter genes, all other probe sets with scores above 500 (i.e. which segregated together for at least 500 partitionings of the tree) were considered to be co-regulated. Where the cloned genes were represented by multiple probe sets, the union of these sets was taken as the co-regulated cluster. The genes mapped to the probe sets in each cluster were identified through Ensembl using the BioMart tool. At this stage, around 50% of all probe sets in the clusters failed to match an Ensembl gene. This is because Ensembl apply more stringent criteria for mapping Affymetrix probes to the genome than Affymetrix themselves, and many probe sets were not considered reliable enough to map to a gene. Of the 77 promoters with at least one Affymetrix probe set, 5 did not cluster with any other genes at a score above the threshold, and were therefore discarded from the analysis. The majority of the remaining cloned promoter genes clustered with between 50 and 125 other genes (Figure 40).

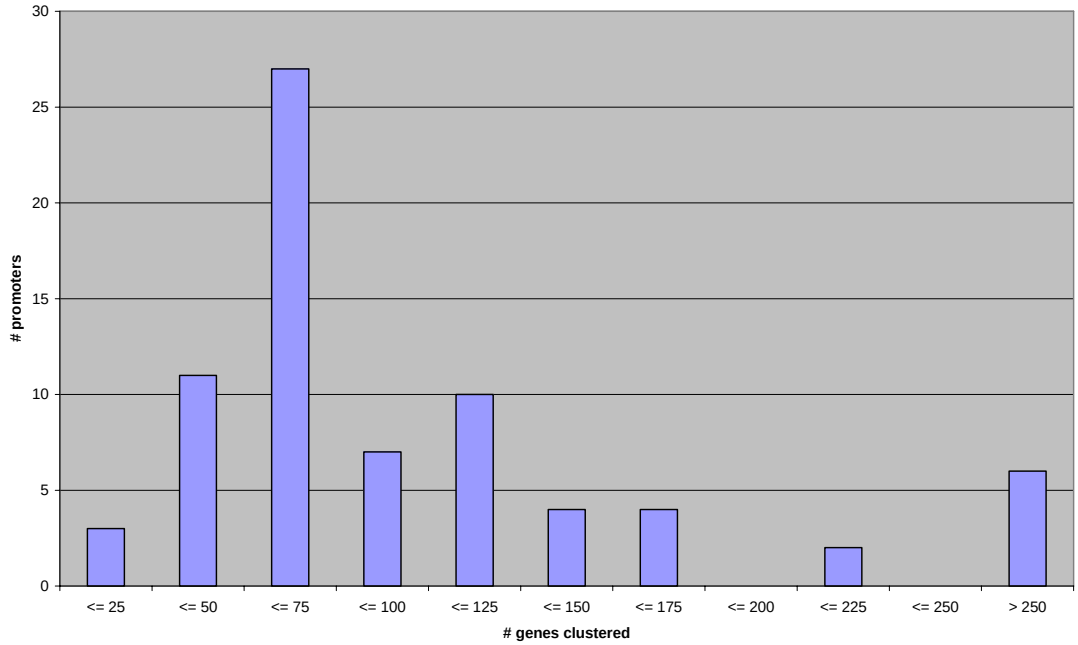


Figure 40. Number of genes clustered with the cloned promoters. The majority of cloned promoter genes clustered with between 50 and 125 other genes.

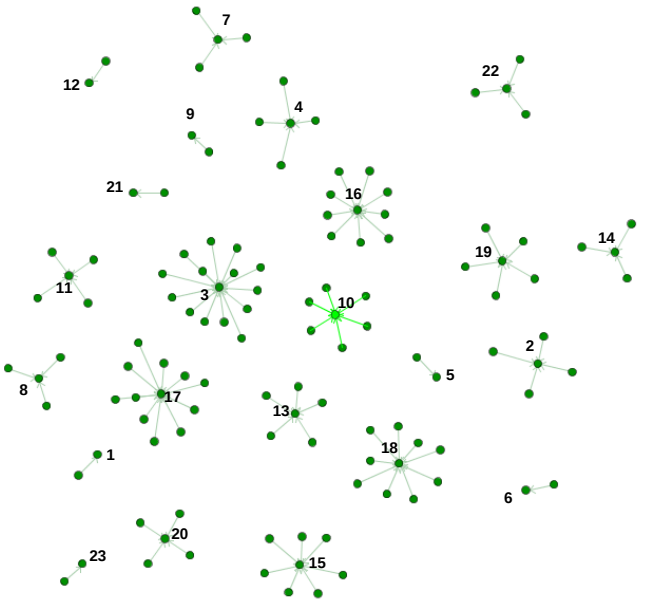
The sequences between 600 base pairs upstream and 100 base pairs downstream of the TSS's of the genes in each cluster were extracted from Ensembl using BioMart. The program nestedMICA (Down and Hubbard 2005) was used to look for motifs within each cluster separately. nestedMICA functions by analysing each set of sequences in terms of a model of significant motifs in a background of noise, and essentially outputs significantly overrepresented motifs in the form of a position weight matrix. In total, 320 motifs were discovered by nestedMICA. However, not all the motifs necessarily occurred in the tested promoters, as there is no requirement for a motif to be present in all genes in a cluster. The cloned promoters were therefore scanned for the presence of the motifs using the program MotifScanner (Aerts et al. 2003). 167/320 motifs were found to match the 72 cloned promoters in a total of 359 separate sites. These sites were then tested using the same method as in section 4.2.15 to see whether there is an enrichment of functional SNPs within these novel motifs. 161 of the 228 cloned polymorphisms were present in the promoters for which motifs were generated, including 45 of the functional polymorphisms discovered in chapter 4. 20/161 (12%) of all cloned polymorphisms were present in at least one of the generated motifs compared to 5/45 (11%) of functional polymorphisms. There is thus

no enrichment for functional SNPs in these motifs, in line with similar analyses in known TFBS and other putative regulatory elements (see section 4.2.15).

The novel motifs were also compared to known TFBS weight matrices using the MotifExplorer tool (Down et al, unpublished) to run a comparison with a downloaded copy of the JASPAR database. Using a threshold score 2 or under (the scoring system in MotifExplorer uses a distance metric to score the cumulative difference between the motifs at each base, with lower scores indicating more similarity than higher scores), 101 of the motifs showed similarity to 23 JASPAR binding site matrix, and these groups of motifs could be visualised using the BioLayout network visualisation tool (Enright and Ouzounis 2001) (Figure 41a). 6 TFs matched only one motif from one gene cluster (Arnt, En1, FOXI1, IRF1, MAX, YY1 and ZNF42_5-13), while others matched a number of motifs from different clusters. The highest number of occurrences were for motifs resembling RUSH1-alfa (9 matches), SPI1 (9 matches), SP1 (11 matches) and c-ETS (14 matches). The fact that 32% of motifs showed similarity to known binding sites suggests that the process was generally producing meaningful motifs. In order to discover whether there were novel motifs that were recurring across multiple clusters, MotifExplorer was again used to compare all novel motifs with each other. Initially, all pairs of motifs that matched with a score of 2 or below were calculated, and the results plotted as a network of similarities using BioLayout. This showed no structure at all, with all motifs contained within one very large amorphous cluster and no obvious subclusters of motifs emerging. If the threshold is made more stringent, some clustering started to emerge. With a highly restrictive threshold of 0.6, several distinct clusters of motifs were detectable (Figure 41b). This included one major cluster composed almost entirely of motifs that matched the SP1 weight matrix in JASPAR, as well as a smaller cluster of 5 motifs including 4 that match JASPAR matrices. This suggests that the other clusters may also consist of meaningful motifs that have a role in regulating multiple genes in the cloned set. In total 173 motifs, slightly over half of the total, were either similar to a known binding site or highly similar to at least one other novel motif.

A

	Transcription Factor	#
1	Arnt	1
2	Arnt-Ahr	4
3	c-ETS	14
4	CREB1	4
5	En1	1
6	FOXI1	1
7	FOXL1	3
8	HAND1-TCF3	3
9	IRF1	1
10	Klf4	6
11	MafB	4
12	MAX	1
13	Myf	5
14	NHLH1	3
15	Pax2	7
16	RUSH1-alfa	9
17	SP1	11
18	SPI1	9
19	SPIB	5
20	TFAP2A	4
21	YY1	1
22	ZNF42_1-4	3
23	ZNF42_5-13	1



B

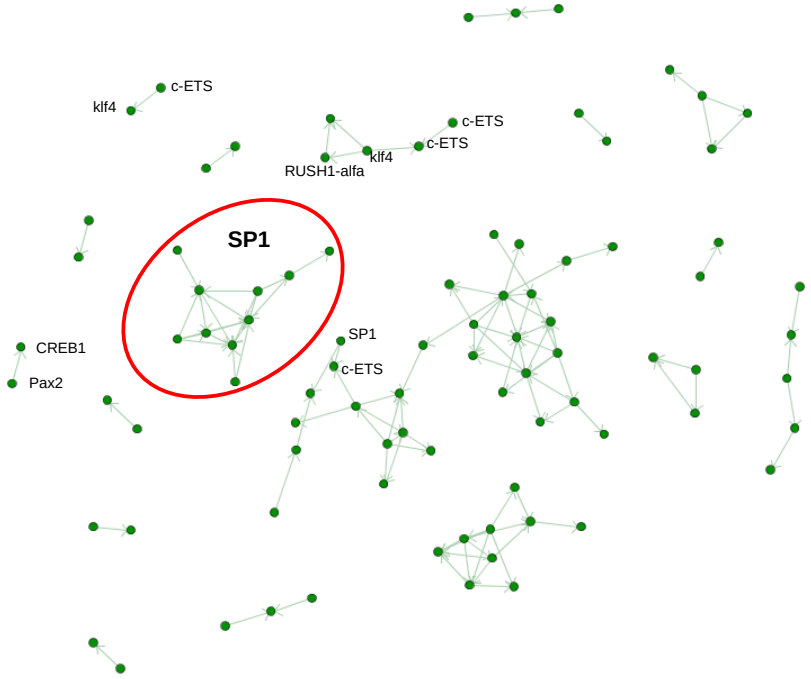


Figure 41. Comparative analysis of the motifs discovered in the clusters of co-regulated genes. Motifs are shown as green nodes joined by lines showing similarity matches. A) Motifs matching TFBS weight matrices in JASPAR with a threshold of 2 or under. 23 different weight matrices were matched to at least one of the novel motifs, with the number of occurrences varying between 1 and 14. The central nodes of each cluster are the JASPAR motifs, and they are marked with a number that links to the adjacent table containing the number of *de novo* motifs that are similar. B) Comparison of all motifs against each other with a threshold of 0.6 or under. Several clusters are visible, including one made up of motifs matching the SP1 weight matrix in A (circled in red). Other motifs outside the SP1 cluster that also matched a JASPAR weight matrix are labelled with the name of the TFBS. These figures were plotted using BioLayout.

5.3 Conclusion

The experiments described in this chapter demonstrate that promoter activity, as measured by luciferase reporter assay, is well-correlated with endogenous gene expression in a qualitative manner. 80% of the promoter activity calls matched the present/marginal/absent calls from the array data. Accepting marginal calls from the array data as confirming expression changed this figure negligibly, with the vast majority of them having active promoters. This confirmed the importance of the promoter sequence to the integration of regulatory inputs, as they largely continued functioning even when taken out of their genomic context. However, this correlation only held for yes/no designations of expression and promoter activity. The correlation between absolute promoter activity and the level of gene expression was much poorer. This contrasts with previous work on the promoters in the ENCODE regions, which showed a moderate but still highly significant quantitative correlation of 0.53 between promoter activity and gene expression, although in this case expression was measured by RT-PCR rather than arrays (Cooper et al. 2006). The difference may reflect the relative abilities of Affymetrix arrays and RT-PCR to accurately determine the gene expression level of a gene, with RT-PCR being the more accurate of the two methods. Another consideration is that this project tested multiple sequences per promoter that often had different promoter activities, whereas the ENCODE study only used a single sequence. This is bound to decrease the amount of correlation given the degree of difference observed in the activities of different promoter haplotypes, making it necessary to decide how to convert these to a single value (in this case, the highest-expressing haplotyope was used).

Where the qualitative promoter and expression calls did not match, there were two possible kinds of discrepancy; promoters active in the reporter assays that were not expressed endogenously, and promoters not active in the reporter assays that were expressed endogenously. The number of discrepancies in the former category outnumbered the latter by a factor of ~2. This suggests that inhibitory regulatory inputs into promoters, such as upstream silencer elements and repressive chromatin, are more common that stimulatory ones, such as upstream enhancers, in modulating the activity of a promoter *in vivo*. Indeed, the difference seen here might well be an underestimate, as the use of differentially regulated alternative promoters *in vivo* may

mask occurrences of cloned and active promoters that are inactive in the cell. This is because the majority of probe sets on the Affymetrix array are unable to distinguish between transcripts with different first exons, as they tend to be biased towards the 3' end where such transcripts would share sequence. There is extensive evidence for widespread use of alternative promoters in humans from ChIP-chip studies of RNA Pol II localisation (Kim et al. 2005b). 22% of promoters in the ENCODE regions contain at least one alternative promoter (Cooper et al. 2006). In contrast, some of the promoters inactive in the luciferase assays may have been due to the real TSS being too far downstream of the annotated TSS for it to be cloned optimally (see section 4.2.11). This effect was relatively minor and could not account for the difference between the two categories.

There are several potential sources of inhibitory inputs into a promoter;

- Transcriptional repressor proteins that inhibit TFs and/or the basal transcription machinery via protein-protein interactions with stimulatory TFs or the pre-initiation complex
- Transcriptional repressor proteins that inhibit TFs and/or the basal transcription machinery by competing for the same binding sites. The inhibition is effected by sterically blocking the action of stimulatory factors at promoters rather than by direct protein-protein interaction
- Epigenetic factors such as histone modifications leading to condensed chromatin, or promoter methylation, causing transcriptional silencing
- Upstream *cis*-acting transcriptional silencer elements that function either by blocking the action of an enhancer or by recruiting transcriptional repressor proteins that then interact with and inhibit proteins on the core and proximal promoters

As both the cloned and endogenous promoters were exposed to the same TF background (within the margins of biological variation between different cultures of each cell line), the first two inhibitory inputs cannot be responsible for the effect observed. This is because they would be expected to act equally on both versions of the promoter. The overrepresentation of negative inputs is thus likely to be caused by a combination of epigenetic repression and upstream transcriptional silencer elements,

as these will affect the endogenous promoter but not the cloned one. The distinction between the two processes is not necessarily clear-cut, as DNA elements can themselves recruit histone modification enzymes that then exert epigenetic effects (Rezai-Zadeh et al. 2003). There is evidence that many promoters have activating elements within the first 500 bases upstream of the TSS, but inhibitory elements between 500 and 1000 bases upstream (Cooper et al. 2006). This was discovered by making serial deletions in a set of cloned promoters from the ENCODE regions. This suggests a significant role for upstream silencing elements in the discrepancy between cloned promoters and endogenous expression, particularly as the fragments cloned in this study only extended to around 600 bases. Interestingly, genes that were only marginally expressed on the array had a median promoter activity twice as high as that of genes that are definitively expressed. This ties in well with the overrepresentation of non-expressed active promoters discussed above, and together these pieces of evidence suggest a prominent role for inhibitory relative to stimulatory inputs. One way to investigate these possibilities is to measure the methylation state of the promoters by bisulphite sequencing or use ChIP-chip to look at the histone modification state of the chromatin around the promoters. These technologies would reveal the extent of the epigenetic component of this possible effect. The presence of upstream silencer elements would be more difficult to prove, as their positional relationship to the promoters is usually unknown. The cloning of larger promoter fragments into luciferase vectors followed by serial deletions and reporter assays could reveal the presence of repressive elements nearby (Cooper et al. 2006).

Analysis of the expression of TFs that had binding sites around functional SNPs seemed to re-iterate the fact that some of these motifs may not be biologically functional regardless of how well they may match known optimal binding sites. In only 8 of 28 instances of a polymorphism in a TFBS was the expression data consistent with a role for the TF. This included the somewhat ambiguous cEBP motif that could have been targeted by any of three isoforms of cEBP, one of which was universally absent and two which were universally present (cEBPB probe set was differentially expressed but was not correlated with the functionality of the SNP). In 14 instances of a binding site around a functional SNP, the expression pattern of the TF seemed to definitively rule out a role in the mechanism behind the functional SNP, as it was not expressed in all the cells in which the function was observed. In only two

cases did differential TF expression correlate with cell-specific SNP function, although it must be stressed that this is not a conclusive piece of evidence. The conclusion to be drawn from this analysis is that the presence of binding sites does not necessarily equate with function, and that the proportion of cases where causality was eliminated on the basis of lack of expression of the factor suggests that using TFBS as predictive entities would unavoidably cause an substantial false positive rate.

Attempts to classify the cell lines according to the profile of their promoter activities seemed to yield very different results to similar classification based on TF expression or the expression of the endogenous genes whose promoters had been cloned. The latter two, in contrast, gave very similar results, and suggested that the cell lines were about equally different from each other. This discrepancy may be due to stochastic or experimental factors influencing the absolute activities of the promoters in each experiment. The fact that patterns of expression between haplotypes within a promoter were more reproducible than the absolute values themselves seems to suggest that the promoter activities on their own are not necessarily definitive. It is also possible that when a promoter is in its correct genomic context it can be more tightly controlled and will thus not be as susceptible to stochastic variation or small differences in experimental conditions.

The novel motifs generated by aligning the promoters of genes with similar expression profiles across the 4 cell lines failed to improve on the performance of previously known motifs. This was disappointing, but not entirely unexpected given the performance of other putative regulatory motifs. While it was hoped that they would perform at least as well as the motifs from other sources, they showed even less enrichment than many of these classes of elements (see section 4.2.15). Several reasons may have contributed to this. Firstly, the number of cell lines was relatively small, and it was possible that this might have led to the alignment of promoters that were not meaningfully co-regulated *in vivo*. This would bias the motif finding algorithm of nestedMICA away from the real signal. However, other studies that have used co-regulation to infer regulatory elements have used as few as two conditions for any single comparison (Roth et al. 1998). The fact that all 4 cell lines were well-established transformed lines may have led to a convergence of expression profiles relative to what would be expected if the tissues of origin (in this case skin, medulla,

embryonic kidney and cervix) were compared. While there is no published information on these particular cell lines and their original tissues, expression profiling of cancer cell lines has shown that they cluster principally according to tissue of origin (Ross et al. 2000), suggesting that this is unlikely to be a factor in this case. The fact that around half of the motifs either matched a known TFBS or clustered with other motifs under stringent conditions indicated that a substantial fraction of these motifs might be real, although manual inspection of some of the motifs did show a substantial number with poor and discontiguous information profiles suggesting that they may not have been biologically meaningful. Perhaps the differences in expression in the set of genes under study were not substantial enough across these four cell lines to reliably cluster them without spuriously including genes that were not really co-regulated, hence giving rise to uninformative motifs.

6 Discussion and Future Work

6.1 Discussion

While a significant number of SNPs in putative promoters are already available as a matter of course from the genome project and SNP ascertainment projects (Sachidanandam et al. 2001; Consortium 2005b; Hinds et al. 2005), there have been almost no efforts of any scale to specifically mine promoter sequences for polymorphisms. Buckland et al were the first group to re-sequence promoters across many genes, but their panel was small, ethnically heterogeneous and gave limited information about allele frequencies, as well as suffering from significant ascertainment bias as reported by the authors themselves (Buckland et al. 2005). This project has carried out the deepest available re-sequencing of promoters currently available, with considerably more power to detect rare polymorphisms than the Buckland project despite there still being some ascertainment bias away from rare SNPs. In a surprising result, essentially no difference was found between overall mutation rates in promoters and in chromosome 22 overall apart from those explainable by elevated GC content. This is despite the naïve assumption that the promoters would have suppressed C-T mutation rates compared to the rest of the chromosome. Some reasons why this might be the case have been outlined in section 4.3. However, an interesting avenue for further investigation would be to look at the history of C-T mutations in order to see whether the rate in the genome as a whole has slowed over time. This could be done by using a measure such as extended haplotype heterozygosity to estimate an age profile for C/T SNPs versus other SNPs, to see whether C/T SNPs are generally older (although this would depend on whether such a slowdown had happened within human evolutionary timescales).

Rockman and Wray have previously estimated a rate of 0.94 functional SNPs per kb in the 850 base pair sequences upstream of TSSs (Rockman and Wray 2002). This was likely to be an underestimate, as the majority of functional variants in the promoters studied have probably not been identified. The chromosome 22 project identified between 0.73 and 0.98 functional SNPs per kb, depending on the number of unconfirmed SNPs that are taken as being real. This is from an average of 630 base pairs upstream. These numbers are in remarkable agreement considering the very different methods used to obtain them, and suggest that the significantly greater

degree of functional variation observed here compared to the Buckland set should not be considered surprising.

What is still unclear is how much of this promoter variation that is detectable by isolating the promoter remains significant when all the other regulatory inputs found in a native genome are added? This work has not been carried out on a consistent set of promoter polymorphisms such as that produced here. However, literature surveys suggest that a significant proportion of SNPs with functional effects in reporter assays also have further evidence of function either on a biochemical or disease level phenotype. Indeed, for a set of 107 genes with published functional promoter polymorphism, 59% and 71% respectively also had published evidence of such phenotypes (Rockman and Wray 2002). These figures may be affected by publication bias as a result of underreporting of negative results, and this is probably not possible to quantify, but nevertheless the link between reporter assays and an *in vivo* function does exist and can be amply demonstrated with current methods, many of which are now being developed to a high-throughput capability (Knight et al. 2003; Linnell et al. 2004). There is also considerable evidence of extensive allele-specific variation in gene expression (Yan et al. 2002b; Pastinen et al. 2005) as well as association between *cis*-acting loci and gene expression levels (Monks et al. 2004; Cheung et al. 2005; Stranger et al. 2005) that suggest the presence of a lot of *cis*-regulatory variation in the genome. Essentially all these studies have been carried out on subsets of the same CEPH families from which the panel for this project was drawn. Even though this does not say anything about the *in vivo* functionality of the particular functional SNPs discovered, it does demonstrate that there is ample potential for them to have phenotypic consequences at least on expression phenotypes in the 48-person CEPH panel, if not at the level of disease and/or organismal phenotype. While no evidence was found for an association of any of the *in vitro* functional SNPs with expression phenotypes in the HapMap individuals in the panel, this may well have been due to the low power afforded from an overlap of only 31 individuals. The lack of power would be exacerbated by the failure to obtain genotypes from the re-sequencing for a subset of individuals in each SNP. This would lead to an even smaller number of informative individuals for whom functional data was available, and was not an uncommon occurrence. The net result was to make it relatively unlikely for any association to survive the correction for multiple testing.

A crucial result of this project was the lack of enrichment of functional SNPs in putative regulatory elements including TFBS and ultraconserved regions. This is surprising given that the traditional model for the action of functional promoter SNPs has been the perturbation of TFBS. Buckland et al reported that only 35% of the functional SNPs were in a TFBS (Buckland et al. 2005). However, the absolute numbers of putative TFBS present in a promoter, as determined by any of a number of possible tools and databases is largely a function of the parameters used for the search and the quality of the position weight matrices in the database. It may therefore be more meaningful to compare the rates of functional and non-functional SNPs in TFBS using consistent parameters and express this as an enrichment factor. To my knowledge, this is the first project to explore the enrichment of putative TFBS for functional SNPs, although others have used TFBS as a criterion to predict functional SNPs (Mottagui-Tabar et al. 2005). The lack of enrichment suggests that current models of TFBS are inadequate and not useful for predicting whether promoter SNPs are likely to be functional. This is despite ample evidence that some regulatory SNPs do function by altering the affinity of a TFBS, as evidenced by EMSA experiments using allelic probes and transient transfection assays in parallel (Rockman and Wray 2002). However, it is often the case in the literature that one set of experiments is done without the other, making it difficult to assess how much known functional variation can be accounted for in this manner. Limited evidence from a small number of experiments has suggested that between 70 and 80% of SNPs in TFBS within conserved regions can alter the binding of a TF *in vitro* according to EMSA experiments (Belanger et al. 2005; Mottagui-Tabar et al. 2005). Even if these results were representative, it is still the case that not all SNPs in binding sites cause functional differences, and indeed it may be that only a minority of sites do so (Rockman and Wray 2002). The lack of functionality of SNPs in some binding sites (even ones experimentally verified by EMSA), as well as the number of functional promoter SNPs apparently not within any known binding sites points to one or both of two possibilities; that there is a significant number of binding sites still to be discovered or that these SNPs are exerting their effects by a mechanism other than direct perturbation of a binding site.

Several analyses of human promoters using various methods, often heavily reliant on evolutionary conservation, have found conserved motifs that are enriched at promoters (Xie et al. 2005; Robertson et al. 2006). This enrichment, and the fact that many known motifs have been re-discovered with these methods, suggests that they may indeed be functional, although the resulting elements have yet to be functionally tested (for example by deletion analysis in reporter constructs). It is therefore not unlikely that our knowledge of the number of regulatory elements is far from complete, although it has been proposed that many of the remaining motifs may be rare and/or only functional in restricted biological conditions (Buckland 2006).

There is also evidence that non-binding site-dependent mechanisms may be important in explaining promoter SNP effects. These SNPs may function by altering the conformational properties of the DNA upstream of the TSS, and thus altering the dynamics of TF interactions with each other and the promoter without necessarily being in a binding site (Buckland 2006). The inherent curvature of DNA is often higher at promoters, and this has been shown to be an important factor in the activation of at least some eukaryotic genes (Nishikawa et al. 2003). Manipulations of cloned promoters in reporter vectors have shown that promoters with higher inherent curvature can promote transcription markedly more efficiently than the same promoter carrying mutations that reduced this curvature (Kim, Klooster, and Shapiro 1995). The addition of intercalators that abrogated this curvature greatly reduced this activity difference (Kim, Klooster, and Shapiro 1995). While structural studies show that some TFs, including TBP and p53 (Nagaich, Appella, and Harrington 1997; de Souza and Ornstein 1998), alter the conformation of DNA on binding, it is also the case that DNA which is already in a favourable conformation pre-binding can drastically increase binding affinity (Parvin et al. 1995). Alteration of TF binding efficiency by the introduction of artificial substitutions outside the TFBS that alter conformation has been demonstrated in yeast (Acton, Zhong, and Vershon 1997), although the presence or extent of natural SNPs that function in this way is unknown. A distinct but related property of the DNA itself that can be important in TF binding is the flexibility, or the ability of DNA conformation to be altered by the binding of proteins. This can be important in allowing multiple protein-DNA interactions in close proximity by relieving steric hindrances (either by one factor binding multiple sites or by multiple factors) or by allowing the DNA to loop and bring distant bound

factors into contact (Mastrangelo et al. 1991; Suzuki and Yagi 1995; Nagaich, Appella, and Harrington 1997).

The results produced in this project and other evidence presented above have important implications for efforts to predict functional polymorphisms by using models of TFBSs. While several such attempts have been made, usually claiming at least moderate success, they are often tested using an inadequately small number of actual functional experiments (Belanger et al. 2005; Mottagui-Tabar et al. 2005). This makes their success hard to quantify, although the fact that even small scale predictions were not confirmed more than 50% of the time suggests there is still some way to go before such predictive methods become reliable. There is some evidence that even using position weight matrices rather than simple consensus sequences may not enable the true deduction of the effect of a base change on a binding site, and that more complete experimental characterization of TFBS may be necessary for this (Bulyk, Johnson, and Church 2002). The presence of an unbiased potential training set of functional polymorphisms may be very important in developing new *in silico* methods for regulatory polymorphism discovery. *In silico* analysis of the effect of the functional SNPs discovered here and by Buckland et al on the DNA conformation may shed more light on the putative importance of this mechanism. Collaboration with other groups to analyse the performance of some of the novel motifs discovered by comparative genomics (Xie et al. 2005) may also shed more light on the utility of conservation for predicting functional variation.

This project has also explored the qualitative relationship between promoter activity and *in vivo* expression. This has confirmed that promoter sequences contain many of the elements that determine whether a gene is expressed or not, and therefore that the promoter really does integrate the majority of signals in the transcription initiation pathway. Other work has found a more quantitative relationship between promoter activity and gene expression (Cooper et al. 2006), but this was not reproduced here. As suggested before, this may be due to the relative quantitative potential of RT-PCR (as used by Cooper et al) and Affymetrix arrays. Another factor may be the difference in the controls used for the luciferase assays, where a single promoterless plasmid was used in this project versus the average of 102 cloned non-functional DNA elements by Cooper et al. This latter control may form a more consistent baseline as any non-

specific activation of transcription due to stochastic biological variation in different cell growths would perturb the baseline by relatively low levels. Indeed, Promega have recently released the pGL4 luciferase plasmid series, where a large number of cryptic TFBS were removed from the vector backbone relative to the pGL3 plasmids. These may have been a source of variation in background levels.

The finding that upregulatory mutations are skewed towards higher derived allele frequencies relative to downregulatory mutations may have implications for the evolutionary mechanisms of gene regulation. The expansion of derived alleles of functional SNPs has been observed previously, with 7 out of 21 known functional SNPs having derived major alleles, and 11 of the remainder having either allele as the major allele in different populations (Rockman and Wray 2002). However, greater tendency for upregulatory changes in promoters to expand relative to downregulatory changes is a novel finding, and suggests that upregulatory promoter changes may be more amenable to positive selection than downregulatory ones, and may therefore be more likely to have positive fitness consequences. If this were the case, it may be important to understanding the mechanistic basis of transcriptome evolution. The known phylogeny between primates is recapitulated by expression variation between species (Gilad et al. 2006), and levels of selective constraint on gene expression levels and coding sequence coincide (Khaitovich et al. 2005). Interestingly, despite more constraint on interspecific gene expression variation in the brain in primates (Khaitovich et al. 2005), there has been an acceleration in gene expression changes in the human lineage (Enard et al. 2002), and this difference is made up largely of upregulations rather than downregulations (Caceres et al. 2003). Upregulations in gene expression in the human lineage have also occurred in human versus chimpanzee TF genes (Gilad et al. 2006) and in fibroblasts (Karaman et al. 2003), although the bias in favour of upregulations is much less clear in the latter case.

The bias towards expansion of upregulatory changes seems at odds with some theoretical models of transcriptome evolution, which propose that downregulatory changes should be more common that upregulatory ones (Khaitovich, Paabo, and Weiss 2005). It also does not agree with recent findings by the Dermitzakis lab at the Sanger Institute that SNPs found by whole genome association to expression phenotypes agree with this model (Stranger et al unpublished). However, it is

important to note that while Stranger et al were measuring mRNA levels, this project was measuring *in vitro* promoter activity, with the latter being a component of the former. A possible explanation for the discrepancy is that these association studies may be finding regulatory SNPs in distant enhancer or silencer elements rather than the promoter, and that such functional SNPs may have more powerful effects than those at promoters. This is suggested by the fact that the majority of SNPs identified by Stranger et al are more than 10 kb away from the TSS of the genes they influence (data not shown). The effects of these elements on transcription may be sufficiently powerful that where they contain functional variation, this dominates over promoter sequence variation, and precludes it from identification in association studies. This may also explain discrepancies in the difference between human and chimpanzee promoter activities and the corresponding difference in transcript levels (Heissig et al. 2005). Heissig and colleagues found seven genes that showed significant differences between chimpanzees and humans both in luciferase reporter assays and measures of transcript abundance. However, in 4/7 genes these differences were in the opposite direction to each other (Heissig et al. 2005). It may therefore be proposed that globally, variation in proximal promoters and in distal regulatory elements are influenced differently by selection.

6.2 Future work

Following on from the generation of a set of functional promoter polymorphisms *in vitro*, the next natural step is to investigate the effects of these SNPs *in vivo* in order to determine whether they are still functional in their native genomic contexts. There are several experimental methods for doing this, all of which give subtly different levels of information on the SNPs under investigation.

The most obvious method would be to look for differences in the mRNA transcripts produced by variant promoters. The most well-established method for doing this is probably quantitative RT-PCR from cell lines or mRNA from heterozygotes (Yan et al. 2002b; Bray et al. 2003; Pastinen and Hudson 2004). This would require the identification of individuals who were heterozygous both for the promoter haplotypes of interest and for a transcribed marker SNP that could be used to distinguish the two transcripts. It would also necessitate knowledge of the phase of the promoter

haplotypes and the maker SNP in order to be able to say which promoter haplotype is driving the expression of which transcript, enabling the assignment of direction to functional changes. With the HapMap project now having completed phase 1, there is a ready source of cell lines from a range of individuals that can be used for this kind of work (Consortium 2005b). The genotype information would also enable the inference of phase between transcribed markers and the promoter SNPs (Stephens, Smith, and Donnelly 2001).

The advent of chromatin immunoprecipitation combined with a quantitative genotyping method also allows direct assay of differential RNA polymerase II loading on polymorphic promoters in a heterozygote, a technique dubbed the haploChIP method (Knight et al. 2003). This would involve chromatin IP with an antibody to RNA Pol II phosphorylated at serine 5, which is enriched at the 5' end of transcripts. This would be followed by quantitative assessment of fragments from the two promoter alleles by primer-extension and mass spectrometry analysis (Knight et al. 2003). This method has the advantage of not requiring a transcribed marker SNP,as well as the ability to yield information on multiple heterozygous promoters in a single chromatin immunoprecipitation sample, hence making it suitable for high throughput applications.

If a complete set of *in vitro* and *in vivo* data for a set of promoter SNPs could be produced, it would then be desirable to explain the mechanistic basis of any functional differences, either in terms of TF binding or mechanisms related to structural conformation of DNA. Again, an established method for this already exists; electrophoretic mobility shift assays (EMSA). To apply it to SNPs, radioactively labelled oligonucleotide probes would be synthesised containing the putative binding site, with one probe per allele per polymorphism. The allelic probes would then be allowed to bind proteins from cellular extracts and run down an agarose gel to look for a band shift indicating binding. Relative binding abilities would be assessed by using a non-specific competitor oligonucleotide. This currently remains a low throughput process, and would probably be a bottleneck in any large scale pipeline. One advantage however is that it introduces the possibility of identifying unknown TFs binding to SNPs that are not in known sites by mass fingerprinting. A more high-throughput possibility would be to use a haploChIP-style method, but assessing

binding of TFs rather than Pol II. However, this will be limited only to the relatively few TFs for which antibodies are available.

Another area of interest would be to study the population history of functional polymorphisms and examine the relative importance of regulatory variation and coding variation. It is now relatively easy to design genotying assays for a known polymorphism, and the facilities available at the Sanger Institute would enable rapid and thorough genotyping of several hundred putative regulatory SNPs across the entire HapMap population panel. This would enable studies both within and across continental populations, and could make possible the use of robust statistical methods for inferring selection. Importantly, full genotyping in the HapMap individuals of a large panel of functional SNPs would make it easy to repeat the association studies with the whole genome expression data and obtain far more robust associations (and where an association couldn't be shown, this would again be a more convincing negative result).

Finally, current knowledge of functional promoter polymorphisms can be used to build a database of polymorphisms for which function is known *a priori*, and use this for meta-analysis to examine the properties of functional SNPs more thoroughly. This database can then be put through the above battery of methods in order to complete the knowledge required for each of the polymorphisms. Two sets of promoter polymorphisms tested *in vitro* under homogeneous experimental conditions are already available; the data presented in the thesis and that produced by Buckland et al. Together, these consist of 79 isolated and confirmed promoter polymorphisms. There have also been two efforts to curate information from the wider literature, which contains data on many more promoter variants distributed among a large number of papers. Rockman and Wray produced a survey of 140 functional SNPs tested in reporter assays in the literature, and in many cases were able to find supporting published evidence in the form of EMSA experiments or associations with expression or disease phenotypes. In addition, the ORegAnno database of regulatory elements (Montgomery et al. 2006) contains a set of 172 promoter polymorphisms that have been partly manually curated and partly submitted by external contributors. In both these curated datasets, the SNPs are not always clearly-mapped to the genome and the evidence supporting each SNP is very heterogeneous (in some cases, for example,

there is differential binding data from EMSA but no luciferase assay). These would then be put through the methods proposed to complete the evidence for them, with the expectation that there would be a high rate of functional confirmation. Indeed, I was able to construct a preliminary database that would hold the integrated results of such a meta-analysis of published functional promoter SNPs, and was able to populate it with data from both the Buckland set and from individual papers. This work was not presented in this thesis, as more work is needed to establish an ontology for populating it with a dataset that can be consistently analysed.

Eventually, these methods would lead to a set of promoter polymorphisms where data was available for every potential step in the process of explaining their mechanistic basis; *in vitro* function in isolation from confounding regulatory inputs, effect on TF binding, carry-over of the *in vitro* effect *in vivo* and population data to study the selection history of the polymorphism. Such a dataset has never been accumulated before, and could be the turning point for efforts to understand the mechanisms of promoter variation effects. It would be an excellent training set for computational methods that could then be used to predict the effect of promoter SNPs. If these methods could be perfected on the strength of such a training dataset, it would have potential implications for human health, allowing better assessments for non-coding pathogenic variants whose function could not be predicted in the same way as deleterious coding functions.

7 References

Acton, T. B., H. Zhong, and A. K. Vershon. 1997. DNA-binding specificity of Mcm1: operator mutations that alter DNA-bending and transcriptional activities by a MADS box protein. Mol Cell Biol **17**:1881-1889.

Aerts, S., G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. De Moor. 2003. Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Res **31**:1753-1764.

Ahituv, N., S. Prabhakar, F. Poulin, E. M. Rubin, and O. Couronne. 2005. Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. Hum Mol Genet **14**:3057-3063.

Al-Shahrour, F., R. Diaz-Uriarte, and J. Dopazo. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics **20**:578-580.

Alam, J., and J. L. Cook. 1990. Reporter genes: application to the study of mammalian gene transcription. Anal Biochem **188**:245-254.

Alves, G., A. Tatro, and T. Fanning. 1996. Differential methylation of human LINE-1 retrotransposons in malignant cells. Gene **176**:39-44.

Antequera, F. 2003. Structure, function and evolution of CpG island promoters. Cell Mol Life Sci **60**:1647-1658.

Antequera, F., and A. Bird. 1993. Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A **90**:11995-11999.

Bailey, T. L., and C. Elkan. 1994. Fitting a mixture model by expectation 30 maximization to discover motifs in biopolymers. Pp. 28-36. International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, California.

Bajic, V. B., and S. H. Seah. 2003a. Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes. Nucleic Acids Res **31**:3560-3563.

Bajic, V. B., and S. H. Seah. 2003b. Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units. Genome Res **13**:1923-1929.

Bajic, V. B., S. H. Seah, A. Chong, S. P. Krishnan, J. L. Koh, and V. Brusic. 2003. Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates. J Mol Graph Model **21**:323-332.

Bajic, V. B., S. L. Tan, Y. Suzuki, and S. Sugano. 2004. Promoter prediction analysis on the whole human genome. Nat Biotechnol **22**:1467-1473.

Bamshad, M. J., S. Mummidi, E. Gonzalez, S. S. Ahuja, D. M. Dunn, W. S. Watkins, S. Wooding, A. C. Stone, L. B. Jorde, R. B. Weiss, and S. K. Ahuja. 2002. A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci U S A **99**:10539-10544.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics **21**:263-265.

Barrett, T. B., R. L. Hauger, J. L. Kennedy, A. D. Sadovnick, R. A. Remick, P. E. Keck, S. L. McElroy, M. Alexander, S. H. Shaw, and J. R. Kelsoe. 2003. Evidence that a single nucleotide polymorphism in the promoter of the G protein receptor kinase 3 gene is associated with bipolar disorder. Mol Psychiatry **8**:546-557.

Baylin, S. B. 2005. DNA methylation and gene silencing in cancer. Nat Clin Pract Oncol **2 Suppl 1**:S4-11.

Belanger, H., P. Beaulieu, C. Moreau, D. Labuda, T. J. Hudson, and D. Sinnett. 2005. Functional promoter SNPs in cell cycle checkpoint genes. Hum Mol Genet **14**:2641-2648.

Bell, A. C., and G. Felsenfeld. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature **405**:482-485.

Bell, A. C., A. G. West, and G. Felsenfeld. 2001. Insulators and boundaries: versatile regulatory elements in the eukaryotic. Science **291**:447-450.

Belting, H. G., C. S. Shashikant, and F. H. Ruddle. 1998. Modification of expression and cis-regulation of Hoxc8 in the evolution of diverged axial morphology. Proc Natl Acad Sci U S A **95**:2355-2360.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. J Roy Stat Soc B Met **57**:289-300.

Bernard, P., and M. Couturier. 1992. Cell killing by the F plasmid CcdB protein involves poisoning of DNA-topoisomerase II complexes. J Mol Biol **226**:735-745.

Blackwood, E. M., and J. T. Kadonaga. 1998. Going the distance: a current view of enhancer action. Science **281**:60-63.

Bray, N. J., P. R. Buckland, M. J. Owen, and M. C. O'Donovan. 2003. Cis-acting variation in the expression of a high proportion of genes in human brain. Hum Genet **113**:149-153.

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science **296**:752-755.

Brown, R. P., and M. E. Feder. 2005. Reverse transcriptional profiling: non-correspondence of transcript level variation and proximal promoter polymorphism. BMC Genomics **6**:110.

Bucher, P. 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol **212**:563-578.

Buckland, P. R. 2006. The importance and identification of regulatory polymorphisms and their mechanisms of action. Biochim Biophys Acta **1762**:17-28.

Buckland, P. R., S. L. Coleman, B. Hoogendoorn, C. Guy, S. K. Smith, and M. C. O'Donovan. 2004a. A high proportion of chromosome 21 promoter polymorphisms influence transcriptional activity. Gene Expr **11**:233-239.

Buckland, P. R., B. Hoogendoorn, S. L. Coleman, C. A. Guy, S. K. Smith, and M. C. O'Donovan. 2005. Strong bias in the location of functional promoter polymorphisms. Hum Mutat **26**:214-223.

Buckland, P. R., B. Hoogendoorn, C. A. Guy, S. L. Coleman, S. K. Smith, J. D. Buxbaum, V. Haroutunian, and M. C. O'Donovan. 2004b. A high proportion of polymorphisms in the promoters of brain expressed genes influences transcriptional activity. Biochim Biophys Acta **1690**:238-249.

Bulger, M., T. Sawado, D. Schubeler, and M. Groudine. 2002. ChIPs of the beta-globin locus: unraveling gene regulation within an active domain. Curr Opin Genet Dev **12**:170-177.

Bulyk, M. L., P. L. Johnson, and G. M. Church. 2002. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res **30**:1255-1261.

Burke, T. W., and J. T. Kadonaga. 1996. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev **10**:711-724.

Butler, J. E., and J. T. Kadonaga. 2001. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. Genes Dev **15**:2515-2519.

Caceres, M., J. Lachuer, M. A. Zapala, J. C. Redmond, L. Kudo, D. H. Geschwind, D. J. Lockhart, T. M. Preuss, and C. Barlow. 2003. Elevated gene expression levels distinguish human from non-human primate brains. Proc Natl Acad Sci U S A **100**:13030-13035.

Cajiao, I., A. Zhang, E. J. Yoo, N. E. Cooke, and S. A. Liebhaber. 2004. Bystander gene activation by a locus control region. Embo J **23**:3854-3863.

Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet.

Carroll, S. B. 2000. Endless forms: the evolution of gene regulation and morphological diversity. Cell **101**:577-580.

Carson, S., and M. V. Wiles. 1993. Far upstream regions of class II MHC Ea are necessary for position-independent, copy-dependent expression of Ea transgene. Nucleic Acids Res **21**:2065-2072.

Carter, D., L. Chakalova, C. S. Osborne, Y. F. Dai, and P. Fraser. 2002. Long-range chromatin regulatory interactions in vivo. Nat Genet **32**:623-626.

Cawley, S., S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, and T. R. Gingeras. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell **116**:499-509.

Chen, L. 1999. Combinatorial gene regulation by eukaryotic transcription factors. Curr Opin Struct Biol **9**:48-55.

Cheng, J., P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D. K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D. S. Gerhard, and T. R. Gingeras. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science **308**:1149-1154.

Cheung, V. G., L. K. Conlin, T. M. Weber, M. Arcaro, K. Y. Jen, M. Morley, and R. S. Spielman. 2003. Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet **33**:422-425.

Cheung, V. G., R. S. Spielman, K. G. Ewens, T. M. Weber, M. Morley, and J. T. Burdick. 2005. Mapping determinants of human gene expression by regional and genome-wide association. Nature **437**:1365-1369.

Chuang, J. H., and H. Li. 2004. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. PLoS Biol **2**:E29.

Collins, J. E., M. E. Goward, C. G. Cole, L. J. Smink, E. J. Huckle, S. Knowles, J. M. Bye, D. M. Beare, and I. Dunham. 2003. Reevaluating human gene annotation: a second-generation analysis of chromosome 22. Genome Res **13**:27-36.

Collins, J. E., C. L. Wright, C. A. Edwards, M. P. Davis, J. A. Grinham, C. G. Cole, M. E. Goward, B. Aguado, M. Mallya, Y. Mokrab, E. J. Huckle, D. M. Beare, and I. Dunham. 2004. A genome annotation-driven approach to cloning the human ORFeome. Genome Biol **5**:R84.

Concino, M. F., R. F. Lee, J. P. Merryweather, and R. Weinmann. 1984. The adenovirus major late promoter TATA box and initiation site are both necessary for transcription in vitro. Nucleic Acids Res **12**:7423-7433.

Consortium, I. H. G. S. 2004a. Finishing the euchromatic sequence of the human genome. Nature **431**:931-945.

Consortium, T. C. S. a. A. 2005a. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**:69-87.

Consortium, T. E. P. 2004b. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science **306**:636-640.

Consortium, T. I. H. 2005b. A haplotype map of the human genome. Nature **437**:1299-1320.

Cooper, S. J., N. D. Trinklein, E. D. Anton, L. Nguyen, and R. M. Myers. 2006. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res **16**:1-10.

Crawford, D. L., J. A. Segal, and J. L. Barnett. 1999. Evolutionary analysis of TATA-less proximal promoter function. Mol Biol Evol **16**:194-207.

Daborn, P. J., J. L. Yen, M. R. Bogwitz, G. Le Goff, E. Feil, S. Jeffers, N. Tijet, T. Perry, D. Heckel, P. Batterham, R. Feyereisen, T. G. Wilson, and R. H. ffrench-Constant. 2002. A single p450 allele associated with insecticide resistance in Drosophila. Science **297**:2253-2256.

Davuluri, R. V., I. Grosse, and M. Q. Zhang. 2001. Computational identification of promoters and first exons in the human genome. Nat Genet **29**:412-417.

Dawson, E., Y. Chen, S. Hunt, L. J. Smink, A. Hunt, K. Rice, S. Livingston, S. Bumpstead, R. Bruskiewich, P. Sham, R. Ganske, M. Adams, K. Kawasaki, N. Shimizu, S. Minoshima, B. Roe, D. Bentley, and I. Dunham. 2001. A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. Genome Res **11**:170-178.

De Gobbi, M., V. Viprakasit, J. R. Hughes, C. Fisher, V. J. Buckle, H. Ayyub, R. J. Gibbons, D. Vernimmen, Y. Yoshinaga, P. de Jong, J. F. Cheng, E. M. Rubin, W. G. Wood, D. Bowden, and D. R. Higgs. 2006. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science **312**:1215-1217.

de Souza, O. N., and R. L. Ornstein. 1998. Inherent DNA curvature and flexibility correlate with TATA box functionality. Biopolymers **46**:403-415.

Dekker, J., K. Rippe, M. Dekker, and N. Kleckner. 2002. Capturing chromosome conformation. Science **295**:1306-1311.

Deng, G., G. A. Song, E. Pong, M. Sleisenger, and Y. S. Kim. 2004. Promoter methylation inhibits APC gene expression by causing changes in chromatin conformation and interfering with the binding of transcription factor CCAAT-binding factor. Cancer Res **64**:2692-2698.

Dermitzakis, E. T., and A. G. Clark. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol **19**:1114-1121.

Dermitzakis, E. T., A. Reymond, and S. E. Antonarakis. 2005. Conserved non-genic sequences - an unexpected feature of mammalian genomes. Nat Rev Genet **6**:151-157.

Deutsch, S., R. Lyle, E. T. Dermitzakis, H. Attar, L. Subrahmanyan, C. Gehrig, L. Parand, M. Gagnebin, J. Rougemont, C. V. Jongeneel, and S. E. Antonarakis. 2005. Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. Hum Mol Genet **14**:3741-3749.

Dierks, P., A. van Ooyen, M. D. Cochran, C. Dobkin, J. Reiser, and C. Weissmann. 1983. Three regions upstream from the cap site are required for efficient and accurate transcription of the rabbit beta-globin gene in mouse 3T6 cells. Cell **32**:695-706.

Down, T. A., and T. J. Hubbard. 2005. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. Nucleic Acids Res **33**:1445-1453.

Down, T. A., and T. J. Hubbard. 2002. Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res **12**:458-461.

Dunham, I., N. Shimizu, B. A. Roe, S. Chissoe, A. R. Hunt, J. E. Collins, R. Bruskiewich, D. M. Beare, M. Clamp, L. J. Smink, R. Ainscough, J. P. Almeida, A. Babbage, C. Bagguley, J. Bailey, K. Barlow, K. N. Bates, O. Beasley, C. P. Bird, S. Blakey, A. M. Bridgeman, D. Buck, J. Burgess, W. D. Burrill, K. P. O'Brien, and et al. 1999. The DNA sequence of human chromosome 22. Nature **402**:489-495.

Dvir, A., J. W. Conaway, and R. C. Conaway. 2001. Mechanism of transcription initiation and promoter escape by RNA polymerase II. Curr Opin Genet Dev **11**:209-214.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A **95**:14863-14868.

Elander, N., P. Soderkvist, and K. Fransen. 2006. Matrix metalloproteinase (MMP) -1, -2, -3 and -9 promoter polymorphisms in colorectal cancer. Anticancer Res **26**:791-795.

Enard, W., P. Khaitovich, J. Klose, S. Zollner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, G. M. Doxiadis, R. E. Bontrop, and S. Paabo. 2002. Intra- and interspecific variation in primate gene expression patterns. Science **296**:340-343.

Enright, A. J., and C. A. Ouzounis. 2001. BioLayout--an automatic graph layout algorithm for similarity visualization. Bioinformatics **17**:853-854.

Euskirchen, G., T. E. Royce, P. Bertone, R. Martone, J. L. Rinn, F. K. Nelson, F. Sayward, N. M. Luscombe, P. Miller, M. Gerstein, S. Weissman, and M. Snyder. 2004. CREB binds to multiple loci on human chromosome 22. Mol Cell Biol **24**:3804-3814.

Evans, R., J. A. Fairley, and S. G. Roberts. 2001. Activator-mediated disruption of sequence-specific DNA contacts by the general transcription factor TFIIB. Genes Dev **15**:2945-2949.

Evans, R. M. 1988. The steroid and thyroid hormone receptor superfamily. Science **240**:889-895.

Ewing, B., and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res **8**:186-194.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res **8**:175-185.

Fay, J. C., H. L. McCullough, P. D. Sniegowski, and M. B. Eisen. 2004. Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae. Genome Biol **5**:R26.

Feinberg, A. P., and B. Vogelstein. 1983. Hypomethylation of ras oncogenes in primary human cancers. Biochem Biophys Res Commun **111**:47-54.

Fickett, J. W., and A. G. Hatzigeorgiou. 1997. Eukaryotic promoter recognition. Genome Res **7**:861-878.

Filippova, G. N., C. P. Thienes, B. H. Penn, D. H. Cho, Y. J. Hu, J. M. Moore, T. R. Klesert, V. V. Lobanenkov, and S. J. Tapscott. 2001. CTCF-binding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus. Nat Genet **28**:335-343.

Garcia-Giralt, N., A. Enjuanes, M. Bustamante, L. Mellibovsky, X. Nogues, R. Carreras, A. Diez-Perez, D. Grinberg, and S. Balcells. 2005. In vitro functional assay of alleles and haplotypes of two COL1A1-promoter SNPs. Bone **36**:902-908.

Garcia-Giralt, N., X. Nogues, A. Enjuanes, J. Puig, L. Mellibovsky, A. Bay-Jensen, R. Carreras, S. Balcells, A. Diez-Perez, and D. Grinberg. 2002. Two new single-nucleotide polymorphisms in the COL1A1 upstream regulatory region and their relationship to bone mineral density. J Bone Miner Res **17**:384-393.

Gardiner-Garden, M., and M. Frommer. 1987. CpG islands in vertebrate genomes. J Mol Biol **196**:261-282.

Ge, B., S. Gurd, T. Gaudin, C. Dore, P. Lepage, E. Harmsen, T. J. Hudson, and T. Pastinen. 2005. Survey of allelic expression using EST mining. Genome Res **15**:1584-1591.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5**:R80.

Giardine, B., C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko. 2005. Galaxy: a platform for interactive large-scale genome analysis. Genome Res **15**:1451-1455.

Gibson, A. W., J. C. Edberg, J. Wu, R. G. Westendorp, T. W. Huizinga, and R. P. Kimberly. 2001. Novel single nucleotide polymorphisms in the distal IL-10 promoter affect IL-10 production and enhance the risk of systemic lupus erythematosus. J Immunol **166**:3915-3922.

Gilad, Y., A. Oshlack, G. K. Smyth, T. P. Speed, and K. P. White. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature **440**:242-245.

Grosschedl, R., and M. L. Birnstiel. 1980. Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. Proc Natl Acad Sci U S A **77**:1432-1436.

Hain, J., W. D. Reiter, U. Hudepohl, and W. Zillig. 1992. Elements of an archaeal promoter defined by mutational analysis. Nucleic Acids Res **20**:5423-5428.

Hamblin, M. T., and A. Di Rienzo. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet **66**:1669-1679.

Hark, A. T., C. J. Schoenherr, D. J. Katz, R. S. Ingram, J. M. Levorse, and S. M. Tilghman. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature **405**:486-489.

Harr, B., and C. Schlotterer. 2006. Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. Nucleic Acids Res **34**:e8.

Hartl, D. L., and A. G. Clark. 1997. Principles of population genetics. Sinauer, Sunderland, Mass.

Hashimoto, S., Y. Suzuki, Y. Kasai, K. Morohoshi, T. Yamada, J. Sese, S. Morishita, S. Sugano, and K. Matsushima. 2004. 5'-end SAGE for the analysis of transcriptional start sites. Nat Biotechnol **22**:1146-1149.

Heissig, F., J. Krause, J. Bryk, P. Khaitovich, W. Enard, and S. Paabo. 2005. Functional analysis of human and chimpanzee promoters. Genome Biol **6**:R57.

Herman, J. G., F. Latif, Y. Weng, M. I. Lerman, B. Zbar, S. Liu, D. Samid, D. S. Duan, J. R. Gnarra, W. M. Linehan, and et al. 1994. Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. Proc Natl Acad Sci U S A **91**:9700-9704.

Hertz, G. Z., and G. D. Stormo. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics **15**:563-577.

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin, D. G. Ballinger, K. A. Frazer, and D. R. Cox. 2005. Whole-genome patterns of common DNA variation in three human populations. Science **307**:1072-1079.

Ho, Y., F. Elefant, N. Cooke, and S. Liebhaber. 2002. A defined locus control region determinant links chromatin domain acetylation with long-range gene activation. Mol Cell **9**:291-302.

Ho, Y., S. A. Liebhaber, and N. E. Cooke. 2004. Activation of the human GH gene cluster: roles for targeted chromatin modification. Trends Endocrinol Metab **15**:40-45.

Hoffmann, R., T. Seidl, and M. Dugas. 2002. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. Genome Biol **3**:RESEARCH0033.

Hoogendoorn, B., S. L. Coleman, C. A. Guy, K. Smith, T. Bowen, P. R. Buckland, and M. C. O'Donovan. 2003. Functional analysis of human promoter polymorphisms. Hum Mol Genet **12**:2249-2254.

Hu, S. L., and J. L. Manley. 1981. DNA sequence required for initiation of transcription in vitro from the major late promoter of adenovirus 2. Proc Natl Acad Sci U S A **78**:820-824.

Huppert, J. L., and S. Balasubramanian. 2005. Prevalence of quadruplexes in the human genome. Nucleic Acids Res **33**:2908-2916.

Huttley, G. A., M. W. Smith, M. Carrington, and S. J. O'Brien. 1999. A scan for linkage disequilibrium across the human genome. Genetics **152**:1711-1722.

Irizarry, R. A., Z. Wu, and H. A. Jaffee. 2006. Comparison of Affymetrix GeneChip expression measures. Bioinformatics **22**:789-794.

Iwama, H., and T. Gojobori. 2004. Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network. Proc Natl Acad Sci U S A **101**:17156-17161.

Javahery, R., A. Khachi, K. Lo, B. Zenzie-Gregory, and S. T. Smale. 1994. DNA sequence requirements for transcriptional initiator activity in mammalian cells. Mol Cell Biol **14**:116-127.

Jin, V. X., G. A. Singer, F. J. Agosto-Perez, S. Liyanarachchi, and R. V. Davuluri. 2006. Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. BMC Bioinformatics **7**:114.

Kanduri, C., V. Pant, D. Loukinov, E. Pugacheva, C. F. Qi, A. Wolffe, R. Ohlsson, and V. V. Lobanenkov. 2000. Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. Curr Biol **10**:853-856.

Kapranov, P., S. E. Cawley, J. Drenkow, S. Bekiranov, R. L. Strausberg, S. P. Fodor, and T. R. Gingeras. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. Science **296**:916-919.

Karaman, M. W., M. L. Houck, L. G. Chemnick, S. Nagpal, D. Chawannakul, D. Sudano, B. L. Pike, V. V. Ho, O. A. Ryder, and J. G. Hacia. 2003. Comparative analysis of gene-expression patterns in human and African great ape cultured fibroblasts. Genome Res **13**:1619-1630.

Kawaji, H., T. Kasukawa, S. Fukuda, S. Katayama, C. Kai, J. Kawai, P. Carninci, and Y. Hayashizaki. 2006. CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis. Nucleic Acids Res **34**:D632-636.

Kel, A. E., E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. 2003. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res **31**:3576-3579.

Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Paabo. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science **309**:1850-1854.

Khaitovich, P., S. Paabo, and G. Weiss. 2005. Toward a neutral evolutionary model of gene expression. Genetics **170**:929-939.

Khambata-Ford, S., Y. Liu, C. Gleason, M. Dickson, R. B. Altman, S. Batzoglou, and R. M. Myers. 2003. Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay. Genome Res **13**:1765-1774.

Khoury, G., and P. Gruss. 1983. Enhancer elements. Cell **33**:313-314.

Kim, J., S. Klooster, and D. J. Shapiro. 1995. Intrinsically bent DNA in a eukaryotic transcription factor recognition sequence potentiates transcription activation. J Biol Chem **270**:1282-1288.

Kim, T. H., L. O. Barrera, C. Qu, S. Van Calcar, N. D. Trinklein, S. J. Cooper, R. M. Luna, C. K. Glass, M. G. Rosenfeld, R. M. Myers, and B. Ren. 2005a. Direct isolation and identification of promoters in the human genome. Genome Res **15**:830-839.

Kim, T. H., L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green, and B. Ren. 2005b. A high-resolution map of active promoters in the human genome. Nature **436**:876-880.

King, D. C., J. Taylor, L. Elnitski, F. Chiaromonte, W. Miller, and R. C. Hardison. 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res **15**:1051-1060.

King, M. C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science **188**:107-116.

Kleinjan, D. A., A. Seawright, A. Schedl, R. A. Quinlan, S. Danes, and V. van Heyningen. 2001. Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. Hum Mol Genet **10**:2049-2059.

Knight, J. C. 2005. Regulatory polymorphisms underlying complex disease traits. J Mol Med **83**:97-109.

Knight, J. C., B. J. Keating, K. A. Rockett, and D. P. Kwiatkowski. 2003. In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. Nat Genet **33**:469-475.

Knudsen, S. 1999. Promoter2.0: for the recognition of PolII promoter sequences. Bioinformatics **15**:356-361.

Kolbe, D., J. Taylor, L. Elnitski, P. Eswara, J. Li, W. Miller, R. Hardison, and F. Chiaromonte. 2004. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. Genome Res **14**:700-707.

Kong, S., D. Bohl, C. Li, and D. Tuan. 1997. Transcription of the HS2 enhancer toward a cis-linked gene is independent of the orientation, position, and distance of the enhancer relative to the gene. Mol Cell Biol **17**:3955-3965.

Kumar, A., Y. Li, S. Patil, and S. Jain. 2005. A haplotype of the angiotensinogen gene is associated with hypertension in african americans. Clin Exp Pharmacol Physiol **32**:495-502.

Kummerfeld, S. K., and S. A. Teichmann. 2006. DBD: a transcription factor prediction database. Nucleic Acids Res **34**:D74-81.

Kutach, A. K., and J. T. Kadonaga. 2000. The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. Mol Cell Biol **20**:4754-4764.

Kuzmichev, A., and D. Reinberg. 2001. Role of histone deacetylase complexes in the regulation of chromatin metabolism. Curr Top Microbiol Immunol **254**:35-58.

Lagrange, T., A. N. Kapanidis, H. Tang, D. Reinberg, and R. H. Ebright. 1998. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. Genes Dev **12**:34-44.

Lagrange, T., T. K. Kim, G. Orphanides, Y. W. Ebright, R. H. Ebright, and D. Reinberg. 1996. High-resolution mapping of nucleoprotein complexes by site-specific protein-DNA photocrosslinking: organization of the human TBP-TFIIA-TFIIB-DNA quaternary complex. Proc Natl Acad Sci U S A **93**:10620-10625.

Landy, A. 1989. Dynamic, structural, and regulatory aspects of lambda site-specific recombination. Annu Rev Biochem **58**:913-949.

Lee, T. I., N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science **298**:799-804.

Legraverend, C., P. Antonson, P. Flodby, and K. G. Xanthopoulos. 1993. High level activity of the mouse CCAAT/enhancer binding protein (C/EBP alpha) gene promoter involves autoregulation and several ubiquitous transcription factors. Nucleic Acids Res **21**:1735-1742.

Lemon, B., and R. Tjian. 2000. Orchestrated response: a symphony of transcription factors for gene control. Genes Dev **14**:2551-2569.

Lerman, D. N., P. Michalak, A. B. Helin, B. R. Bettencourt, and M. E. Feder. 2003. Modification of heat-shock gene expression in Drosophila melanogaster populations via transposable elements. Mol Biol Evol **20**:135-144.

Lettice, L. A., T. Horikoshi, S. J. Heaney, M. J. van Baren, H. C. van der Linde, G. J. Breedveld, M. Joosse, N. Akarsu, B. A. Oostra, N. Endo, M. Shibata, M. Suzuki, E. Takahashi, T. Shinka, Y. Nakahori, D. Ayusawa, K. Nakabayashi, S. W. Scherer, P. Heutink, R. E. Hill, and S. Noji. 2002. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A **99**:7548-7553.

Lewin, B. 2003. Genes VIII. Prentice Hall.

Li, Q., S. Harju, and K. R. Peterson. 1999. Locus control regions: coming of age at a decade plus. Trends Genet **15**:403-408.

Li, S., F. C. MacLaughlin, J. G. Fewell, M. Gondo, J. Wang, F. Nicol, D. A. Dean, and L. C. Smith. 2001. Muscle-specific enhancement of gene expression by incorporation of SV40 enhancer in the expression plasmid. Gene Ther **8**:494-497.

Li, Y., P. M. Flanagan, H. Tschochner, and R. D. Kornberg. 1994. RNA polymerase II initiation factor interactions and transcription start site selection. Science **263**:805-807.

Li, Y., S. Jain, S. Patil, and A. Kumar. 2006. A haplotype of angiotensinogen gene that is associated with essential hypertension increases its promoter activity in adipocytes. Vascul Pharmacol **44**:29-33.

Li, Z., S. Van Calcar, C. Qu, W. K. Cavenee, M. Q. Zhang, and B. Ren. 2003. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. Proc Natl Acad Sci U S A **100**:8164-8169.

Lim, C. Y., B. Santoso, T. Boulay, E. Dong, U. Ohler, and J. T. Kadonaga. 2004. The MTE, a new core promoter element for transcription by RNA polymerase II. Genes Dev **18**:1606-1617.

Lin, C. H., S. Y. Hsieh, I. S. Sheen, W. C. Lee, T. C. Chen, W. C. Shyu, and Y. F. Liaw. 2001. Genome-wide hypomethylation in hepatocellular carcinogenesis. Cancer Res **61**:4238-4243.

Lin, S., D. J. Cutler, M. E. Zwick, and A. Chakravarti. 2002. Haplotype inference in random population samples. Am J Hum Genet **71**:1129-1137.

Linnell, J., R. Mott, S. Field, D. P. Kwiatkowski, J. Ragoussis, and I. A. Udalova. 2004. Quantitative high-throughput analysis of transcription factor binding specificities. Nucleic Acids Res **32**:e44.

Lo, H. S., Z. Wang, Y. Hu, H. H. Yang, S. Gere, K. H. Buetow, and M. P. Lee. 2003. Allelic variation in gene expression is common in the human genome. Genome Res **13**:1855-1862.

Ludwig, M. Z., C. Bergman, N. H. Patel, and M. Kreitman. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403**:564-567.

Martone, R., G. Euskirchen, P. Bertone, S. Hartman, T. E. Royce, N. M. Luscombe, J. L. Rinn, F. K. Nelson, P. Miller, M. Gerstein, S. Weissman, and M. Snyder. 2003. Distribution of NF-kappaB-binding sites across human chromosome 22. Proc Natl Acad Sci U S A **100**:12247-12252.

Mastrangelo, I. A., A. J. Courey, J. S. Wall, S. P. Jackson, and P. V. Hough. 1991. DNA looping and Sp1 multimer links: a mechanism for transcriptional synergism and enhancement. Proc Natl Acad Sci U S A **88**:5670-5674.

McGuire, W., A. V. Hill, C. E. Allsopp, B. M. Greenwood, and D. Kwiatkowski. 1994. Variation in the TNF-alpha promoter region associated with susceptibility to cerebral malaria. Nature **371**:508-510.
McKusick, V. A. 1998. Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. Johns Hopkins University Press, Baltimore.
Monks, S. A., A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak, S. Edwards, J. W. Phillips, A. Sachs, and E. E. Schadt. 2004. Genetic inheritance of gene expression in human cell lines. Am J Hum Genet **75**:1094-1105.
Montgomery, S. B., O. L. Griffith, M. C. Sleumer, C. M. Bergman, M. Bilenky, E. D. Pleasance, Y. Prychyna, X. Zhang, and S. J. Jones. 2006. ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. Bioinformatics **22**:637-640.
Morison, I. M., J. P. Ramsay, and H. G. Spencer. 2005. A census of mammalian imprinting. Trends Genet **21**:457-465.
Morley, M., C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung. 2004. Genetic analysis of genome-wide variation in human gene expression. Nature **430**:743-747.
Mottagui-Tabar, S., M. A. Faghihi, Y. Mizuno, P. G. Engstrom, B. Lenhard, W. W. Wasserman, and C. Wahlestedt. 2005. Identification of functional SNPs in the 5-prime flanking sequences of human genes. BMC Genomics **6**:18.
Nagaich, A. K., E. Appella, and R. E. Harrington. 1997. DNA bending is essential for the site-specific recognition of DNA response elements by the DNA binding domain of the tumor suppressor protein p53. J Biol Chem **272**:14842-14849.
Nakashima, K., T. Hirota, K. Obara, M. Shimizu, S. Doi, K. Fujita, T. Shirakawa, T. Enomoto, S. Yoshihara, M. Ebisawa, K. Matsumoto, H. Saito, Y. Suzuki, Y. Nakamura, and M. Tamari. 2006. A functional polymorphism in MMP-9 is associated with childhood atopic asthma. Biochem Biophys Res Commun **344**:300-307.
Nightingale, K. P., L. P. O'Neill, and B. M. Turner. 2006. Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. Curr Opin Genet Dev **16**:125-136.
Nikolov, D. B., H. Chen, E. D. Halay, A. A. Usheva, K. Hisatake, D. K. Lee, R. G. Roeder, and S. K. Burley. 1995. Crystal structure of a TFIIB-TBP-TATA-element ternary complex. Nature **377**:119-128.
Nishikawa, J., M. Amano, Y. Fukue, S. Tanaka, H. Kishi, Y. Hirota, K. Yoda, and T. Ohyama. 2003. Left-handedly curved DNA regulates accessibility to cis-DNA elements in chromatin. Nucleic Acids Res **31**:6651-6662.
Niu, T., Z. S. Qin, X. Xu, and J. S. Liu. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet **70**:157-169.
O'Shea-Greenfield, A., and S. T. Smale. 1992. Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription. J Biol Chem **267**:6450.
Odom, D. T., N. Zizlsperger, D. B. Gordon, G. W. Bell, N. J. Rinaldi, H. L. Murray, T. L. Volkert, J. Schreiber, P. A. Rolfe, D. K. Gifford, E. Fraenkel, G. I. Bell, and R. A. Young. 2004. Control of pancreas and liver gene expression by HNF transcription factors. Science **303**:1378-1381.
Ohler, U., G. C. Liao, H. Niemann, and G. M. Rubin. 2002. Computational analysis of core promoters in the Drosophila genome. Genome Biol **3**:RESEARCH0087.

Park, J. Y., J. M. Park, J. S. Jang, J. E. Choi, K. M. Kim, S. I. Cha, C. H. Kim, Y. M. Kang, W. K. Lee, S. Kam, R. W. Park, I. S. Kim, J. T. Lee, and T. H. Jung. 2006. Caspase 9 promoter polymorphisms and risk of primary lung cancer. Hum Mol Genet **15**:1963-1971.

Parvin, J. D., R. J. McCormick, P. A. Sharp, and D. E. Fisher. 1995. Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. Nature **373**:724-727.

Pastinen, T., B. Ge, S. Gurd, T. Gaudin, C. Dore, M. Lemire, P. Lepage, E. Harmsen, and T. J. Hudson. 2005. Mapping common regulatory variants to human haplotypes. Hum Mol Genet **14**:3963-3971.

Pastinen, T., B. Ge, and T. J. Hudson. 2006. Influence of human genome polymorphism on gene expression. Hum Mol Genet **15 Suppl 1**:R9-R16.

Pastinen, T., and T. J. Hudson. 2004. Cis-acting regulatory variation in the human genome. Science **306**:647-650.

Pastinen, T., R. Sladek, S. Gurd, A. Sammak, B. Ge, P. Lepage, K. Lavergne, A. Villeneuve, T. Gaudin, H. Brandstrom, A. Beck, A. Verner, J. Kingsley, E. Harmsen, D. Labuda, K. Morgan, M. C. Vohl, A. K. Naumova, D. Sinnett, and T. J. Hudson. 2004. A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics **16**:184-193.

Pazin, M. J., and J. T. Kadonaga. 1997. What's up and down with histone deacetylation and transcription? Cell **89**:325-328.

Ponger, L., and D. Mouchiroud. 2002. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics **18**:631-633.

Pritchard, J. K., and M. Przeworski. 2001. Linkage disequilibrium in humans: models and data. Am J Hum Genet **69**:1-14.

Przeworski, M., R. R. Hudson, and A. Di Rienzo. 2000. Adjusting the focus on human variation. Trends Genet **16**:296-302.

Ptashne, M. 1992. A Genetic Switch: Phage (Lambda) and Higher Organisms. Cell Press, Cambridge, MA.

Ragoczy, T., A. Telling, T. Sawado, M. Groudine, and S. T. Kosak. 2003. A genetic analysis of chromosome territory looping: diverse roles for distal regulatory elements. Chromosome Res **11**:513-525.

Rastegar, M., F. P. Lemaigre, and G. G. Rousseau. 2000. Control of gene expression by growth hormone in liver: key role of a network of transcription factors. Mol Cell Endocrinol **164**:1-4.

Reese, M. G. 2001. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. Comput Chem **26**:51-56.

Reik, W., and J. Walter. 2001. Genomic imprinting: parental influence on the genome. Nat Rev Genet **2**:21-32.

Reiter, W. D., U. Hudepohl, and W. Zillig. 1990. Mutational analysis of an archaebacterial promoter: essential role of a TATA box for transcription efficiency and start-site selection in vitro. Proc Natl Acad Sci U S A **87**:9509-9513.

Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. 2000. Genome-wide location and function of DNA binding proteins. Science **290**:2306-2309.

Rezai-Zadeh, N., X. Zhang, F. Namour, G. Fejer, Y. D. Wen, Y. L. Yao, I. Gyory, K. Wright, and E. Seto. 2003. Targeted recruitment of a histone H4-specific methyltransferase by the transcription factor YY1. Genes Dev **17**:1019-1029.

Rhodes, S. J., R. Chen, G. E. DiMattia, K. M. Scully, K. A. Kalla, S. C. Lin, V. C. Yu, and M. G. Rosenfeld. 1993. A tissue-specific enhancer confers Pit-1-dependent morphogen inducibility and autoregulation on the pit-1 gene. Genes Dev **7**:913-932.

Rice, J. C., S. D. Briggs, B. Ueberheide, C. M. Barber, J. Shabanowitz, D. F. Hunt, Y. Shinkai, and C. D. Allis. 2003. Histone methyltransferases direct different degrees of methylation to define distinct chromatin domains. Mol Cell **12**:1591-1598.

Rifkin, S. A., J. Kim, and K. P. White. 2003. Evolution of gene expression in the Drosophila melanogaster subgroup. Nat Genet **33**:138-144.

Rinn, J. L., G. Euskirchen, P. Bertone, R. Martone, N. M. Luscombe, S. Hartman, P. M. Harrison, F. K. Nelson, P. Miller, M. Gerstein, S. Weissman, and M. Snyder. 2003. The transcriptional activity of human Chromosome 22. Genes Dev **17**:529-540.

Robertson, G., M. Bilenky, K. Lin, A. He, W. Yuen, M. Dagpinar, R. Varhol, K. Teague, O. L. Griffith, X. Zhang, Y. Pan, M. Hassel, M. C. Sleumer, W. Pan, E. D. Pleasance, M. Chuang, H. Hao, Y. Y. Li, N. Robertson, C. Fjell, B. Li, S. B. Montgomery, T. Astakhova, J. Zhou, J. Sander, A. S. Siddiqui, and S. J. Jones. 2006. cisRED: a database system for genome-scale computational discovery of regulatory elements. Nucleic Acids Res **34**:D68-73.

Rockman, M. V., and G. A. Wray. 2002. Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol **19**:1991-2004.

Ross, D. T., U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. 2000. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet **24**:227-235.

Roth, F. P., J. D. Hughes, P. W. Estep, and G. M. Church. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol **16**:939-945.

Roth, S. Y., J. M. Denu, and C. D. Allis. 2001. Histone acetyltransferases. Annu Rev Biochem **70**:81-120.

Rothenburg, S., F. Koch-Nolte, A. Rich, and F. Haag. 2001a. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. Proc Natl Acad Sci U S A **98**:8985-8990.

Rothenburg, S., F. Koch-Nolte, H. G. Thiele, and F. Haag. 2001b. DNA methylation contributes to tissue- and allele-specific expression of the T-cell differentiation marker RT6. Immunogenetics **52**:231-241.

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol **132**:365-386.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, and D.

Altshuler. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409**:928-933.

Saito, T., F. Guan, D. F. Papolos, N. Rajouria, C. S. Fann, and H. M. Lachman. 2001. Polymorphism in SNAP29 gene promoter region associated with schizophrenia. Mol Psychiatry **6**:193-201.

Sandelin, A., W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res **32**:D91-94.

Sawado, T., J. Halow, M. A. Bender, and M. Groudine. 2003. The beta -globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. Genes Dev **17**:1009-1018.

Schubeler, D., M. Groudine, and M. A. Bender. 2001. The murine beta-globin locus control region regulates the rate of transcription but not the hyperacetylation of histones at the active genes. Proc Natl Acad Sci U S A **98**:11432-11437.

Schubeler, D., D. M. MacAlpine, D. Scalzo, C. Wirbelauer, C. Kooperberg, F. van Leeuwen, D. E. Gottschling, L. P. O'Neill, B. M. Turner, J. Delrow, S. P. Bell, and M. Groudine. 2004. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev **18**:1263-1271.

Seenisamy, J., E. M. Rezler, T. J. Powell, D. Tye, V. Gokhale, C. S. Joshi, A. Siddiqui-Jain, and L. H. Hurley. 2004. The dynamic character of the G-quadruplex element in the c-MYC promoter and modification by TMPyP4. J Am Chem Soc **126**:8702-8709.

Segal, J. A., J. L. Barnett, and D. L. Crawford. 1999. Functional analyses of natural variation in Sp1 binding sites of a TATA-less promoter. J Mol Evol **49**:736-749.

Shaw, M. A., I. J. Donaldson, A. Collins, C. S. Peacock, Z. Lins-Lainson, J. J. Shaw, F. Ramos, F. Silveira, and J. M. Blackwell. 2001. Association and linkage of leprosy phenotypes with HLA class II and tumour necrosis factor genes. Genes Immun **2**:196-204.

Shedden, K., W. Chen, R. Kuick, D. Ghosh, J. Macdonald, K. R. Cho, T. J. Giordano, S. B. Gruber, E. R. Fearon, J. M. Taylor, and S. Hanash. 2005. Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. BMC Bioinformatics **6**:26.

Shin, H. D., C. Winkler, J. C. Stephens, J. Bream, H. Young, J. J. Goedert, T. R. O'Brien, D. Vlahov, S. Buchbinder, J. Giorgi, C. Rinaldo, S. Donfield, A. Willoughby, S. J. O'Brien, and M. W. Smith. 2000. Genetic restriction of HIV-1 pathogenesis to AIDS by promoter alleles of IL10. Proc Natl Acad Sci U S A **97**:14467-14472.

Shiraki, T., S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A **100**:15776-15781.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res **15**:1034-1050.

Smale, S. T. 1997. Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. Biochim Biophys Acta **1351**:73-88.
Smale, S. T., and D. Baltimore. 1989. The "initiator" as a transcription control element. Cell **57**:103-113.
Snoussi, K., W. Mahfoudh, N. Bouaouina, S. B. Ahmed, A. N. Helal, and L. Chouchane. 2006. Genetic variation in IL-8 associated with increased risk and poor prognosis of breast carcinoma. Hum Immunol **67**:13-21.
Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell **9**:3273-3297.
Spiecker, M., H. Darius, T. Hankeln, M. Soufi, A. M. Sattler, J. R. Schaefer, K. Node, J. Borgel, A. Mugge, K. Lindpaintner, A. Huesing, B. Maisch, D. C. Zeldin, and J. K. Liao. 2004. Risk of coronary artery disease associated with polymorphism of the cytochrome P450 epoxygenase CYP2J2. Circulation **110**:2132-2136.
Spitz, F., F. Gonzalez, and D. Duboule. 2003. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. Cell **113**:405-417.
Sprecher, E., A. Ishida-Yamamoto, M. Mizrahi-Koren, D. Rapaport, D. Goldsher, M. Indelman, O. Topaz, I. Chefetz, H. Keren, J. O'Brien T, D. Bercovich, S. Shalev, D. Geiger, R. Bergman, M. Horowitz, and H. Mandel. 2005. A mutation in SNAP29, coding for a SNARE protein involved in intracellular trafficking, causes a novel neurocutaneous syndrome characterized by cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma. Am J Hum Genet **77**:242-251.
Stanford, W. L., J. B. Cohn, and S. P. Cordes. 2001. Gene-trap mutagenesis: past, present and beyond. Nat Rev Genet **2**:756-768.
Stephens, M., and P. Scheet. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet **76**:449-462.
Stephens, M., N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet **68**:978-989.
Sterner, D. E., and S. L. Berger. 2000. Acetylation of histones and transcription-related factors. Microbiol Mol Biol Rev **64**:435-459.
Stranger, B. E., M. S. Forrest, A. G. Clark, M. J. Minichiello, S. Deutsch, R. Lyle, S. Hunt, B. Kahl, S. E. Antonarakis, S. Tavare, P. Deloukas, and E. T. Dermitzakis. 2005. Genome-Wide Associations of Gene Expression Variation in Humans. PLoS Genet **1**:e78.
Strathdee, G., A. Sim, and R. Brown. 2004. Control of gene expression by CpG island methylation in normal cells. Biochem Soc Trans **32**:913-915.
Strathdee, G., A. Sim, R. Soutar, T. L. Holyoake, and R. Brown. 2006. HOXA5 is targeted by cell type specific CpG island methylation in normal cells and during the development of acute myeloid leukaemia. Carcinogenesis.
Suzuki, M., and N. Yagi. 1995. Stereochemical basis of DNA bending by transcription factors. Nucleic Acids Res **23**:2083-2091.
Suzuki, Y., T. Tsunoda, J. Sese, H. Taira, J. Mizushima-Sugano, H. Hata, T. Ota, T. Isogai, T. Tanaka, Y. Nakamura, A. Suyama, Y. Sakaki, S. Morishita, K. Okubo, and S. Sugano. 2001. Identification and characterization of the

potential promoter regions of 1031 kinds of human genes. Genome Res **11**:677-684.
Suzuki, Y., R. Yamashita, M. Shirota, Y. Sakakibara, J. Chiba, J. Mizushima-Sugano, K. Nakai, and S. Sugano. 2004. Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. Genome Res **14**:1711-1718.
Talkington, C. A., and P. Leder. 1982. Rescuing the in vitro function of a globin pseudogene promoter. Nature **298**:192-195.
Thijs, G., K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. 2002. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol **9**:447-464.
Timchenko, N., D. R. Wilson, L. R. Taylor, S. Abdelsayed, M. Wilde, M. Sawadogo, and G. J. Darlington. 1995. Autoregulation of the human C/EBP alpha gene by stimulation of upstream stimulatory factor binding. Mol Cell Biol **15**:1192-1202.
Ting, J. P., and J. Trowsdale. 2002. Genetic control of MHC class II expression. Cell **109 Suppl**:S21-33.
Tolhuis, B., R. J. Palstra, E. Splinter, F. Grosveld, and W. de Laat. 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell **10**:1453-1465.
Tournamille, C., Y. Colin, J. P. Cartron, and C. Le Van Kim. 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat Genet **10**:224-228.
Trinklein, N. D., S. F. Aldred, S. J. Hartman, D. I. Schroeder, R. P. Otillar, and R. M. Myers. 2004. An abundance of bidirectional promoters in the human genome. Genome Res **14**:62-66.
Trinklein, N. D., S. J. Aldred, A. J. Saldanha, and R. M. Myers. 2003. Identification and functional analysis of human transcriptional promoters. Genome Res **13**:308-312.
Umbricht, C. B., E. Evron, E. Gabrielson, A. Ferguson, J. Marks, and S. Sukumar. 2001. Hypermethylation of 14-3-3 sigma (stratifin) is an early event in breast cancer. Oncogene **20**:3348-3353.
Wang, R. L., A. Stec, J. Hey, L. Lukens, and J. Doebley. 1999. The limits of selection during maize domestication. Nature **398**:236-239.
Warren, P., J. Bienkowska, P. G. V. Martini, J. Jackson, and D. M. Taylor. 2006. PANP - a New Method of Gene Detection on Oligonucleotide Expression Arrays. Research in Computational Molecular Biology, Venice, Italy.
Wasserman, W. W., and A. Sandelin. 2004. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet **5**:276-287.
Wasylyk, B., R. Derbyshire, A. Guy, D. Molko, A. Roget, R. Teoule, and P. Chambon. 1980. Specific in vitro transcription of conalbumin gene is drastically decreased by single-point mutation in T-A-T-A box homology sequence. Proc Natl Acad Sci U S A **77**:7024-7028.
Weisberg, R., and A. Landy. 1983. Site-Specific Recombination in Phage Lambda. Pp. 211-250 *in* R. Weisberg, ed. Lambda II. Cold Spring Harbor Press, Cold Spring Harbor, NY.
West, A. G., and P. Fraser. 2005. Remote control of gene transcription. Hum Mol Genet **14 Spec No 1**:R101-111.
West, A. G., M. Gaszner, and G. Felsenfeld. 2002. Insulators: many functions, many mechanisms. Genes Dev **16**:271-288.

Whitehead, A. S., and R. Sackstein. 1985. Molecular biology of the human and mouse MHC class III genes: phylogenetic conservation, genetics and regulation of expression. Immunol Rev **87**:185-208.
Wingender, E., P. Dietze, H. Karas, and R. Knuppel. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res **24**:238-241.
Wittkopp, P. J., B. K. Haerum, and A. G. Clark. 2004. Evolutionary changes in cis and trans gene regulation. Nature **430**:85-88.
Wonodi, I., L. E. Hong, M. T. Avila, R. W. Buchanan, W. T. Carpenter, Jr., O. C. Stine, B. D. Mitchell, and G. K. Thaker. 2005. Association between polymorphism of the SNAP29 gene promoter region and schizophrenia. Schizophr Res **78**:339-341.
Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. 2003. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol **20**:1377-1419.
Wright, S., E. deBoer, A. Rosenthal, R. A. Flavell, and F. Grosveld. 1984. DNA sequences required for regulated expression of beta-globin genes in murine erythroleukaemia cells. Philos Trans R Soc Lond B Biol Sci **307**:271-282.
Wu, Z., R. Irizarry, R. Gentlemen, F. Martinez-Murillo, and F. Spencer. 2004. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of the American Statistical Association **99**:909-917.
Xie, X., J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature **434**:338-345.
Xu, L. C., M. Thali, and W. Schaffner. 1991. Upstream box/TATA box order is the major determinant of the direction of transcription. Nucleic Acids Res **19**:6699-6704.
Yamamoto, K. R. 1985. Steroid receptor regulated transcription of specific genes and gene networks. Annu Rev Genet **19**:209-252.
Yan, H., Z. Dobbie, S. B. Gruber, S. Markowitz, K. Romans, F. M. Giardiello, K. W. Kinzler, and B. Vogelstein. 2002a. Small changes in expression affect predisposition to tumorigenesis. Nat Genet **30**:25-26.
Yan, H., W. Yuan, V. E. Velculescu, B. Vogelstein, and K. W. Kinzler. 2002b. Allelic variation in human gene expression. Science **297**:1143.

7 Appendices

Appendix A – Genes targeted for promoter re-sequencing

RRP22	*CHEK2*	*SERHL*
AP1B1	*ZBED4*	*POLDIP3*
GAS2L1	*RBX1*	*CYB5R3*
GNAZ	*OTTHUMG00000030682*	*NDUFA6*
RTDR1	*TFIP11*	*AK123630*
NEFH	*ASPHD2*	*SEZ6L*
ACR	*SRR1L*	*TIMP3*
MRPL40	*HPS4*	*PACSIN2*
UFD1L	*BIK*	*TTLL1*
OTTHUMG00000030822	*MT*	*DNAL4*
CDC45L	*MYO18B*	*NULL*
TXNRD2	*CGI-96*	*CBX6*
COMT	*APOBEC3A*	*RPL3*
HIRA	*ZNRF3*	*SYNGR1*
CLDN5	*XRCC6*	*FLJ2358*
GNB1L	*NHP2L1*	*CSDC2*
TBX1	*OTTHUMG00000030205*	*PMM1*
SEPT5	*NUP50*	*D15Wsu75e*
GP1BB	*CRYBB3*	*POLR3H*
DGCR2	*CRYBB2*	*PITPNB*
RAB36	*OTTHUMG00000030164*	*MN1b*
LIMK2	*TOB2*	*SLC25A17*
MGC17330	*PHF5A*	*FLJ33814*
HORMAD2	*PLA2G6*	*HSC20*
CRKL	*GPR24*	*GRAP2*
P2RXL1	*SREBF2*	*FAM83F*
LZTR1	*C22orf18*	*SH3BP1*
SLC7A4	*SEPT3*	*LGALS1*
AIFL	*NAGA*	*PDXP*
SLC25A18	*MGC26816*	*MGC3731*
PCQAP	*TNFRSF13C*	*SULT4A1*
OSM	*OTTHUMG00000030501*	*PNPLA5*
LIF	*C22orf9*	*MAPK12*
OTTHUMG00000030137	*UPK3A*	*MAPK11*
CLTCL1	*C22orf8*	*PP2447*
SLC25A1	*SMC1L2*	*SELO*
DGCR14	*LARGE*	*TUBGCP6*
TSSK2	*HMOX1*	*HDAC10*
MORC2	*MCM5*	*MOV10L1*
OTTHUMG00000030444	*XBP1*	*PANX2*
MTP18	*OTTHUMG00000030528*	*MAP3K7IP1*

SEC14L2
ASCC2
TBC1D10A
AB051443
TBC1D10A
SF3A1
LOC200312
LOC550631
ZNF278
OTTHUMG00000030356
DEPDC5
PIB5PA
SMTN
PLA2G3
RNF185
PES1
GAL3st1
TCN2
SLC35E4
CECR1
XKR3
ZNF74
SCARF2
KLHL22
NF2
NIPSNAP1
ZMAT5
CABP7
ARVCF
BID
BCL2L13
ATP6V1E1
RANBP1
C22orf25
DGCR8
ZDHHC8
AK057137
CECR5
CECR6
IL17R
DKFZp434N035
PIK4CA
SNAP29
SERPIND1
CERK
MFNG
PARVB
MN1
CRYBB1
CRYBA4
TPst2
MAFF
KREMEN1
HS747E2A
NCF4
CSF2RB
FBLN1
ATXN10
EMID1
C22orf3
EWSR1
MPPED1
SLC5A1
A4GALT
OTTHUMG00000030672
Tst
MPst
OTTHUMG00000030172
PDGFB
OTTHUMG00000030676
SYN3
YWHAH
FLJ20699
OTTHUMG00000030329
PPM1F
SUHW1
PRAME
TOP3B
VPREB1
EIF3S6IP
MICAL-L1
RIBC2
LDOC1L
PRKCABP
SLC16A8
BAIAP2L2
SOX10
POLR2F
DNAJB7
RBM9
APOL5
OSBP2
DDX17
KCNJ4
KDELR3
GGA1
OTTHUMG00000030404
H1F0
GCAT
GALR3
OTTHUMG00000030664
PGEA1
JOSD1
GTPBP1
UNC84B
TOM1
HMG2L1
BRD1
BZRP
TTLL12
SCUBE1
ADPN
RASD2
SBF1
MIOX
ADM2
TCF20
CSNK1E
C22orf5
PSCD4
KIAA1904
PARVG
BC104183
MYH9
OTTHUMG00000030139
PKDREJ
FLJ27365
APOBEC3B
CBX7
RANGAP1
L3MBTL2
SAMM50

USP18
TUBA8
PEX26
AK000085
ZNF70
VPREB3
C22orf16
MMP11
SMARCB1
DERL3
SLC2A11
MIF
GstT1
CABIN1
ADORA2A
UPB1
C22orf13
SNRPD3
LOC388886
UBE2L3
PPIL2
YPEL1
SDF2L1
LOC150223
FLJ36046
MAPK1
DRG1
EIF4ENIF1
EIF4ENIF1
FAM19A5
C22orf23
ADSL
RUTBC3
MKL1
MGAT3
ATF4
NULL
FLJ20232
TXN2
EIF3S7
FLJ23322
TRMU
GTSE1
CELSR1
KCTD17
TMPRSS6
IL2RB
CDC42EP1
LGALS2
PHF21B
OTTHUMG00000058273
HSPC117
FBXO7
RAC2
C1QTNF6
CACNA1I
ARHGAP8
OTTHUMG00000030175
FAM109B
RAXLX
MEI1
FLJ22349
SFI1
PISD
EP300
CARD10
RFPL2
SLC5A4
RUTBC2
TEF
ZC3H7B
PVALB
MB
APOL6
TBC1D22A
BCR
AK057318
OTTHUMG00000030810
KLHDC7B
OTTHUMG00000030815
hCAP-H2
SCO2
ECGF1
CPT1B
CHKB
MAPK8IP2
ARSA
TMEM153

Appendix B - SNPs discovered in promoter re-sequencing

The names of the genes whose promoters were re-sequences are listed in order of occurrence on chromosome 22, from the centromeric to the telomeric end of the q arm.

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
XKR3	15677239	-103	rs2891848	T	A	0.45
XKR3	15677328	-192	rs5016127	G	C	0.46
XKR3	15677357	-221	rs5016128	G	A	0.48
XKR3	15677377	-241	rs12484826	T	C	0.28
XKR3	15677594	-458	rs5992556	T	C	0.45
XKR3	15677830	-694	NT_011519.10_455426	A	T	0.09
XKR3	15677937	-801	rs5994031	T	C	0.46
XKR3	15678018	-882	rs2215841	A	C	0.42
XKR3	15678150	-1014	rs2192431	T	G	0.09
XKR3	15678475	-1339	rs9606477	A	T	0.43
XKR3	15678686	-1550	rs175138	A	G	0.35
XKR3	15678800	-1664	rs175139	T	C	0.45
XKR3	15679011	-1875	rs5994033	T	C	0.33
IL17R	15939543	-877	NT_011519.10_717139	G	A	0.01
IL17R	15939567	-853	rs4819958	G	A	0.16
IL17R	15939589	-831	rs4819554	A	G	0.16
CECR6	15977393	-698	rs5992629	A	G	0.14
CECR6	15977999	-1304	NT_011519.10_755595	C	T	0.06
CECR6	15978031	-1336	rs5748871	A	G	0.31

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
CECR6	15978057	-1362	rs28360663	G	A	0.07
CECR6	15978355	-1660	rs5746996	C	A	0.36
CECR6	15978377	-1682	rs5748872	G	A	0.06
CECR6	15978426	-1731	NT_011519.10_756022	G	A	0.01
CECR5	16015633	-914	rs5748917	C	T	0.41
CECR5	16016239	-1520	NT_011519.10_793835	C	T	0.12
CECR5	16016466	-1747	rs5747015	A	G	0.35
CECR1	16075858	-991	rs9619019	C	A	0.09
CECR1	16076365	-1498	rs737970	C	A	0.36
CECR1	16076600	-1733	rs737969	A	G	0.45
CECR1	16076804	-1937	rs1807519	T	C	0.44
SLC25A18	16416647	-1061	rs174357	T	A	0.11
SLC25A18	16416930	-778	rs1296805	T	G	0.15
BCL2L13	16484411	-1824	NT_011519.10_1262007	G	A	0.05
BCL2L13	16484502	-1733	rs5992769	C	G	0.48
BCL2L13	16484517	-1718	rs4449236	T	C	0.43
BCL2L13	16485985	-250	rs17526598	T	C	0.27
BCL2L13	16486110	-125	rs5746448	A	C	0.13
BCL2L13	16486197	-38	NT_011519.10_1263793	C	T	0.01
BID	16632898	-1089	rs8190256	T	G	0.03
BID	16632936	-1127	rs366542	C	T	0.38
PEX26	16933643	-1599	NT_011519.10_1711239	T	C	0.05
PEX26	16935124	-118	rs462055	C	T	0.22
PEX26	16935151	-91	NT_011519.10_1712747	T	C	0.01

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
PEX26	16935165	-77	rs12157958	T	G	0.02
AK000085	16937288	-1086	NT_011519.10_1714884	T	C	0.01
AK000085	16937841	-1639	rs362146	G	A	0.07
AK000085	16937856	-1654	NT_011519.10_1715452	T	C	0.01
USP18	17005919	-1344	rs9618216	C	T	0.07
USP18	17006608	-655	rs9617680	G	C	0.09
DGCR2	17484963	-467	rs17526612	A	G	0.32
DGCR2	17486114	-1618	NT_011519.10_2263710	C	A	0.14
TSSK2	17491378	-1015	NT_011519.10_2268974	T	C	0.15
TSSK2	17491441	-952	NT_011519.10_2269037	T	A	0.12
TSSK2	17491821	-572	rs8139221	T	C	0.36
TSSK2	17492200	-193	NT_011519.10_2269796	A	G	0.03
DGCR14	17506879	-152	rs737923	A	G	0.39
DGCR14	17506933	-206	NT_011519.10_2284529	G	T	0.01
DGCR14	17506934	-207	rs1936950	A	T	0.12
DGCR14	17506939	-212	rs1936951	G	A	0.12
DGCR14	17507658	-931	rs5748005	C	A	0.23
DGCR14	17508159	-1432	rs715544	G	A	0.13
DGCR14	17508470	-1743	rs4819776	C	T	0.29
DGCR14	17508576	-1849	rs4819777	G	A	0.28
DGCR14	17508637	-1910	rs4819778	C	T	0.41
DGCR14	17508713	-1986	rs7289913	C	T	0.14
SLC25A1	17541633	-798	rs5746674	G	A	0.01
SLC25A1	17541939	-1104	rs738904	C	A	0.41

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
SLC25A1	17541963	-1128	rs712958	T	C	0.48
SLC25A1	17541975	-1140	rs5746675	G	A	0.01
SLC25A1	17542355	-1520	rs2793062	C	T	0.16
SLC25A1	17542641	-1806	rs2800974	A	G	0.35
CLTCL1	17654305	-556	NT_011519.10_2431901	A	T	0.04
CLTCL1	17654722	-973	NT_011519.10_2432318	A	G	0.15
CLTCL1	17654957	-1208	rs3810597	G	T	0.08
CDC45L	17840789	-1186	rs13447177	C	T	0.07
CDC45L	17840903	-1072	rs5748231	G	T	0.46
CDC45L	17840939	-1036	rs5993649	A	G	0.5
UFD1L	17841440	-70	rs5992403	C	T	0.46
CDC45L	17841783	-192	rs13447180	T	G	0.01
CDC45L	17841851	-124	rs4141528	C	G	0.01
UFD1L	17842016	-646	rs13447182	G	T	0.1
UFD1L	17842206	-836	rs5748232	C	T	0.34
UFD1L	17842271	-901	rs4141527	C	T	0.01
UFD1L	17842441	-1071	rs13447184	G	A	0.01
UFD1L	17843068	-1698	rs13447189	T	C	0.03
UFD1L	17843209	-1839	rs5993650	T	C	0.46
UFD1L	17843226	-1856	rs5746745	G	A	0.03
CLDN5	17887472	-32	rs5748258	C	T	0.03
CLDN5	17888043	-603	rs11705109	T	G	0.03
CLDN5	17888246	-806	NT_011519.10_2665842	C	T	0.02
CLDN5	17888430	-990	rs9606048	C	T	0.48

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
CLDN5	17888451	-1011	rs16983833	C	T	0.01
CLDN5	17888484	-1044	rs2871029	G	A	0.08
CLDN5	17888552	-1112	rs739371	G	C	0.22
CLDN5	17888567	-1127	rs9604969	A	G	0.07
CLDN5	17888805	-1365	NT_011519.10_2666401	G	A	0.01
CLDN5	17888950	-1510	rs739370	T	C	0.5
CLDN5	17889021	-1581	NT_011519.10_2666617	C	T	0.03
TBX1	18117978	-801	NT_011519.10_2895574	C	T	0.28
GNB1L	18217302	-288	rs28451568	A	G	0.06
COMT	18303438	-411	rs2020917	T	C	0.48
COMT	18303581	-268	rs13306278	C	T	0.25
TXNRD2	18304675	-608	rs737865	A	G	0.26
TXNRD2	18304713	-646	rs737864	C	T	0.31
TXNRD2	18305364	-1297	NT_011519.10_3082960	T	C	0.06
TXNRD2	18305727	-1660	rs933270	T	A	0.28
ARVCF	18379491	-630	rs2531717	C	G	0.4
ARVCF	18380698	-1837	rs2531700	C	T	0.3
ARVCF	18380908	-2047	rs5748501	G	C	0.35
C22orf25	18381266	-1919	rs2531702	G	A	0.02
C22orf25	18382153	-1032	rs7288996	G	A	0.27
C22orf25	18382173	-1012	rs5746851	C	T	0.34
DGCR8	18441352	-973	NT_011519.10_3218948	T	G	0.04
RANBP1	18477970	-1503	NT_011519.10_3255566	A	G	0.01
RANBP1	18478399	-1074	NT_011519.10_3255995	G	A	0.01

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
RANBP1	18479333	-140	NT_011519.10_3256929	C	T	0.02
RANBP1	18479373	-100	rs713982	C	A	0.25
RANBP1	18479407	-66	rs2286929	T	G	0.27
ZDHHC8	18492272	-1735	NT_011519.10_3269868	G	C	0.01
ZDHHC8	18492906	-1101	NT_011519.10_3270502	G	C	0.01
ZDHHC8	18493004	-1003	NT_011519.10_3270600	T	C	0.02
AK057137	18566885	-1588	rs9605084	C	G	0.2
AK057137	18567098	-1375	rs12485013	G	A	0.13
AK057137	18567653	-820	NT_011519.10_3345249	C	T	0.08
AK057137	18567668	-805	rs640836	C	A	0.14
ZNF74	19072317	-696	rs17551325	G	T	0.44
ZNF74	19072607	-406	rs17551339	G	A	0.31
ZNF74	19072807	-206	rs17551346	C	T	0.41
SCARF2	19118468	-1801	rs1035239	T	C	0.38
SCARF2	19118476	-1809	NT_011520.10_184491	A	G	0.02
KLHL22	19175978	-1305	rs9608041	C	T	0.06
KLHL22	19176004	-1331	NT_011520.10_242019	C	T	0.03
PCQAP	19185485	-937	rs738092	C	T	0.2
PCQAP	19186060	-362	NT_011520.10_252075	A	G	0.02
DKFZp434N035	19378097	-1858	NT_011520.10_444112	C	G	0.06
DKFZp434N035	19378214	-1741	NT_011520.10_444229	G	A	0.03
DKFZp434N035	19378224	-1731	rs5760087	C	T	0.45
DKFZp434N035	19378404	-1551	rs6003940	G	C	0.07
DKFZp434N035	19378496	-1459	NT_011520.10_444511	A	G	0.2

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
DKFZp434N035	19378548	-1407	NT_011520.10_444563	C	T	0.01
DKFZp434N035	19378884	-1071	rs2329472	T	C	0.05
DKFZp434N035	19378920	-1035	NT_011520.10_444935	G	C	0.01
DKFZp434N035	19379142	-813	rs2908694	T	C	0.44
SERPIND1	19456708	-1039	rs165912	C	T	0.43
SNAP29	19536098	-1747	NT_011520.10_602113	C	T	0.07
SNAP29	19536134	-1711	NT_011520.10_602149	C	T	0.01
PIK4CA	19537615	-104	NT_011520.10_603630	G	A	0.01
PIK4CA	19537740	-229	NT_011520.10_603755	G	C	0.02
CRKL	19595377	-790	rs7288034	C	G	0.37
CRKL	19595474	-693	NT_011520.10_661489	G	A	0.41
AIFL	19644875	-1197	rs5761567	T	C	0.47
AIFL	19645082	-990	rs6005061	T	C	0.47
AIFL	19645218	-854	rs17526584	G	A	0.41
AIFL	19645260	-812	rs17555251	T	C	0.32
LZTR1	19659015	-1827	rs8140475	T	G	0.01
LZTR1	19659427	-1415	rs13057408	G	C	0.42
LZTR1	19659813	-1029	rs178278	C	A	0.47
SLC7A4	19713159	-1757	rs2541956	C	T	0.34
UBE2L3	20245848	-651	rs140489	G	A	0.13
UBE2L3	20246240	-259	rs140490	G	T	0.09
FLJ36046	20310655	-903	NT_011520.10_1376670	G	T	0.01
FLJ36046	20310727	-831	NT_011520.10_1376742	G	T	0.03
FLJ36046	20310950	-608	rs12158334	T	A	0.01

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
FLJ36046	20311381	-177	rs861855	A	C	0.27
SDF2L1	20319585	-1543	rs861849	G	A	0.31
PPM1F	20631935	-174	rs875344	C	T	0.09
PPM1F	20632073	-312	rs412050	G	C	0.25
PPM1F	20633334	-1573	rs9610704	C	A	0.13
TOP3B	20655563	-670	rs17759988	C	T	0.11
TOP3B	20655841	-948	NT_011520.10_1721856	C	G	0.02
TOP3B	20655959	-1066	NT_011520.10_1721974	A	G	0.17
TOP3B	20656300	-1407	rs9607467	C	G	0.1
TOP3B	20656373	-1480	rs2877004	T	C	0.5
VPREB1	20921845	-1908	rs5750798	C	G	0.47
VPREB1	20921884	-1869	rs6001551	A	G	0.25
VPREB1	20921960	-1793	rs5750799	A	G	0.33
VPREB1	20922150	-1603	rs11574456	T	C	0.12
VPREB1	20922155	-1598	rs5757629	G	A	0.34
VPREB1	20923355	-398	rs6001558	G	A	0.04
VPREB1	20923436	-317	rs5757641	A	G	0.2
VPREB1	20923611	-142	rs5757643	G	A	0.36
SUHW1	21199209	-65	rs4822092	T	A	0.21
SUHW1	21199347	-203	rs362241	C	T	0.23
SUHW1	21199361	-217	rs9607985	G	C	0.3
SUHW1	21199519	-375	rs362208	G	A	0.26
SUHW1	21199606	-462	rs361755	C	G	0.38
SUHW1	21200091	-947	rs361660	G	T	0.11

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
SUHW1	21200344	-1200	rs361864	A	C	0.19
SUHW1	21200705	-1561	NT_011520.10_2266720	C	G	0.06
SUHW1	21200790	-1646	rs361535	C	T	0.25
SUHW1	21200809	-1665	rs361940	C	T	0.11
SUHW1	21200940	-1796	rs361828	G	T	0.13
SUHW1	21200984	-1840	rs9620107	G	A	0.17
GNAZ	21735472	-1088	rs13054904	T	A	0.24
GNAZ	21735606	-954	NT_011520.10_2801621	C	T	0.01
GNAZ	21736571	11	rs3788337	G	A	0.28
ZNF70	22418343	-514	NT_011520.10_3484358	C	A	0.15
ZNF70	22418769	-940	rs5759991	G	C	0.31
ZNF70	22419030	-1201	rs731545	T	G	0.37
SMARCB1	22453683	-15	rs11704810	G	T	0.14
DERL3	22506721	-942	NT_011520.10_3572736	G	A	0.01
SLC2A11	22522724	-1897	rs9754326	T	C	0.5
SLC2A11	22523868	-753	NT_011520.10_3589883	G	T	0.1
MIF	22559218	-1662	rs2012124	C	T	0.19
MIF	22559288	-1592	rs2012133	G	C	0.19
MIF	22559361	-1519	rs12483859	T	C	0.19
MIF	22559413	-1467	rs12485058	A	G	0.18
CABIN1	22731332	-982	rs422674	C	A	0.33
CABIN1	22732037	-277	rs11090305	T	C	0.23
CABIN1	22732121	-193	rs7289998	G	A	0.15
ADORA2A	23136931	-1331	rs3747115	C	T	0.39

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
UPB1	23187292	-114	rs738813	G	A	0.31
LOC388886	23314474	-877	rs2154611	C	T	0.25
LOC388886	23314767	-1170	rs4820599	A	G	0.36
LOC388886	23314902	-1305	rs5760488	T	A	0.32
LOC388886	23314912	-1315	rs2070476	A	G	0.35
LOC388886	23315200	-1603	rs5760489	A	G	0.3
RUTBC2	23524830	-1859	NT_011520.10_4590845	A	G	0.01
RUTBC2	23524980	-1709	NT_011520.10_4590995	C	T	0.03
RUTBC2	23525871	-818	rs175662	G	A	0.36
OTTHUMG00000030682	23668352	-1237	rs6004364	C	T	0.15
OTTHUMG00000030164	23671624	-1626	NT_011520.10_4737639	C	T	0.04
CRYBB3	23918554	-1824	rs6004479	A	G	0.48
CRYBB3	23918573	-1805	NT_011520.10_4984588	G	A	0.03
CRYBB3	23920110	-268	NT_011520.10_4986125	G	T	0.07
MYO18B	24461214	-1457	rs133849	T	C	0.37
MYO18B	24461881	-790	rs133851	G	C	0.4
ASPHD2	25152676	-1264	rs9608490	T	C	0.04
HPS4	25202535	-1869	rs3747134	A	G	0.12
HPS4	25202578	-1826	rs5761557	G	A	0.11
HPS4	25202837	-1567	NT_011520.10_6268852	A	G	0.06
HPS4	25203521	-883	rs5752333	C	T	0.11
HPS4	25203645	-759	rs9620611	G	A	0.18
HPS4	25203851	-553	rs6005059	C	G	0.1
HPS4	25204054	-350	NT_011520.10_6270069	G	C	0.11

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
HPS4	25204117	-287	rs3747136	T	C	0.04
HPS4	25204245	-159	rs3747137	T	C	0.08
HPS4	25204377	-27	rs968425	A	C	0.1
SRR1L	25205110	-756	rs4822724	C	T	0.48
SRR1L	25205685	-1331	rs13054869	C	T	0.02
SRR1L	25205700	-1346	rs5761560	C	T	0.47
CRYBB1	25339021	-429	rs5761635	C	T	0.42
CRYBA4	25341854	-627	rs2283843	G	T	0.39
CRYBA4	25341989	-492	rs5997109	C	G	0.37
PITPNB	26641672	-1633	rs9625361	G	T	0.09
PITPNB	26641771	-1732	NT_011520.10_7707786	A	T	0.02
PITPNB	26641955	-1916	rs12170161	C	A	0.14
MN1b	26711174	-1836	rs470100	A	G	0.43
MN1b	26711799	-1211	NT_011520.10_7777814	T	G	0.01
MN1b	26712303	-707	rs138642	G	A	0.08
MN1b	26712933	-77	rs138644	G	A	0.3
HSC20	27461711	-882	rs17883375	G	A	0.35
HSC20	27462498	-95	rs17436064	C	G	0.3
FLJ33814	27491783	-1477	NT_011520.10_8557798	G	T	0.01
FLJ33814	27492893	-367	NT_011520.10_8558908	G	A	0.02
FLJ33814	27493141	-119	NT_011520.10_8559156	G	C	0.02
HS747E2A	27782995	-536	rs134559	G	A	0.29
HS747E2A	27783342	-883	rs16987014	G	A	0.09
HS747E2A	27783399	-940	rs134560	G	A	0.17

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
KREMEN1	27793010	-651	rs134594	T	C	0.35
C22orf3	27989460	-948	rs2857460	C	T	0.22
GAS2L1	28026689	-840	NT_011520.10_9092704	C	T	0.03
GAS2L1	28026693	-836	rs174761	C	T	0.42
RRP22	28036603	-136	NT_011520.9_9102618	G	A	0.14
AP1B1	28110920	-1860	rs5763189	G	T	0.04
AP1B1	28111051	-1991	rs5763190	C	T	0.05
NEFH	28191010	-450	rs28380	C	G	0.42
NEFH	28191054	-406	NT_011520.10_9257069	G	A	0.01
NEFH	28191148	-312	NT_011520.10_9257163	T	G	0.22
NIPSNAP1	28302132	-254	rs13057041	C	T	0.4
NIPSNAP1	28302156	-278	rs12484392	C	A	0.42
NIPSNAP1	28302580	-702	rs5763346	A	T	0.33
NIPSNAP1	28303303	-1425	rs16987832	C	G	0.18
NIPSNAP1	28303312	-1434	rs5763347	G	C	0.09
NIPSNAP1	28303613	-1735	rs5763348	T	C	0.08
NF2	28322671	-1445	NT_011520.10_9388686	T	A	0.01
NF2	28323790	-326	rs1800538	G	C	0.4
ZMAT5	28487684	-95	NT_011520.10_9553699	G	T	0.01
ZMAT5	28487886	-297	rs17526577	A	G	0.4
ZMAT5	28488389	-800	rs140135	G	C	0.25
ZMAT5	28489170	-1581	rs140136	G	T	0.47
ASCC2	28559924	-1144	NT_011520.10_9625939	C	T	0.01
ASCC2	28560046	-1266	rs4820820	C	T	0.13

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
OTTHUMG00000030137	28964768	-1984	rs737921	G	A	0.32
OTTHUMG00000030137	28965149	-1603	rs929273	G	A	0.21
OTTHUMG00000030137	28965893	-859	NT_011520.10_10031908	C	T	0.05
LIF	28968163	-883	rs2267153	C	G	0.17
TBC1D10A	29010780	-648	rs4823085	C	T	0.29
TBC1D10A	29010808	-676	rs7284531	T	C	0.14
LOC550631	29076181	-1242	rs4820834	G	A	0.11
LOC550631	29076729	-694	NT_011520.10_10142744	T	G	0.05
LOC550631	29076895	-528	rs17657653	A	C	0.23
LOC550631	29076921	-502	rs5997619	C	T	0.1
SF3A1	29078539	-865	rs17657701	A	G	0.16
SF3A1	29078920	-1246	rs17730978	T	C	0.03
AB051443	29096409	-85	rs5753106	A	G	0.16
AB051443	29097514	-1190	NT_011520.10_10163529	T	C	0.2
AB051443	29098285	-1961	rs5997628	C	A	0.3
LOC200312	29109091	-1167	rs7284527	G	A	0.4
SEC14L2	29116381	-1105	rs1076271	G	A	0.14
SEC14L2	29116669	-817	rs715504	C	T	0.17
MTP18	29144312	-1864	rs1061664	G	A	0.13
MTP18	29145261	-915	rs5994305	A	G	0.2
GAL3st1	29285602	-170	rs42932	T	C	0.41
GAL3st1	29285738	-306	rs11704774	T	A	0.06
GAL3st1	29286106	-674	NT_011520.10_10352121	C	T	0.03
GAL3st1	29286199	-767	rs42933	C	T	0.13

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
GAL3st1	29286714	-1282	rs4820867	G	A	0.04
GAL3st1	29286760	-1328	rs42934	T	C	0.21
GAL3st1	29287296	-1864	rs42935	G	A	0.46
PES1	29312415	64	rs17526675	C	G	0.42
PES1	29312487	-8	rs17526668	T	A	0.1
PES1	29312514	-35	rs17526661	C	T	0.13
PES1	29312936	-457	rs17526654	G	A	0.36
PES1	29313072	-593	rs12484495	G	T	0.11
PES1	29313231	-752	NT_011520.10_10379246	G	C	0.01
PES1	29313316	-837	rs12484511	C	T	0.05
PES1	29314472	-1993	rs16988814	G	C	0.12
TCN2	29326376	-1260	rs5749131	G	A	0.39
SLC35E4	29354727	-1465	NT_011520.10_10420742	C	A	0.01
SLC35E4	29354767	-1425	rs5749148	C	T	0.45
SLC35E4	29355082	-1110	rs5753259	C	T	0.39
SLC35E4	29355146	-1046	rs5753260	G	A	0.43
OTTHUMG00000030444	29688253	-1943	NT_011520.10_10754268	G	A	0.01
PIB5PA	29841988	-1282	NT_011520.10_10908003	C	T	0.44
PIB5PA	29842616	-654	NT_011520.10_10908631	T	C	0.03
PLA2G3	29861234	-89	rs2232173	G	C	0.02
PLA2G3	29861338	-193	rs2232172	C	T	0.08
PLA2G3	29861676	-531	rs2232170	G	A	0.33
PLA2G3	29861896	-751	rs9619169	T	G	0.48
PLA2G3	29861920	-775	NT_011520.10_10927935	G	A	0.01

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
PLA2G3	29862959	-1814	NT_011520.10_10928974	G	T	0.01
LIMK2	29931503	-1301	NT_011520.10_10997518	C	G	0.05
LIMK2	29931738	-1066	rs9606827	A	G	0.06
LIMK2	29932388	-416	rs2073858	G	C	0.25
MGC17330	30014259	-869	rs5997948	T	A	0.14
OTTHUMG00000030356	30065621	-1698	rs714909	G	A	0.33
ZNF278	30067991	-1251	NT_011520.10_11134006	C	T	0.01
ZNF278	30068563	-1823	NT_011520.10_11134578	T	C	0.04
ZNF278	30068626	-1886	rs5997959	G	A	0.09
EIF4ENIF1	30209870	-582	rs12106594	C	T	0.06
DEPDC5	30474409	-199	NT_011520.10_11540424	G	C	0.01
YWHAH	30663739	-1285	rs9609391	A	G	0.11
YWHAH	30663767	-1257	rs929036	C	T	0.43
SLC5A4	30976117	-336	rs16990066	G	A	0.02
SLC5A4	30976548	-767	NT_011520.10_12042563	G	A	0.03
SLC5A4	30976563	-782	NT_011520.10_12042578	G	T	0.04
SLC5A4	30976583	-802	NT_011520.10_12042598	T	C	0.12
SLC5A4	30976958	-1177	rs12160790	C	T	0.03
SLC5A4	30976999	-1218	rs12157791	G	A	0.01
HSPC117	31132909	-115	rs17555307	G	A	0.42
HSPC117	31133380	-586	rs12167726	C	T	0.03
HSPC117	31133552	-758	rs734809	C	T	0.15
HSPC117	31134386	-1592	rs9609538	T	C	0.27
OTTHUMG00000058273	31185346	-361	rs9609562	C	T	0.29

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
OTTHUMG00000058273	31185578	-593	rs12628767	T	A	0.1
OTTHUMG00000058273	31185637	-652	rs9606957	G	T	0.42
OTTHUMG00000058273	31185772	-787	NT_011520.10_12251787	A	G	0.01
OTTHUMG00000058273	31185827	-842	NT_011520.10_12251842	C	T	0.17
OTTHUMG00000058273	31186740	-1755	NT_011520.10_12252755	T	A	0.01
OTTHUMG00000058273	31186936	-1951	rs738263	G	A	0.44
FBXO7	31194049	-1168	rs9609566	G	T	0.33
FBXO7	31195128	-89	rs3761435	A	G	0.09
TIMP3	31520938	-1302	rs5749511	C	T	0.1
TIMP3	31521247	-993	rs9619311	T	C	0.13
TIMP3	31521644	-596	rs2234920	T	A	0.01
SYN3	31779474	-566	NT_011520.10_12845489	G	A	0.11
SYN3	31780353	-1445	rs5749552	T	C	0.29
SYN3	31780644	-1736	rs5998713	C	A	0.05
LARGE	32642177	-1913	rs2267328	T	G	0.02
RAXLX	33784921	-1662	rs362198	G	A	0.32
RAXLX	33785475	-1108	rs361738	C	A	0.34
RAXLX	33785831	-752	rs361788	T	C	0.46
RAXLX	33785892	-691	rs362214	G	A	0.38
RAXLX	33785911	-672	rs361813	C	A	0.21
RAXLX	33785937	-646	rs362166	C	T	0.08
RAXLX	33786043	-540	rs361969	T	G	0.48
RAXLX	33786084	-499	rs361750	A	G	0.46
RAXLX	33786089	-494	rs361837	G	A	0.44

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
RAXLX	33786174	-409	rs361805	C	G	0.45
RAXLX	33786507	-76	rs5755566	C	A	0.24
HMG2L1	33977249	-783	rs2413323	T	C	0.43
HMG2L1	33977325	-707	rs11703542	G	A	0.1
HMG2L1	33978027	-5	rs5755674	T	C	0.43
TOM1	34018763	-1635	rs4509	A	G	0.42
TOM1	34018849	-1549	rs138726	G	A	0.42
TOM1	34019447	-951	rs138727	C	A	0.29
TOM1	34020356	-42	rs17526640	A	G	0.33
HMOX1	34100257	-650	NT_011520.10_15166272	T	G	0.34
MCM5	34119844	-825	rs4645726	G	C	0.25
MCM5	34119967	-702	rs1078979	A	G	0.36
RASD2	34259975	-1439	rs2092195	A	G	0.5
MB	34338762	-643	rs5750135	G	A	0.34
MB	34339399	-1280	NT_011520.10_15405414	C	T	0.1
APOL6	34368087	-827	rs5995133	G	C	0.1
APOL6	34368135	-779	rs4820207	G	A	0.03
MYH9	35108649	-115	NT_011520.10_16174664	G	A	0.01
MYH9	35109797	-1263	NT_011520.10_16175812	C	T	0.04
FLJ23322	35227546	-351	rs2277842	G	A	0.19
EIF3S7	35249997	-16	rs17555300	G	A	0.21
EIF3S7	35250238	-257	rs9607351	A	C	0.24
EIF3S7	35251614	-1633	NT_011520.10_16317629	G	A	0.24
EIF3S7	35251702	-1721	rs6519015	G	A	0.26

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
EIF3S7	35251725	-1744	rs8136248	A	G	0.25
EIF3S7	35251920	-1939	rs6000305	G	A	0.28
PVALB	35540258	-316	rs2001064	C	G	0.3
PVALB	35540982	-1040	rs2267362	C	G	0.32
PVALB	35541672	-1730	rs2213429	C	T	0.18
NCF4	35579121	-1963	rs9680702	G	A	0.2
NCF4	35579375	-1709	NT_011520.10_16645390	G	T	0.01
NCF4	35580042	-1042	rs9607387	T	C	0.48
NCF4	35580353	-731	rs4820258	T	C	0.22
NCF4	35580851	-233	rs9607388	G	T	0.15
NCF4	35580976	-108	rs10854694	G	A	0.4
CSF2RB	35640541	-2034	rs9607398	C	T	0.39
CSF2RB	35640625	-1950	rs5750338	A	G	0.5
CSF2RB	35640633	-1942	NT_011520.10_16706648	G	A	0.1
CSF2RB	35640888	-1687	rs4821567	T	G	0.46
CSF2RB	35640976	-1599	rs4821568	T	C	0.46
CSF2RB	35641373	-1202	rs4821569	A	G	0.45
CSF2RB	35641449	-1126	rs4820261	G	A	0.45
CSF2RB	35641609	-966	rs4820262	T	C	0.45
CSF2RB	35642348	-227	rs5756408	T	C	0.47
CSF2RB	35642414	-161	NT_011520.10_16708429	C	T	0.03
CSF2RB	35642514	-61	rs10222238	G	T	0.47
CSF2RB	35642533	-42	NT_011520.10_16708548	C	T	0.02
OTTHUMG00000030172	35728749	-379	rs5756471	G	A	0.41

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
OTTHUMG00000030172	35728909	-539	rs6000531	G	C	0.03
OTTHUMG00000030172	35729337	-967	rs6000532	G	A	0.44
OTTHUMG00000030172	35729710	-1340	rs130597	C	G	0.06
OTTHUMG00000030172	35729908	-1538	NT_011520.10_16795923	G	T	0.23
OTTHUMG00000030172	35729917	-1547	rs5756474	A	G	0.5
OTTHUMG00000030172	35729941	-1571	rs5756475	G	A	0.48
OTTHUMG00000030172	35730130	-1760	rs9610629	T	C	0.48
OTTHUMG00000030172	35730171	-1801	NT_011520.10_16796186	C	T	0.11
OTTHUMG00000030172	35730365	-1995	rs9610630	T	C	0.42
MPst	35743720	-965	rs4821585	G	T	0.4
MPst	35743775	-910	rs10427747	C	T	0.06
MPst	35743842	-843	rs10427778	T	C	0.06
MPst	35743977	-708	NT_011520.10_16809992	G	A	0.01
MPst	35744037	-648	NT_011520.10_16810052	T	C	0.05
MPst	35744082	-603	rs10427757	A	G	0.04
MPst	35744623	-62	rs11704682	C	G	0.16
KCTD17	35771546	-569	rs11913810	G	C	0.08
TMPRSS6	35830910	-699	rs228917	T	C	0.46
TMPRSS6	35831347	-1136	rs228920	T	C	0.45
TMPRSS6	35831362	-1151	rs2092169	C	T	0.14
TMPRSS6	35831376	-1165	rs228921	G	A	0.39
TMPRSS6	35831492	-1281	rs228922	C	T	0.41
TMPRSS6	35831519	-1308	rs1861947	G	A	0.16
C1QTNF6	35909186	-362	NT_011520.10_16975201	G	C	0.03

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
C1QTNF6	35909596	-772	rs6000600	T	C	0.07
C1QTNF6	35909738	-914	NT_011520.10_16975753	G	A	0.01
C1QTNF6	35910236	-1412	rs10427849	C	T	0.43
RAC2	35965910	-1104	rs5995406	T	C	0.32
RAC2	35965940	-1134	rs5995407	T	C	0.02
RAC2	35966124	-1318	rs7292284	A	G	0.32
PSCD4	36001167	-1756	rs5756586	C	G	0.27
PSCD4	36001892	-1031	rs2267363	T	G	0.34
PSCD4	36001904	-1019	rs2267364	C	T	0.34
PSCD4	36001990	-933	rs6000649	T	C	0.47
PSCD4	36002219	-704	rs727047	G	A	0.25
PSCD4	36002392	-531	rs727048	C	T	0.17
PSCD4	36002596	-327	rs11705401	G	A	0.29
PSCD4	36002704	-219	rs5756587	T	C	0.48
PSCD4	36002825	-98	rs3213554	T	C	0.41
PSCD4	36002878	-45	rs3213555	A	G	0.45
KIAA1904	36096485	-408	rs4821653	G	A	0.39
KIAA1904	36097369	-1292	rs4821654	T	G	0.25
MFNG	36207059	-112	rs2071839	T	C	0.09
MFNG	36207231	-284	rs11089844	C	T	0.09
MFNG	36208358	-1411	rs3761441	T	C	0.09
CARD10	36240605	-722	NT_011520.10_17306620	T	C	0.02
CARD10	36240613	-730	NT_011520.10_17306628	T	C	0.01
CARD10	36240617	-734	NT_011520.10_17306632	A	G	0.01

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
CDC42EP1	36279697	-1293	rs7284657	A	G	0.26
SH3BP1	36358351	-1822	NT_011520.10_17424366	C	G	0.03
SH3BP1	36358410	-1763	rs7289275	C	T	0.06
SH3BP1	36358650	-1523	rs13058685	A	G	0.05
PDXP	36378409	-824	NT_011520.10_17444424	C	T	0.4
PDXP	36378417	-816	NT_011520.10_17444432	G	A	0.01
PDXP	36378538	-695	NT_011520.10_17444553	A	G	0.32
PDXP	36378762	-471	rs9622677	A	C	0.38
PDXP	36378825	-408	rs7287340	C	T	0.37
MGC3731	36405089	-1754	rs7286269	C	T	0.26
MGC3731	36405119	-1724	rs9610831	G	A	0.39
MGC3731	36406071	-772	rs5756764	G	C	0.25
MGC3731	36406247	-596	rs5750472	A	C	0.27
H1F0	36523554	-2059	rs12160750	A	G	0.36
H1F0	36523864	-1749	rs6000897	G	A	0.11
H1F0	36524450	-1163	rs5756825	T	C	0.5
H1F0	36524624	-989	rs11703407	C	T	0.29
GCAT	36526899	-1538	rs6000898	T	C	0.34
GCAT	36527052	-1385	rs1894644	C	T	0.18
GCAT	36527142	-1295	rs1894645	T	G	0.11
GCAT	36527234	-1203	rs1894646	C	G	0.15
GALR3	36542435	-1453	rs5995502	T	C	0.44
GALR3	36543137	-751	rs7290156	T	C	0.42
GALR3	36543752	-136	NT_011520.10_17609767	T	C	0.02

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
OTTHUMG00000030664	36565932	-996	rs6000905	G	A	0.45
EIF3S6IP	36568490	-884	rs5756836	G	A	0.07
EIF3S6IP	36568716	-658	rs4821721	T	C	0.07
MICAL-L1	36624974	-1907	rs4346487	T	C	0.43
MICAL-L1	36625467	-1414	rs4821723	A	T	0.38
C22orf23	36675400	-1248	NT_011520.10_17741415	G	T	0.02
SOX10	36706804	-1725	rs12170378	G	A	0.06
PRKCABP	36776754	-945	rs742396	G	C	0.38
PRKCABP	36776796	-903	NT_011520.10_17842811	G	A	0.01
PRKCABP	36776865	-834	rs737662	C	G	0.28
PRKCABP	36776982	-717	rs17555334	G	A	0.29
PRKCABP	36777635	-64	rs11089858	G	A	0.01
SLC16A8	36804507	-860	NT_011520.10_17870522	C	T	0.13
BAIAP2L2	36831547	-402	rs5756916	G	A	0.38
PLA2G6	36903894	-1633	rs4821752	A	G	0.4
PLA2G6	36904069	-1808	rs4820321	T	A	0.33
MAFF	36920600	-1897	rs9607517	A	G	0.48
C22orf5	36994388	-894	rs11705672	T	A	0.28
CSNK1E	37039722	-1937	rs1997644	A	G	0.3
PGEA1	37377067	-77	rs9622836	T	C	0.01
GTPBP1	37425045	-1423	rs7291524	C	T	0.03
GTPBP1	37425146	-1322	NT_011520.10_18491161	A	G	0.02
GTPBP1	37425167	-1301	NT_011520.10_18491182	A	G	0.01
GTPBP1	37426119	-349	rs2267393	G	C	0.34

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
GTPBP1	37426133	-335	rs2267394	T	C	0.31
UNC84B	37476912	-512	rs7284966	A	G	0.38
DNAL4	37515324	-653	rs4821825	T	C	0.06
APOBEC3B	37817263	-455	rs113023	C	T	0.44
APOBEC3B	37817751	33	NT_011520.10_18883766	T	A	0.05
CBX7	37873673	-520	NT_011520.10_18939688	C	T	0.11
CBX7	37873833	-680	NT_011520.10_18939848	C	T	0.1
CBX7	37873926	-773	rs12158877	T	G	0.39
CBX7	37874026	-873	rs139405	C	A	0.38
MAP3K7IP1	38119728	-443	rs5750813	C	G	0.4
MAP3K7IP1	38120183	12	rs4821892	C	G	0.34
MGAT3	38176084	-1764	rs5757684	A	G	0.26
MGAT3	38176470	-1378	rs1557541	A	C	0.27
MGAT3	38176850	-998	rs1557542	C	G	0.17
FLJ20232	38221663	-1112	rs7287617	A	G	0.41
FLJ20232	38221913	-862	NT_011520.10_19287928	G	A	0.01
FLJ20232	38222292	-483	rs2294360	G	A	0.29
FLJ20232	38222349	-426	NT_011520.10_19288364	G	T	0.18
RPS19BP1	38254446	-1160	rs1109793	T	C	0.25
RPS19BP1	38254697	-1411	rs2413595	T	C	0.5
CACNA1I	38277133	-1177	rs5757726	G	C	0.14
FAM83F	38713843	-1703	rs9611241	A	G	0.28
FAM83F	38714352	-1194	NT_011520.10_19780367	C	T	0.06
FAM83F	38714359	-1187	NT_011520.10_19780374	C	T	0.09

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
FAM83F	38714738	-808	rs28607928	G	A	0.3
ADSL	39065844	-1150	rs12484610	T	C	0.22
MKL1	39357363	-151	rs4140512	A	G	0.15
MKL1	39358209	-997	rs6001989	G	A	0.38
MKL1	39358215	-1003	rs5758029	A	G	0.25
MKL1	39358221	-1009	rs10582736	G	A	0.15
GPR24	39399345	-291	NT_011520.10_20465360	T	A	0.01
GPR24	39399458	-178	NT_011520.10_20465473	T	C	0.01
EP300	39811718	-471	rs5995992	T	C	0.38
EP300	39811844	-345	rs4822002	A	G	0.37
RANGAP1	40006803	-76	NT_011520.10_21072818	T	A	0.04
RANGAP1	40007644	-917	rs11704524	G	C	0.13
RANGAP1	40008306	-1579	rs1969666	A	C	0.4
RANGAP1	40008503	-1776	rs6002312	A	T	0.34
RANGAP1	40008593	-1866	rs1535048	C	T	0.4
ZC3H7B	40020472	-1613	NT_011520.10_21086487	G	T	0.01
TEF	40100953	-1511	NT_011520.10_21166968	C	T	0.03
TEF	40101115	-1349	NT_011520.10_21167130	A	C	0.01
TEF	40101146	-1318	rs4822025	G	C	0.22
PHF5A	40189834	-619	NT_011520.10_21255849	G	A	0.04
PHF5A	40189892	-677	NT_011520.10_21255907	C	G	0.01
PHF5A	40190309	-1094	rs19573	A	C	0.19
POLR3H	40265372	-235	rs5758387	G	A	0.08
POLR3H	40265979	-842	rs9607813	G	A	0.04

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
CSDC2	40281444	-59	NT_011520.10_21347459	G	A	0.11
XRCC6	40340265	-1493	rs5751129	T	C	0.22
XRCC6	40341026	-732	rs28384701	C	T	0.02
OTTHUMG00000030205	40409136	-1535	NT_011520.10_21475151	G	C	0.01
MEI1	40419640	-362	rs2003816	G	T	0.23
MEI1	40419964	-38	rs743832	C	T	0.04
FLJ22349	40519910	-1201	rs139561	T	C	0.21
FLJ22349	40520967	-144	rs738248	G	A	0.32
FLJ22349	40521184	73	rs139562	G	C	0.18
SREBF2	40551650	-1952	NT_011520.10_21617665	A	G	0.08
SREBF2	40552656	-946	NT_011520.10_21618671	T	C	0.08
SREBF2	40552891	-711	NT_011520.10_21618906	A	C	0.01
SREBF2	40553592	-10	NT_011520.10_21619607	G	C	0.12
TNFRSF13C	40648413	-1133	rs5996088	C	T	0.07
TNFRSF13C	40648625	-1345	NT_011520.10_21714640	G	T	0.01
TNFRSF13C	40648963	-1683	NT_011520.10_21714978	G	A	0.09
TNFRSF13C	40649021	-1741	NT_011520.10_21715036	G	C	0.09
TNFRSF13C	40649122	-1842	rs12158335	T	G	0.25
TNFRSF13C	40649200	-1920	NT_011520.10_21715215	T	C	0.09
TNFRSF13C	40649225	-1945	NT_011520.10_21715240	G	A	0.03
OTTHUMG00000030501	40666434	-1853	NT_011520.10_21732449	A	G	0.07
OTTHUMG00000030501	40666577	-1710	NT_011520.10_21732592	C	T	0.01
OTTHUMG00000030501	40667265	-1022	rs3752592	G	T	0.12
OTTHUMG00000030501	40667370	-917	NT_011520.10_21733385	C	G	0.1

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
OTTHUMG00000030501	40667591	-696	rs6002555	G	A	0.26
C22orf18	40668667	-896	NT_011520.10_21734682	T	C	0.01
C22orf18	40668797	-1026	rs8140869	G	A	0.26
C22orf18	40668908	-1137	rs5996092	G	A	0.47
SEPT3	40695491	-1946	rs5751191	T	C	0.39
SEPT3	40695833	-1604	rs4820451	C	T	0.25
MGC26816	40717311	-1959	rs1062753	G	A	0.32
NAGA	40791405	-61	rs133376	C	T	0.37
NAGA	40792195	-851	rs6519305	C	G	0.35
NAGA	40793035	-1691	rs133379	A	G	0.46
FAM109B	40794563	-198	NT_011520.10_21860578	C	T	0.04
FAM109B	40794817	56	rs13057094	C	T	0.37
OTTHUMG00000030175	40798638	-1556	rs1807494	C	G	0.33
OTTHUMG00000030175	40800068	-126	rs8135801	A	G	0.28
OTTHUMG00000030175	40800203	9	rs2269524	T	G	0.29
AK123630	40809792	-1725	rs4147640	T	C	0.35
AK123630	40809855	-1662	NT_011520.10_21875870	C	T	0.01
AK123630	40810171	-1346	rs2284087	C	T	0.37
AK123630	40811223	-294	rs1801311	G	A	0.36
AK123630	40811261	-256	NT_011520.10_21877276	T	G	0.01
NDUFA6	40812400	-939	rs4147638	G	A	0.35
TCF20	40936906	-962	rs5758652	T	C	0.18
CGI-96	41240713	-116	rs5758781	C	G	0.46
CGI-96	41241652	-1055	rs7287384	C	T	0.26

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
SERHL	41273947	-450	NT_011520.10_22339962	G	A	0.2
POLDIP3	41335744	-281	rs137114	C	T	0.09
POLDIP3	41335901	-438	rs137115	C	T	0.16
CYB5R3	41371015	-1116	rs11705269	G	A	0.35
CYB5R3	41371205	-1306	rs6002862	C	T	0.46
A4GALT	41417284	-1149	rs3761462	C	T	0.42
A4GALT	41417349	-1214	rs130396	A	C	0.22
A4GALT	41417641	-1506	rs130397	T	C	0.17
A4GALT	41417687	-1552	NT_011520.10_22483702	T	A	0.06
A4GALT	41417794	-1659	rs3761465	G	T	0.06
A4GALT	41417955	-1820	rs916231	G	A	0.11
A4GALT	41418094	-1959	rs135108	G	A	0.08
PACSIN2	41736385	-709	rs5759095	A	C	0.35
BIK	41829086	-2019	rs2013863	T	C	0.34
BIK	41829569	-1536	rs4988360	C	T	0.25
BIK	41829641	-1464	NT_011520.10_22895656	A	G	0.04
BIK	41829655	-1450	rs4988362	A	G	0.37
BIK	41829942	-1163	rs11574525	C	A	0.07
MT	41864149	-249	rs5759182	A	G	0.11
MT	41865671	-1771	rs926329	C	T	0.48
TTLL12	41908653	-1050	NT_011520.10_22974668	A	C	0.01
TTLL12	41908727	-1124	rs138957	T	G	0.31
TTLL12	41908745	-1142	NT_011520.10_22974760	A	G	0.01
TTLL12	41908779	-1176	rs138958	T	G	0.32

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
TTLL12	41908910	-1307	NT_011520.10_22974925	A	G	0.04
TTLL12	41909129	-1526	rs138959	A	G	0.28
TTLL12	41909209	-1606	rs7286832	C	T	0.06
TTLL12	41909473	-1870	NT_011520.10_22975488	C	T	0.03
SCUBE1	42064564	-741	rs2859446	C	T	0.13
SCUBE1	42064589	-766	NT_011520.10_23130604	G	C	0.03
SCUBE1	42064742	-919	rs2744880	T	C	0.47
SCUBE1	42065635	-1812	NT_011520.10_23131650	C	T	0.04
SCUBE1	42065684	-1861	NT_011520.10_23131699	C	T	0.01
MPPED1	42130698	-1294	NT_011520.10_23196713	C	G	0.06
MPPED1	42131452	-540	rs5759322	G	T	0.2
FLJ2358	42534189	-1073	rs5764317	T	G	0.16
SULT4A1	42585073	-1839	rs138111	G	A	0.05
PNPLA5	42613207	-418	rs11913819	G	C	0.05
PNPLA5	42614724	-1935	NT_011521.3_393437	G	A	0.1
ADPN	42643599	-920	NT_011521.3_422312	A	G	0.01
SAMM50	42675318	-862	rs1474746	G	C	0.37
SAMM50	42676159	-21	NT_011521.3_454872	A	C	0.01
PARVG	42900039	-1718	rs139122	G	C	0.47
PARVG	42900479	-1278	rs878405	G	C	0.32
PARVG	42900769	-988	rs3747208	A	G	0.14
PARVG	42900986	-771	rs3747209	G	A	0.41
PARVG	42901033	-724	rs139124	G	A	0.35
PARVG	42901114	-643	rs139125	G	C	0.48

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
PARVG	42901521	-236	NT_011521.3_680234	T	A	0.01
PARVG	42901556	-201	rs7287117	G	A	0.02
BC104183	43019513	-1950	rs8140742	A	T	0.48
LDOC1L	43215940	-1729	NT_011522.5_161984	G	A	0.03
LDOC1L	43216104	-1893	rs131167	G	A	0.43
ARHGAP8	43416759	-1911	rs5765914	A	G	0.33
ARHGAP8	43417386	-1284	rs5765915	G	C	0.25
PHF21B	43726928	-812	rs1997890	G	A	0.43
PHF21B	43726976	-860	rs12169401	A	G	0.05
PHF21B	43727037	-921	NT_011522.5_673081	G	C	0.01
PHF21B	43727064	-948	rs4823435	C	T	0.3
PHF21B	43727070	-954	rs140552	C	T	0.13
PHF21B	43727655	-1539	rs131989	C	T	0.11
NUP50	43879730	-591	rs132846	G	A	0.32
NUP50	43880168	-153	rs132847	G	C	0.31
NUP50	43880278	-43	rs3788634	G	T	0.15
NUP50	43880308	-13	rs132848	A	C	0.3
C22orf9	43958169	-960	rs6007507	C	G	0.44
C22orf9	43958702	-1493	rs6007508	C	T	0.46
C22orf9	43958798	-1589	NT_011522.5_904842	G	A	0.48
C22orf9	43958803	-1594	NT_011522.5_904847	A	G	0.48
C22orf9	43958883	-1674	rs4823286	G	A	0.47
C22orf9	43959072	-1863	rs5766584	T	G	0.43
UPK3A	44001198	-201	rs2742631	C	G	0.42

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
C22orf8	44025165	-1186	rs6007582	G	A	0.07
C22orf8	44025778	-573	rs226503	T	G	0.13
C22orf8	44025920	-431	rs226504	A	T	0.33
RIBC2	44128989	-1131	NT_011522.5_1075033	T	A	0.48
RIBC2	44129732	-388	NT_011522.5_1075776	G	A	0.06
RIBC2	44129999	-121	NT_011522.5_1076043	A	G	0.01
SMC1L2	44130161	-126	rs2272804	C	A	0.37
SMC1L2	44130235	-200	rs2272805	G	A	0.15
ATXN10	44386503	-1712	rs134858	A	G	0.22
ATXN10	44386580	-1635	rs134859	G	C	0.14
FLJ27365	44812355	-1969	NT_011523.10_104548	C	T	0.18
FLJ27365	44812565	-1759	rs3747242	G	C	0.39
FLJ27365	44812956	-1368	rs17576497	G	A	0.17
FLJ27365	44813080	-1244	NT_011523.10_105273	C	A	0.01
FLJ27365	44813105	-1219	rs8136389	A	G	0.34
FLJ27365	44813249	-1075	rs8136639	A	G	0.28
FLJ27365	44813989	-335	rs9615411	T	C	0.27
OTTHUMG00000030672	44967128	-421	rs6008320	T	C	0.22
OTTHUMG00000030672	44967507	-800	rs9627287	A	G	0.02
OTTHUMG00000030672	44967948	-1241	rs8142080	G	T	0.13
OTTHUMG00000030672	44968039	-1332	NT_011523.10_260232	T	C	0.01
OTTHUMG00000030672	44968363	-1656	NT_011523.10_260556	G	A	0.09
OTTHUMG00000030672	44968428	-1721	rs3087501	A	G	0.11
PKDREJ	44980461	-725	rs12167567	G	C	0.14

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
PKDREJ	44980533	-797	NT_011523.10_272726	T	C	0.14
PKDREJ	44980689	-953	NT_011523.10_272882	G	A	0.11
FLJ20699	44982505	-1874	rs7410393	G	A	0.2
FLJ20699	44983536	-843	NT_011523.10_275729	A	G	0.14
OTTHUMG00000030329	45014273	-1151	NT_011523.10_306466	C	T	0.04
TRMU	45050538	-1591	rs9615952	A	T	0.15
TRMU	45050672	-1457	rs7287689	A	G	0.24
TRMU	45051005	-1124	rs6008749	T	A	0.18
TRMU	45051543	-586	rs9615953	C	G	0.16
CELSR1	45255120	-1536	rs1009156	A	G	0.1
CELSR1	45255223	-1639	rs3788728	G	C	0.32
CELSR1	45255332	-1748	rs1883189	G	A	0.33
OTTHUMG00000030404	45328827	-1594	rs138507	C	T	0.06
OTTHUMG00000030404	45328951	-1470	rs138506	C	T	0.2
OTTHUMG00000030404	45329257	-1164	rs138503	T	C	0.15
OTTHUMG00000030404	45329274	-1147	rs138502	T	C	0.17
OTTHUMG00000030404	45330390	-31	rs9917583	C	T	0.2
CERK	45455246	-713	rs7349028	C	T	0.18
CERK	45455373	-840	NT_011523.10_747566	G	C	0.01
CERK	45455724	-1191	rs5769126	A	C	0.46
CERK	45455791	-1258	rs5769127	G	A	0.46
CERK	45456482	-1949	rs801581	C	T	0.32
CERK	45456518	-1985	rs4823874	A	G	0.42
CERK	45456626	-2093	NT_011523.10_748819	C	T	0.29

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
TBC1D22A	45477222	-1832	rs5769136	T	C	0.39
TBC1D22A	45477658	-1396	rs801643	C	T	0.45
TBC1D22A	45477740	-1314	rs5769138	G	C	0.39
TBC1D22A	45478131	-923	rs5769139	C	A	0.41
TBC1D22A	45478698	-356	rs12389	G	A	0.4
TBC1D22A	45478854	-200	rs11703936	C	A	0.06
TBC1D22A	45478963	-91	rs2295441	T	C	0.41
TBC1D22A	45478995	-59	rs801641	C	G	0.04
AK057318	45632453	-20	rs9616151	C	T	0.1
AK057318	45632617	-184	rs9616152	C	T	0.1
AK057318	45632637	-204	NT_011523.10_924830	C	G	0.01
AK057318	45632836	-403	rs9616153	C	T	0.1
AK057318	45633075	-642	NT_011523.10_925268	G	A	0.03
AK057318	45633601	-1168	rs5769244	A	G	0.32
AK057318	45633834	-1401	NT_011523.10_926027	G	A	0.01
AK057318	45633869	-1436	rs5767412	C	T	0.24
AK057318	45634176	-1743	NT_011523.10_926369	C	T	0.01
AK057318	45634407	-1974	NT_011523.10_926600	A	G	0.01
OTTHUMG00000030676	46856280	-1039	rs9637353	C	G	0.4
OTTHUMG00000030676	46856287	-1032	rs2285091	A	G	0.47
OTTHUMG00000030676	46856513	-806	NT_011523.10_2148706	C	T	0.02
OTTHUMG00000030676	46856772	-547	rs135610	C	T	0.05
OTTHUMG00000030676	46856993	-326	rs131132	G	A	0.06
OTTHUMG00000030676	46857192	-127	rs1018793	C	T	0.42

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
OTTHUMG00000030528	47263876	-160	rs6008753	T	C	0.01
OTTHUMG00000030528	47265529	-1813	rs736786	T	C	0.05
OTTHUMG00000030528	47265580	-1864	rs133470	C	A	0.42
FAM19A5	47361959	-1742	rs874087	A	G	0.04
FAM19A5	47362137	-1564	rs9617404	C	T	0.35
FAM19A5	47362321	-1380	rs9617471	C	A	0.34
FAM19A5	47362447	-1254	NT_011525.6_52727	C	A	0.02
FAM19A5	47362666	-1035	rs4925418	G	A	0.18
FAM19A5	47363051	-650	rs9617405	G	A	0.29
FAM19A5	47363409	-292	NT_011525.6_53689	C	T	0.04
FAM19A5	47363493	-208	rs9617406	G	T	0.28
FAM19A5	47363587	-114	rs9617472	C	T	0.3
FAM19A5	47363609	-92	rs9617407	T	C	0.35
BRD1	48540813	-406	NT_011525.6_1231093	C	T	0.13
BRD1	48540887	-480	NT_011525.6_1231167	G	A	0.03
BRD1	48541343	-936	rs138881	G	A	0.09
ZBED4	48567556	-833	rs8139718	G	A	0.02
MOV10L1	48830197	-753	NT_019197.4_102731	A	G	0.37
MOV10L1	48830275	-675	rs4838820	G	A	0.04
PP2447	48924861	-1946	NT_019197.4_197395	G	C	0.14
PP2447	48926539	-268	NT_019197.4_199073	C	G	0.23
SELO	48940379	-1815	rs6712	G	C	0.14
SELO	48940543	-1651	rs2272845	C	G	0.2
SELO	48940567	-1627	NT_019197.4_213101	G	A	0.02

Gene Name	SNP Position (build35)	SNP Position (relative to TSS)	SNP Name	Major Allele	Minor Allele	Minor Allele Frequency
SELO	48941516	-678	rs28668373	G	T	0.33
TUBGCP6	48986753	-877	rs13058062	C	T	0.31
TUBGCP6	48986769	-893	rs5771271	G	A	0.15
TUBGCP6	48986872	-996	rs11553697	T	C	0.01
TUBGCP6	48987165	-1289	rs2294404	G	A	0.12
TUBGCP6	48987204	-1328	NT_019197.4_259738	C	T	0.01
HDAC10	48992779	-490	rs2341111	C	G	0.4
HDAC10	48992915	-626	NT_019197.4_265449	C	A	0.02
HDAC10	48993922	-1633	NT_019197.4_266456	G	A	0.01
SBF1	49205305	-1706	NT_011526.6_135224	C	T	0.01
SBF1	49205429	-1830	rs5770843	C	T	0.16
SBF1	49205466	-1867	NT_011526.6_135385	A	G	0.07
ADM2	49208550	-1578	rs9616854	A	G	0.44
hCAP-H2	49235400	-1376	rs131824	A	G	0.32
hCAP-H2	49235943	-833	NT_011526.6_165862	C	G	0.04
hCAP-H2	49236077	-699	NT_011526.6_165996	G	T	0.01
TMEM153	49237340	-1067	NT_011526.6_167259	T	C	0.01
KLHDC7B	49274852	-1754	NT_011526.6_204771	C	G	0.05
OTTHUMG00000030815	49309855	-1726	NT_011526.6_239774	G	C	0.01
OTTHUMG00000030815	49311402	-179	NT_011526.6_241321	C	A	0.01

Appendix C - SNPs and indels in cloned promoter fragments

The genes are listed in order of occurrence on chromosome 22, from the centromeric to the telomeric end of the q arm.

Gene Name	SNP Position (build35)	SNP Position (Relative to TSS)	dbSNP id	Alleles
XKR3	15677594	-458	rs5992556	G/A
XKR3	15677526	-390	rs5994030	G/T
XKR3	15677440	-304	rs5994029	G/A
XKR3	15677390	-254	rs5992555	G/A
XKR3	15677377	-241	rs12484826	G/A
XKR3	15677357	-221	rs5016128	C/T
XKR3	15677328	-192	rs5016127	G/C
XKR3	15677317	-181	rs12484164	G/C
XKR3	15677239	-103	rs2891848	A/T
XKR3	15677067	69		AA/--
SLC25A18	16417644	-64	rs17555265	C/G
BCL2L13	16485756	-479		G/A
BCL2L13	16485985	-250	rs17526598	C/T
BCL2L13	16486110	-125	rs5746448	C/A
PEX26	16934759	-483		G/T
PEX26	16935124	-118	rs462055	C/T
PEX26	16935151	-91		C/T
PEX26	16935165	-77	rs12157958	G/T
DGCR2	17484963	-467	rs17526612	C/T
DGCR2	17484509	-13	rs17526619	C/T
TSSK2	17492088	-305		C/T
TSSK2	17492200	-193		G/A
TSSK2	17492446	53	rs12233351	C/T
DGCR14	17507135	-408		[A]n
DGCR14	17506949	-222		C/T
DGCR14	17506939	-212	rs1936951	C/T
DGCR14	17506934	-207	rs1936950	T/A
DGCR14	17506933	-206		C/A
DGCR14	17506879	-152	rs737923	C/T
UFD1L	17841440	-70	rs5992403	G/A
CDC45L	17841783	-192	rs13447180	G/T
CDC45L	17841851	-124	rs4141528	C/G
CDC45L	17842057	82		C/A
CLDN5	17887472	-32	rs5748258	G/A
TBX1	18118583	-196		C/T
GNB1L	18217302	-288	rs28451568	C/T
COMT	18303362	-487		T/-
COMT	18303438	-411	rs2020917	C/T
COMT	18303581	-268	rs13306278	C/T
RANBP1	18478960	-513		C/T
RANBP1	18479333	-140		C/T
RANBP1	18479373	-100	rs713982	C/A
RANBP1	18479407	-66	rs2286929	G/T
AK057137	18568149	-324		G/A
AK057137	18568415	-58	rs654389	G/C

Gene Name	SNP Position (build35)	SNP Position (Relative to TSS)	dbSNP id	Alleles
ZNF74	19072607	-406	rs17551339	G/A
ZNF74	19072807	-206	rs17551346	C/T
PCQAP	19186060	-362		G/A
PIK4CA	19537970	-459	rs17858053	C/T
PIK4CA	19537934	-423	rs1061063	G/A
PIK4CA	19537740	-229		C/G
PIK4CA	19537615	-104		C/T
PIK4CA	19537526	-15		AGGCGG/------
UBE2L3	20245975	-524		G/T
UBE2L3	20245981	-518		C/T
UBE2L3	20246020	-479	rs9623962	T/-
PPM1F	20632073	-312	rs412050	G/C
PPM1F	20631935	-174	rs875344	G/A
VPREB1	20923278	-475	rs5757639	G/T
VPREB1	20923297	-456	rs5757640	C/T
VPREB1	20923355	-398	rs6001558	G/A
VPREB1	20923436	-317	rs5757641	G/A
VPREB1	20923611	-142	rs5757643	G/A
VPREB1	20923632	-121		G/A
VPREB1	20923683	-70	rs11574461	G/A
VPREB1	20923848	95		C/A
SUHW1	21199660	-516		C/T
SUHW1	21199606	-462	rs361755	C/G
SUHW1	21199588	-444	rs9607987	G/A
SUHW1	21199565	-421		TTGAGA/------
SUHW1	21199519	-375	rs362208	C/T
SUHW1	21199397	-253	rs9607986	C/G
SUHW1	21199361	-217	rs9607985	C/G
SUHW1	21199347	-203	rs362241	G/A
SUHW1	21199209	-65	rs4822092	A/T
SUHW1	21199194	-50		G/A
SUHW1	21199092	52	rs361986	C/G
SMARCB1	22453559	-139	rs2073387	G/T
SMARCB1	22453683	-15	rs11704810	G/T
SMARCB1	22453768	70		C/A
CABIN1	22731888	-426	rs9624386	G/C
CABIN1	22732037	-277	rs11090305	C/T
CABIN1	22732077	-237		[CA]n
CABIN1	22732120	-194		Complex*
CRYBB3	23920110	-268		G/T
HPS4	25204466	62		T/A
SRR1L	25204500	-146	rs13340064	TCTCCCCCGGGGCGCCGCCTC /----------
SRR1L	25204377	-23	rs968425	G/T
MN1b	26712933	-77	rs138644	G/A
MN1b	26712976	-34	rs13057353	C/A
FLJ33814	27493141	-119		G/C
RRP22	28036603	-136		C/T
AP1B1	28109340	-280	rs5752906	C/G
NEFH	28191054	-406		G/A
NEFH	28191316	-144		C/T

NIPSNAP1	28302156	-278		T/G
Gene Name	**SNP Position (build35)**	**SNP Position (Relative to TSS)**	**dbSNP id**	**Alleles**
NIPSNAP1	28302132	-254		A/G
ZMAT5	28487886	-297	rs17526577	C/T
ZMAT5	28487684	-95		C/A
HORMAD2	28800825	123	rs6519802	G/T
LIMK2	29932388	-416	rs2073858	G/C
DEPDC5	30474409	-199		G/C
HSPC117	31133091	-297		T/-
HSPC117	31132909	-115	rs17555307	C/T
OTTHUMG00000058273	31185346	-361	rs9609562	G/A
OTTHUMG00000058273	31185189	-204		[GT]n
FBXO7	31194867	-350		C/-
FBXO7	31194868	-349		[G]n
FBXO7	31195128	-89	rs3761435	G/A
HMG2L1	33978027	-5	rs5755674	C/T
TOM1	34020096	-302		C/T
TOM1	34020231	-167	rs17526626	C/-
TOM1	34020356	-42	rs17526640	G/A
MYH9	35108649	-115		C/T
NCF4	35580762	-322	rs11089806	G/A
NCF4	35580851	-233	rs9607388	G/T
NCF4	35580976	-108	rs10854694	G/A
CSF2RB	35642348	-227	rs5756408	C/T
CSF2RB	35642414	-161		C/T
CSF2RB	35642470	-105	rs10222232	G/A
CSF2RB	35642514	-61	rs10222238	G/T
CSF2RB	35642533	-42		C/T
CSF2RB	35642534	-41		G/A
CSF2RB	35642573	-2		C/T
CSF2RB	35642621	46		G/T
OTTHUMG00000030172	35728749	-379	rs5756471	C/T
OTTHUMG00000030172	35728667	-297	rs16997638	G/T
MPst	35744623	-62	rs11704682	G/C
PSCD4	36002449	-474	rs8141057	C/A
PSCD4	36002596	-327	rs11705401	G/A
PSCD4	36002704	-219	rs5756587	C/T
PSCD4	36002824	-99		CGTTTGTT/--------
PSCD4	36002825	-98		[GTTT]n
PSCD4	36002878	-45	rs3213555	G/A
KIAA1904	36096485	-408	rs4821653	C/T
MFNG	36207231	-284	rs11089844	G/A
MFNG	36207059	-112	rs2071839	G/A
PDXP	36378883	-350		C/T
GALR3	36543572	-316		C/-
GALR3	36543752	-136		C/T
PRKCABP	36777635	-64	rs11089858	G/A
C22orf5	36993495	-1		G/A
PGEA1	37376620	-524		C/T
GTPBP1	37426119	-349	rs2267393	C/G
GTPBP1	37426133	-335	rs2267394	C/T
APOBEC3B	37817263	-455	rs113023	C/T

Gene Name	SNP Position (build35)	SNP Position (Relative to TSS)	dbSNP id	Alleles
APOBEC3B	37817352	-366		C/T
APOBEC3B	37817399	-319		C/T
APOBEC3B	37817434	-284		G/A
APOBEC3B	37817748	30		T/C
APOBEC3B	37817751	33		A/T
APOBEC3B	37817794	76		C/T
FLJ20232	38222292	-483	rs2294360	A/G
FLJ20232	38222349	-426		G/T
FLJ20232	38222546	-229		G/A
PHF5A	40189740	-525		C/T
PHF5A	40189357	-142		G/A
OTTHUMG00000030205	40410353	-318		C/A
OTTHUMG00000030205	40410483	-188		[A]n
MEI1	40419563	-439		G/T
MEI1	40419640	-362	rs2003816	G/T
MEI1	40419916	-86		G/A
MEI1	40419964	-38	rs743832	C/T
MEI1	40420043	41	rs6003024	GA/--
FLJ22349	40520811	-300		C/T
FLJ22349	40520859	-252		G/A
FLJ22349	40520967	-144	rs738248	G/A
FLJ22349	40521184	73	rs139562	C/G
SREBF2	40553592	-10		G/C
MGC26816	40719112	-158	rs4822079	C/G
NAGA	40791879	-535		G/A
NAGA	40791480	-136	rs2859438	A/T
NAGA	40791450	-106	rs133377	G/A
NAGA	40791405	-61	rs133376	G/A
OTTHUMG00000030175	40799699	-495		C/T
OTTHUMG00000030175	40799715	-479		C/T
OTTHUMG00000030175	40800068	-126	rs8135801	G/A
OTTHUMG00000030175	40800203	9	rs2269524	G/T
CGI-96	41240889	-292	rs5751295	C/T
CGI-96	41240713	-116	rs5758781	G/C
SERHL	41273912	-485		C/T
SERHL	41273947	-450		G/A
SERHL	41274041	-356		G/C
POLDIP3	41335901	-438	rs137115	G/A
POLDIP3	41335744	-281	rs137114	G/A
MT	41864247	-347		C/A
MT	41864149	-249	rs5759182	C/T
MT	41863929	-29	rs13056026	T/A
MPPED1	42131622	-370		T/C
PNPLA5	42613207	-418	rs11913819	C/G
SAMM50	42675665	-515		C/A
SAMM50	42676159	-21		C/A
PARVG	42901521	-236		T/A
PARVG	42901556	-201	rs7287117	G/A
NUP50	43879802	-519		G/A
NUP50	43879807	-514		C/A
NUP50	43880168	-153	rs132847	G/C

NUP50	43880278	-43	rs3788634	G/T
Gene Name	**SNP Position (build35)**	**SNP Position (Relative to TSS)**	**dbSNP id**	**Alleles**
NUP50	43880308	-13	rs132848	C/A
UPK3A	44001021	-378		G/T
UPK3A	44001198	-201	rs2742631	G/A
C22orf8	44025920	-431	rs226504	A/T
C22orf8	44026241	-110		GGGCG/-----
RIBC2	44129732	-388		G/A
RIBC2	44129999	-121		G/A
RIBC2	44130161	41	rs2272804	C/A
SMC1L2	44130393	-358		C/T
SMC1L2	44130303	-268		G/A
SMC1L2	44130284	-249		G/A
SMC1L2	44130235	-200	rs2272805	C/T
SMC1L2	44130161	-126	rs2272804	G/T
SMC1L2	44129995	40		G/A
FLJ27365	44813989	-335	rs9615411	C/T
FLJ27365	44814307	-17		C/T
FLJ27365	44814371	47	rs3747243	C/T
OTTHUMG00000030672	44967128	-421	rs6008320	G/A
OTTHUMG00000030672	44966694	13		G/A
PKDREJ	44980195	-459		CAAA/----
PKDREJ	44979778	-42		G/A
PKDREJ	44979705	31		G/T
TBC1D22A	45478698	-356	rs12389	G/A
TBC1D22A	45478854	-200	rs11703936	C/A
TBC1D22A	45478963	-91	rs2295441	C/T
TBC1D22A	45478995	-59	rs801641	G/C
AK057318	45632836	-403	rs9616153	G/A
AK057318	45632637	-204		G/C
AK057318	45632617	-184	rs9616152	G/A
AK057318	45632453	-20	rs9616151	G/A

***** Sequence of the hypervariable region for each of the 7 haplotypes. It must be noted that the low complexity of the sequence means that it may be influenced by sequencing perturbations. The real sequence may not be identical to that elucidated experimentally

c1 CACACACACACACACACACACACACACACACACACACACACACACACACGCGCACGCGCACGCGCACGCGCGCCG

c2 ----------GCACACGCACACGCACACGCACACGCACGCGCGCGCGCGCCG

c3 CACACACACACACGCACACGCACACGCGCACGCGCGCACACACACGCGCACACACACGCGCGCAC--GCGCGCGCGCCG

c4 CGCACACGCACACGCACACGCACACGCGCACGCGCGCACACACGCGCACACACACGCGCGCACGCGCGCGCGCGCCG

c5 CGCACACGCACACGCACACGCACACGCGCACGCGCGCACACACGCGCACACACACGCGCGCAC--GCGCGCGCGCCGC

c6 ----------CACACACACACACACACACACACACGCGCACGCGCACGCGCACGCGCGCCG

c7 ----------CACACGCACACACACACGCACGCGCGCGCGCGCCG

Appendix D - Haplotypes cloned into Gateway-modified pGL3 Basic

The names of the genes are listed in order of occurrence on chromosome 22, from the centromeric to the telomeric end of the q arm. Bold haplotypes are those which are only differentiated from another haplotype by the presence of unconfirmed SNPs. Minor alleles of unconfirmed SNPs are marked with an *. Minor alleles of unconfirmed SNPs that match a dbSNP entry with the same alleles are marked with a §.

Gene Name	SNP Position (build35)	SNP Position (Relative to TSS)	Alleles	Haplotype									
				1	2	3	4	5	6	7	8	9	10
XKR3	15677067	69	AA/--	AA	AA	--	AA						
XKR3	15677239	-103	A/T	A	T	A	A						
XKR3	15677317	-181	G/C	G	C	C	C						
XKR3	15677328	-192	G/C	C	G	C	G						
XKR3	15677357	-221	C/T	C	T	C	T						
XKR3	15677377	-241	G/A	G	A	A	A						
XKR3	15677390	-254	G/A	G	A	G	A						
XKR3	15677440	-304	G/A	G	A	G	A						
XKR3	15677526	-390	G/T	G	T	G	T						
XKR3	15677594	-458	G/A	A	G	A	G						
SLC25A18	16417644	-64	C/G	G	C								
BCL2L13	16485756	-479	G/A	G		**A***	G						
BCL2L13	16485985	-250	C/T	C		**T**	C						
BCL2L13	16486110	-125	C/A	A		**A**	C						
PEX26	16934759	-483	G/T	G	G	G	G	**T***					
PEX26	16935124	-118	C/T	C	T	C	T	**T**					
PEX26	16935151	-91	C/T	C	T	T	T	**T**					
PEX26	16935165	-77	G/T	T	G	T	T	**G**					
DGCR2	17484509	-13	C/T	**C§**	T	T							
DGCR2	17484963	-467	C/T	**T**	T	C							
TSSK2	17492088	-305	C/T	**T***	C	C	C						
TSSK2	17492200	-193	G/A	**A**	A	G	A						
TSSK2	17492446	53	C/T	**C**	T	T	C						

Gene Name	SNP Position (build35)	SNP Position (Relative to TSS)	Alleles	Haplotype									
				1	2	3	4	5	6	7	8	9	10
DGCR14	17506879	-152	C/T	T	C	C	C	**C**	T	C	T		
DGCR14	17506933	-206	C/A	C	A	C	C	**A**	C	C	C		
DGCR14	17506934	-207	T/A	T	T	T	T	**T**	A	A	T		
DGCR14	17506939	-212	C/T	C	C	C	C	**C**	T	T	C		
DGCR14	17506949	-222	C/T	C	C	C	C	**T***	C	C	C		
DGCR14	17507135	-408	[A]n	8	10	9	11	**10**	10	10	12		
UFD1L	17841440	-70	G/A	A	G								
CDC45L	17841783	-192	G/T	T	T	G	**T**						
CDC45L	17841851	-124	C/G	G	G	C	**C**						
CDC45L	17842057	82	C/A	C	C	C	**A***						
CLDN5	17887472	-32	G/A	G	A								
TBX1	18118583	-196	C/T	C	T								
GNB1L	18217302	-288	C/T	T	C								
COMT	18303362	-487	T/-		T	T		T	-				
COMT	18303438	-411	C/T		C	T		T	T				
COMT	18303581	-268	C/T		T	C		T	C				
RANBP1	18478960	-513	C/T	C	C	**T***	C	C					
RANBP1	18479333	-140	C/T	C	T	**T**	C	C					
RANBP1	18479373	-100	C/A	A	A	**A**	C	C					
RANBP1	18479407	-66	G/T	T	T	**T**	T	G					
OTTHUMG00000030620	18568149	-324	G/A	G	G	**A***							
OTTHUMG00000030620	18568415	-58	G/C	C	G	**G**							
ZNF74	19072607	-406	G/A	A		G		G					
ZNF74	19072807	-206	C/T	C		C		T					
PCQAP	19186060	-362	G/A	G	A								

Gene Name	SNP Position (build35)	SNP Position (Relative to TSS)	Alleles	Haplotype									
				1	2	3	4	5	6	7	8	9	10
PIK4CA	19537526	-15	AGGCGG/------	in	del	in	**in**	in					
PIK4CA	19537615	-104	C/T	T	C	C	**C**	T					
PIK4CA	19537740	-229	C/G	G	C	C	**C**	C					
PIK4CA	19537934	-423	G/A	G	A[§]	G	**A**[§]	G					
PIK4CA	19537970	-459	C/T	T	C[§]	T	**C**[§]	T					
UBE2L3	20245975	-524	G/T	G	G	**T**[*]							
UBE2L3	20245981	-518	C/T	C	C	**T**[*]							
UBE2L3	20246020	-479	T/-	T	-	**-**							
PPM1F	20631935	-174	G/A	A	G								
PPM1F	20632073	-312	G/C	C	G								
VPREB1	20923278	-475	G/T	T	T	G[§]	**G**[§]	T	G[§]		**T**		
VPREB1	20923297	-456	C/T	T	T	T	**T**	C[§]	T		**C**[§]		
VPREB1	20923355	-398	G/A	G	A	G	**G**	G	G		**G**		
VPREB1	20923436	-317	G/A	G	A	G	**G**	A	A		**A**		
VPREB1	20923611	-142	G/A	G	G	G	**G**	G	A		**G**		
VPREB1	20923632	-121	G/A	G	G	G	**A**[*]	G	G		**G**		
VPREB1	20923683	-70	G/A	G	G	A	**A**	G	G		**G**		
VPREB1	20923848	95	C/A	A	A	A	**A**	A	A		**C**[*]		
SUHW1	21199092	52	C/G	G	G	G	C	G	C	G	**G**	G	
SUHW1	21199194	-50	G/A	A	A	A	A	A	A	A	**G**[*]	A	
SUHW1	21199209	-65	A/T	T	A	A	A	T	T	T	**T**	A	
SUHW1	21199347	-203	G/A	G	G	G	A	G	A	G	**G**	G	
SUHW1	21199361	-217	C/G	G	G	C	G	C	G	G	**C**	C	
SUHW1	21199397	-253	C/G	C	C	C	G	C	C	C	**C**	C	
SUHW1	21199519	-375	C/T	T	T	C	T	C	T	T	**C**	C	
SUHW1	21199565	-421	TTGAGA/------	del	del	del	del	in	del	del	**in**	in	
SUHW1	21199588	-444	G/A	A	A	G	G	G	G	G	**G**	G	
SUHW1	21199606	-462	C/G	C	C	G	C	C	C	C	**C**	C	
SUHW1	21199660	-516	C/T	C	C	C	C	C	T[*]	C	**C**	C	

Gene Name	SNP Position (build35)	SNP Position (Relative to TSS)	Alleles	Haplotype									
				1	2	3	4	5	6	7	8	9	10
SMARCB1	22453559	-139	G/T	**T**	G		G						
SMARCB1	22453683	-15	G/T	**G**	G		T						
SMARCB1	22453768	70	C/A	**C**[*]	A		A						
OTTHUMG00000030257	22731888	-426	G/C	C	C	C	C	C			G[§]	C	
OTTHUMG00000030257	22732037	-277	C/T	C	C	T	T	T			C	T	
OTTHUMG00000030257	22732077	-237	[CA]n	34	7	19	18	19			19	14	
OTTHUMG00000030257	22732120	-194	complex	c1	c2	c3	c4	c5			c6	c7	
CRYBB3	23920110	-268	G/T	G	T								
HPS4	25204377	-23	G/T	G	T								
HPS4	25204500	-146	TCTCCCCCGGGGCGCCGCCTC /----------	in	del								
SRR1L	25204466	62	T/A	**A**[*]	**T**[*]								
MN1	26712933	-77	G/A	A	G	A							
MN1	26712976	-34	C/A	C	A	A							
OTTHUMG00000030143	27493141	-119	G/C	C	G								
RR22_HUMAN	28036603	-136	C/T	T	C								
AP1B1	28109340	-280	C/G	G	C								
NEFH	28191054	-406	G/A	A[*]	G[*]								
NEFH	28191316	-144	C/T	T	C								
NIPSNAP1	28302132	-254	A/G	A	G								
NIPSNAP1	28302156	-278	T/G	T	G								
ZMAT5	28487684	-95	C/A	A	C	C							
ZMAT5	28487886	-297	C/T	T	C	T							
HORMAD2	28800825	123	G/T	T	G								
LIMK2	29932388	-416	G/C	C	G								
DEPDC5	30474409	-199	G/C	G	C								
HSPC117	31132909	-115	C/T	T	C								
HSPC117	31133091	-297	T/-	T	-								

Gene Name	SNP Position (build35)	SNP Position (Relative to TSS)	Alleles	Haplotype									
				1	2	3	4	5	6	7	8	9	10
OTTHUMG00000058273	31185189	-204	[GT]n	17	17	19	18	16					
OTTHUMG00000058273	31185346	-361	G/A	G	A	G	G	A					
FBXO7	31194867	-350	C/-	C	C	-	-	-					
FBXO7	31194868	-349	[G]n	3	3	3	2	4					
FBXO7	31195128	-89	G/A	A	G	A	A	G					
HMG2L1	33978027	-5	C/T	T	C								
TOM1	34020096	-302	C/T	T	C	T	T						
TOM1	34020231	-167	C/-	-	C	-	C						
TOM1	34020356	-42	G/A	A	G	G	G						
MYH9	35108649	-115	C/T	T	C								
NCF4	35580762	-322	G/A	G	G	A	A						
NCF4	35580851	-233	G/T	G	T	G	G						
NCF4	35580976	-108	G/A	G	G	G	A						
CSF2RB	35642348	-227	C/T	T	C	C	**T**	T	T	C	T		**C**
CSF2RB	35642414	-161	C/T	C	C	C	**T**	C	C	C	T		**C**
CSF2RB	35642470	-105	G/A	A	G	G	**A**	A	A	G	A		**G**
CSF2RB	35642514	-61	G/T	G	T	T	**G**	G	T	G	G		**T**
CSF2RB	35642533	-42	C/T	C	C	T	**C**	C	C	C	C		**T**
CSF2RB	35642534	-41	G/A	G	G	G	**G**	G	G	G	G		**A***
CSF2RB	35642573	-2	C/T	T	T	T	**C***	T	T	T	T		**T**
CSF2RB	35642621	46	G/T	G	T	T	**T**	T	T	G	T		**T**
OTTHUMG00000030172	35728667	-297	G/T	T	G	T							
OTTHUMG00000030172	35728749	-379	C/T	T	C	C							
MPST	35744623	-62	G/C	C	G								

Gene Name	SNP Position (build35)	SNP Position (Relative to TSS)	Alleles	Haplotype									
				1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
PSCD4	36002449	-474	C/A	C	A	A	C	A	C	C	C	A	C
PSCD4	36002596	-327	G/A	G	G	G	A	G	G	G	G	A	G
PSCD4	36002704	-219	C/T	T	C	C	T	C	C	C	C	T	C
PSCD4	36002824	-99	CGTTTGTT/--------	in	del	del	in	in	in	del	in	in	del
PSCD4	36002825	-98	[GTTT]n	8	6	5	7	8	7	6	8	7	6
PSCD4	36002878	-45	G/A	G	A	A	G	G	A	A	G	G	A
OTTHUMG00000030683	36096485	-408	C/T	C		T							
MFNG	36207059	-112	G/A	G	A								
MFNG	36207231	-284	G/A	A	G								
PDXP	36378883	-350	C/T	**T***	**C***								
GALR3	36543572	-316	C/-	-	-	C							
GALR3	36543752	-136	C/T	C	T	T							
PRKCABP	36777635	-64	G/A	A	G								
C22orf5	36993495	-1	G/A	**A***	**G***								
PGEA1	37376620	-524	C/T	**C***	**T***								
GTPBP1	37426119	-349	C/G	G	C								
GTPBP1	37426133	-335	C/T	T	C								
APOBEC3B	37817263	-455	C/T	T		**T**	T	C	T				
APOBEC3B	37817352	-366	C/T	C		**C**	T*	C	C				
APOBEC3B	37817399	-319	C/T	T		**T**	T	T	C				
APOBEC3B	37817434	-284	G/A	A		**A**	A	A	G				
APOBEC3B	37817748	30	T/C	C		**T***	C	C	C				
APOBEC3B	37817751	33	A/T	T		**T**	A	T	A				
APOBEC3B	37817794	76	C/T	T		**T**	C*	T	C*				
OTTHUMG00000030194	38222292	-483	A/G	G		A	A	**A**					
OTTHUMG00000030194	38222349	-426	G/T	G		G	T	**T**					
OTTHUMG00000030194	38222546	-229	G/A	G		G	G	**A***					

Gene Name	SNP Position (build35)	SNP Position (Relative to TSS)	Alleles	Haplotype									
				1	2	3	4	5	6	7	8	9	10
PHF5A	40189357	-142	G/A	A	**G***	**A**							
PHF5A	40189740	-525	C/T	C	**C**	**T***							
OTTHUMG00000030205	40410353	-318	C/A	C	C		A						
OTTHUMG00000030205	40410483	-188	[A]n	18	14		17						
MEI1	40419563	-439	G/T	G	G	G	G		**G**	**T***			
MEI1	40419640	-362	G/T	T	T	T	G		**G**	**T**			
MEI1	40419916	-86	G/A	G	G	G	G		**A***	**G**			
MEI1	40419964	-38	C/T	C	T	C	C		**C**	**C**			
MEI1	40420043	41	GA/--	--	--	GA	--		--	**GA**			
OTTHUMG00000030087	40520811	-300	C/T	**T***	**C**	C	C	C					
OTTHUMG00000030087	40520859	-252	G/A	**G**	**A***	G	G	G					
OTTHUMG00000030087	40520967	-144	G/A	**G**	**G**	G	A	G					
OTTHUMG00000030087	40521184	73	C/G	**G**	**G**	G	G	C					
SREBF2	40553592	-10	G/C	G	C								
OTTHUMG00000030498	40719112	-158	C/G	G	C								
NAGA	40791405	-61	G/A	G	G	A	**A**						
NAGA	40791450	-106	G/A	G	A	A	**A**						
NAGA	40791480	-136	A/T	T	A	A	**A**						
NAGA	40791879	-535	G/A	G	G	G	**A***						
OTTHUMG00000030175	40799699	-495	C/T	T	T	T	**C***	**T**					
OTTHUMG00000030175	40799715	-479	C/T	T	T	T	**T**	**C***					
OTTHUMG00000030175	40800068	-126	G/A	A	G	G	**G**	**G**					
OTTHUMG00000030175	40800203	9	G/T	T	T	G	**G**	**G**					
OTTHUMG00000030384	41240713	-116	G/C	C	G			C					
OTTHUMG00000030384	41240889	-292	C/T	T	T			C					
SERHL	41273912	-485	C/T	C	C	**T***							
SERHL	41273947	-450	G/A	G	A	**G**							
SERHL	41274041	-356	G/C	C	G*	**C**							

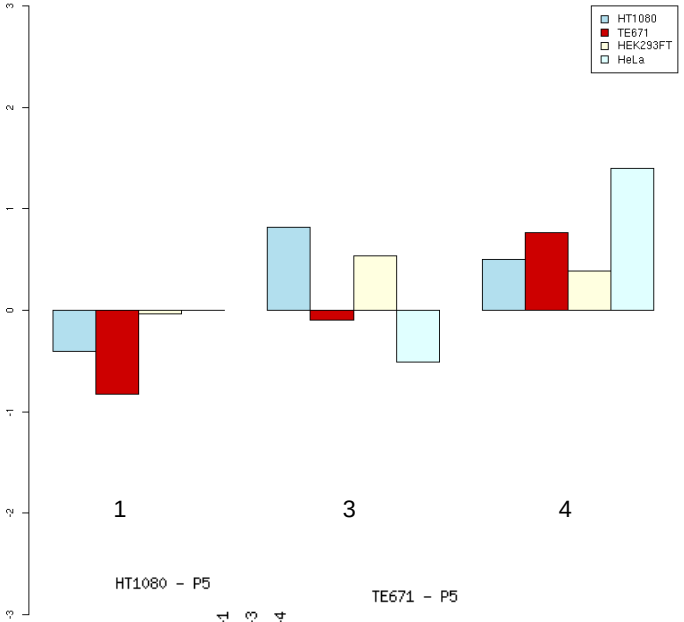
Gene Name	SNP Position (build35)	SNP Position (Relative to TSS)	Alleles	Haplotype									
				1	2	3	4	5	6	7	8	9	10
POLDIP3	41335744	-281	G/A	G	G	A	A						
POLDIP3	41335901	-438	G/A	A	G	G	A						
OTTHUMG00000030962	41863929	-29	T/A	T	T	A	**T**						
OTTHUMG00000030962	41864149	-249	C/T	T	C	T	**T**						
OTTHUMG00000030962	41864247	-347	C/A	C	C	C	**A***						
MPPED1	42131622	-370	T/C	**C***	**T***								
PNPLA5	42613207	-418	C/G	G	C								
SAMM50	42675665	-515	C/A	C	**A***	C							
SAMM50	42676159	-21	C/A	A	**C**	C							
PARVG	42901521	-236	T/A	T		A	T						
PARVG	42901556	-201	G/A	G		A	A						
NUP50	43879802	-519	G/A	A	A	A	A	A	G*				
NUP50	43879807	-514	C/A	A	A	A	A	C*	A				
NUP50	43880168	-153	G/C	G	G	C	C	C	G				
NUP50	43880278	-43	G/T	G	T	G	G	G	G				
NUP50	43880308	-13	C/A	A	A	A	C	C	C				
UPK3A	44001021	-378	G/T	G		**T***	G						
UPK3A	44001198	-201	G/A	G		**G**	A						
C22orf8	44025920	-431	A/T	A	T								
C22orf8	44026241	-110	GGGCG/-----	in	del								
RIBC2	44129732	-388	G/A	A	G	G	A	G					
RIBC2	44129999	-121	G/A	A	A	G	A	A					
RIBC2	44130161	41	C/A	C	C	A	A	A					
SMC1L2	44129995	40	G/A	**G**	G	**A***	**G**	G	G	G			
SMC1L2	44130161	-126	G/T	**T**	G	**T**	**G**	T	T	G			
SMC1L2	44130235	-200	C/T	**T**	T	**T**	**C**	T	C	C			
SMC1L2	44130284	-249	G/A	**A**	A	**A**	**G***	A	A	A			
SMC1L2	44130303	-268	G/A	**A***	G	**G**	**G**	G	G	G			
SMC1L2	44130393	-358	C/T	**C**	C	**T***	**C**	C	C	C			

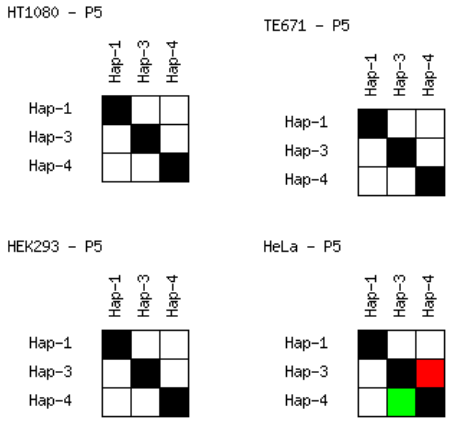
Gene Name	SNP Position (build35)	SNP Position (Relative to TSS)	Alleles	Haplotype									
				1	2	3	4	5	6	7	8	9	10
OTTHUMG00000030109	44813989	-335	C/T	**T**	T	C	T						
OTTHUMG00000030109	44814307	-17	C/T	**C***	T	T	T						
OTTHUMG00000030109	44814371	47	C/T	**T**	T	C	C						
OTTHUMG00000030672	44966694	13	G/A	G	**A***	G							
OTTHUMG00000030672	44967128	-421	G/A	A	**A**	G							
PKDREJ	44979705	31	G/T	G*	T	T							
PKDREJ	44979778	-42	G/A	G	A*	G							
PKDREJ	44980195	-459	CAAA/----	in	in	del							
TBC1D22A	45478698	-356	G/A	G	G	G	G	A					
TBC1D22A	45478854	-200	C/A	C	C	C	A	C					
TBC1D22A	45478963	-91	C/T	T	T	C	T	C					
TBC1D22A	45478995	-59	G/C	G	C	C	C	C					
AK057318	45632453	-20	G/A		A	G	A	G	A				
AK057318	45632617	-184	G/A		A	G	G	G	G				
AK057318	45632637	-204	G/C		G	C	G	G	C				
AK057318	45632836	-403	G/A		A	G	G	G	G				

Appendix E - Luciferase reporter assay results and sequence-confirmed haplotypes

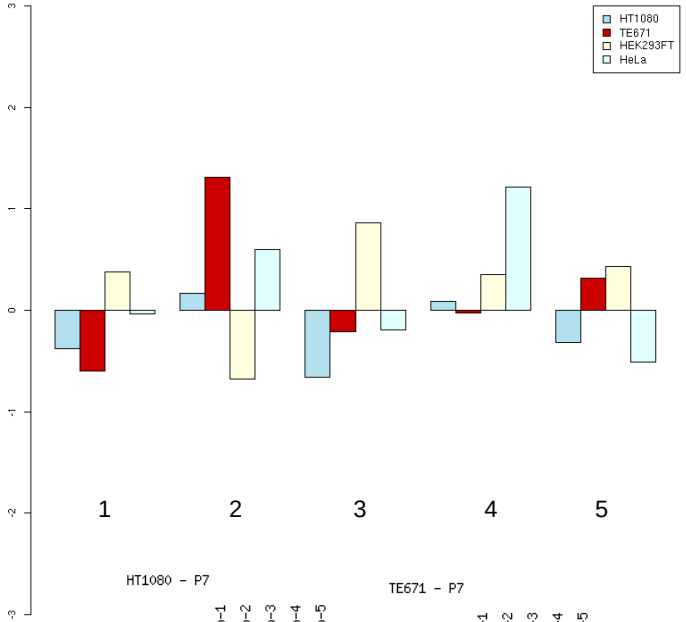
This appendix contains the results of the luciferase reporter assays on the library of 293 cloned promoter haplotypes. Only promoters that were active in at least one of the four cell lines are shown. Each promoter is presented in a separate panel with a bar chart showing the Z score for each cell line, calculated as described in section 4.1.12. The results of the Tukey's HSD tests for the haplotypes in each cell line are shown as four separate matrices, with the coloured cells in each matrix representing a comparison of a pair of haplotypes. Green shaded cells indicate that the activity of the haplotype in the row is significantly higher than that of the haplotype in the corresponding column. Red shading indicates significantly lower activity using the same orientation. Pale green and red shading have the same meaning, but are for comparisons where one of the two haplotypes fell below the 7x activity threshold. The diagonal of each matrix shows the activity level of the haplotype, with black cells indicating active haplotypes and grey cells indicating inactive haplotypes. White cells signify no reproducibly significant activity level differences between the two haplotypes. Significance was inferred as described in section 4.1.12.

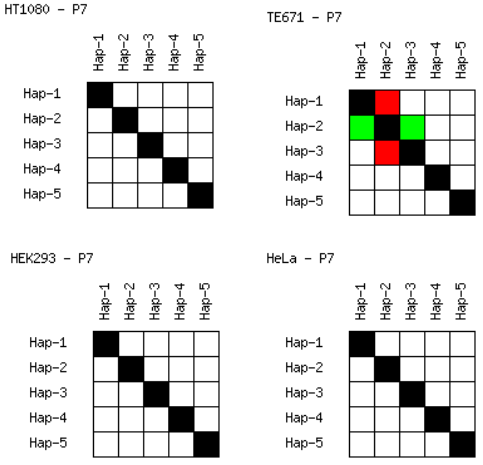
P5 – BCL2L13



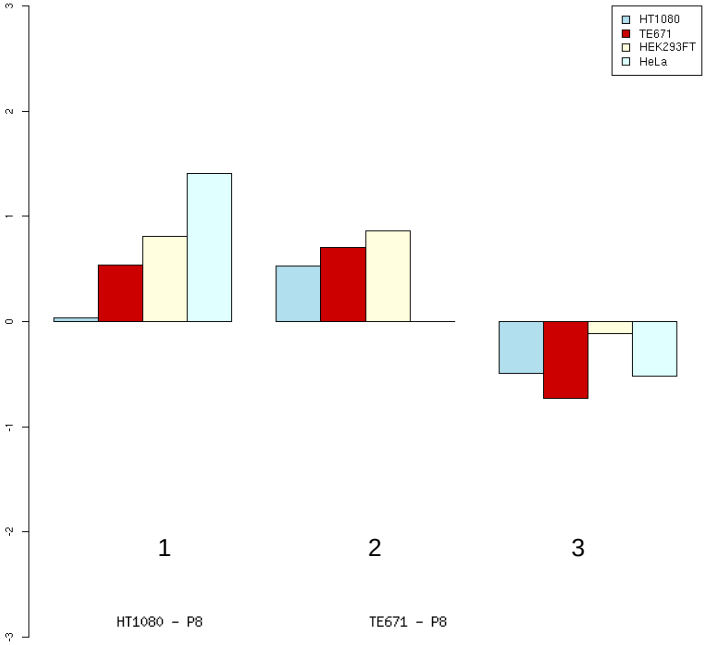


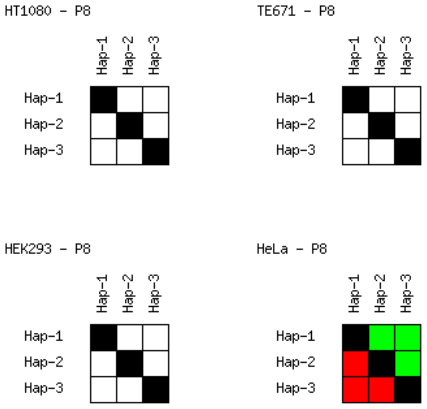
P7 – PEX26



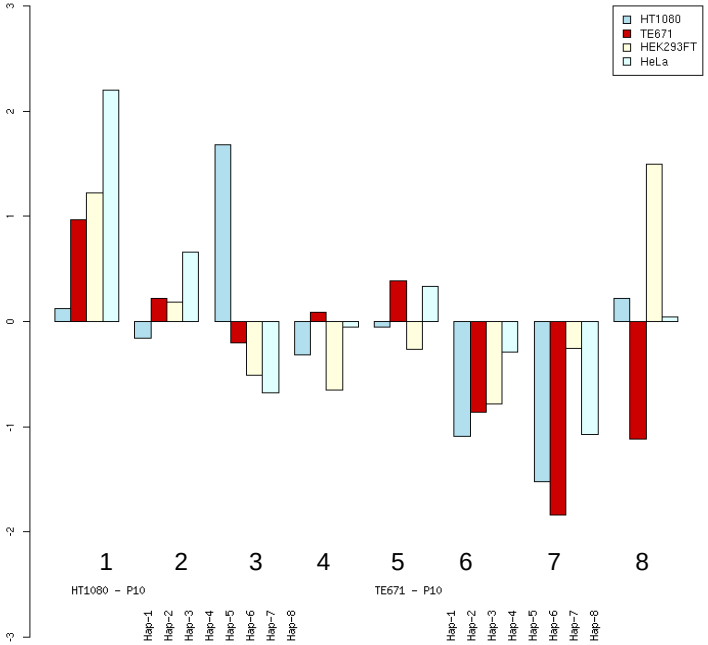


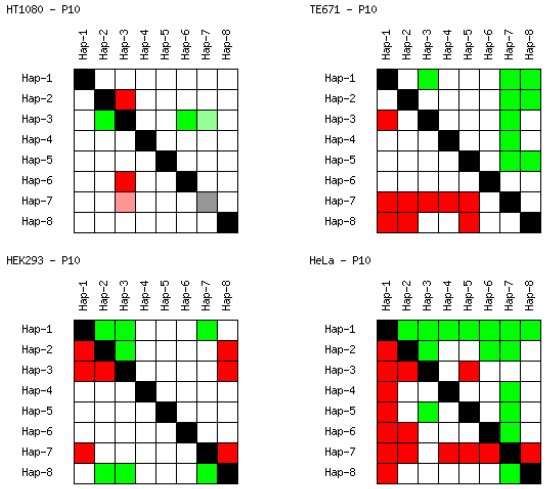
P8 – DGCR2



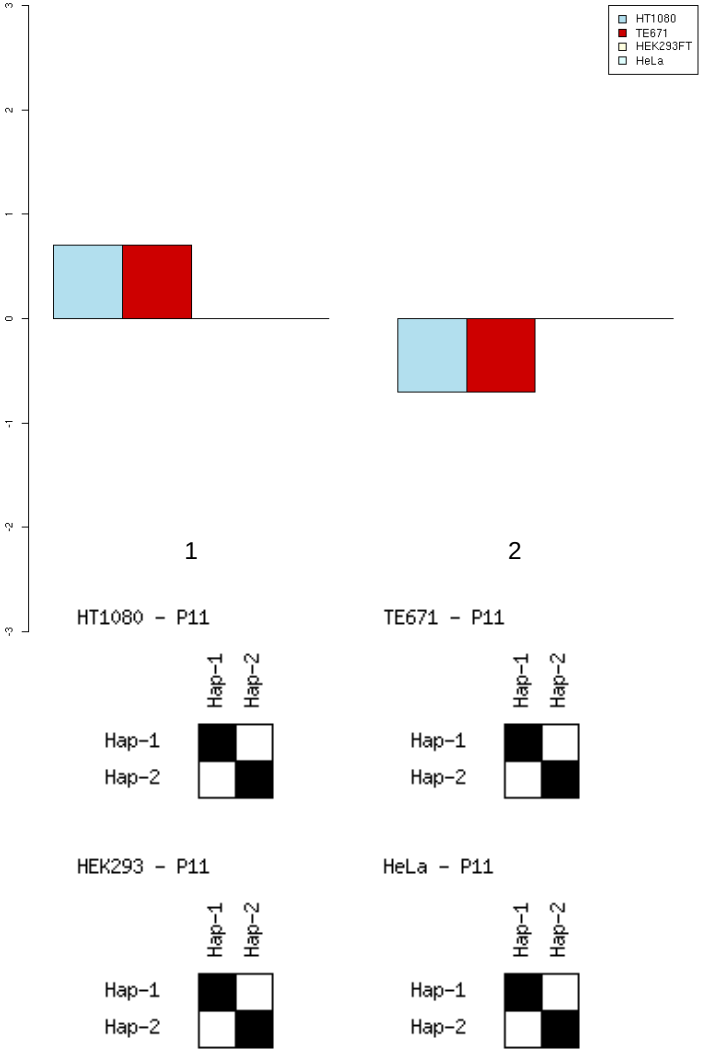


P10 – DGCR14

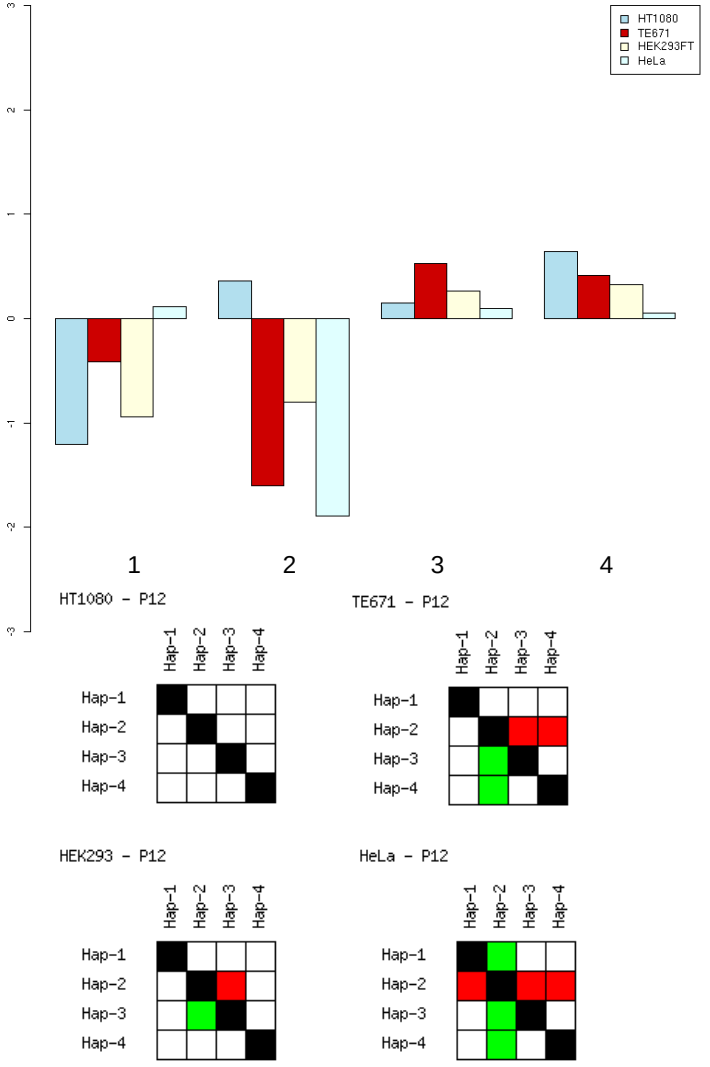




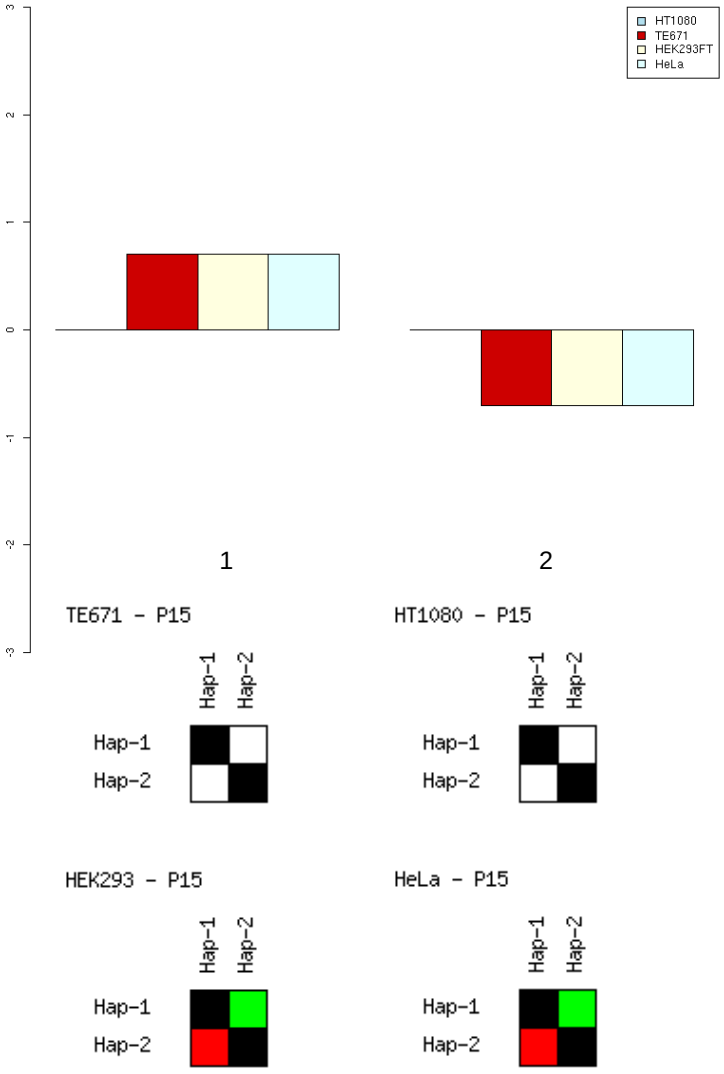
P11 – UFD1L



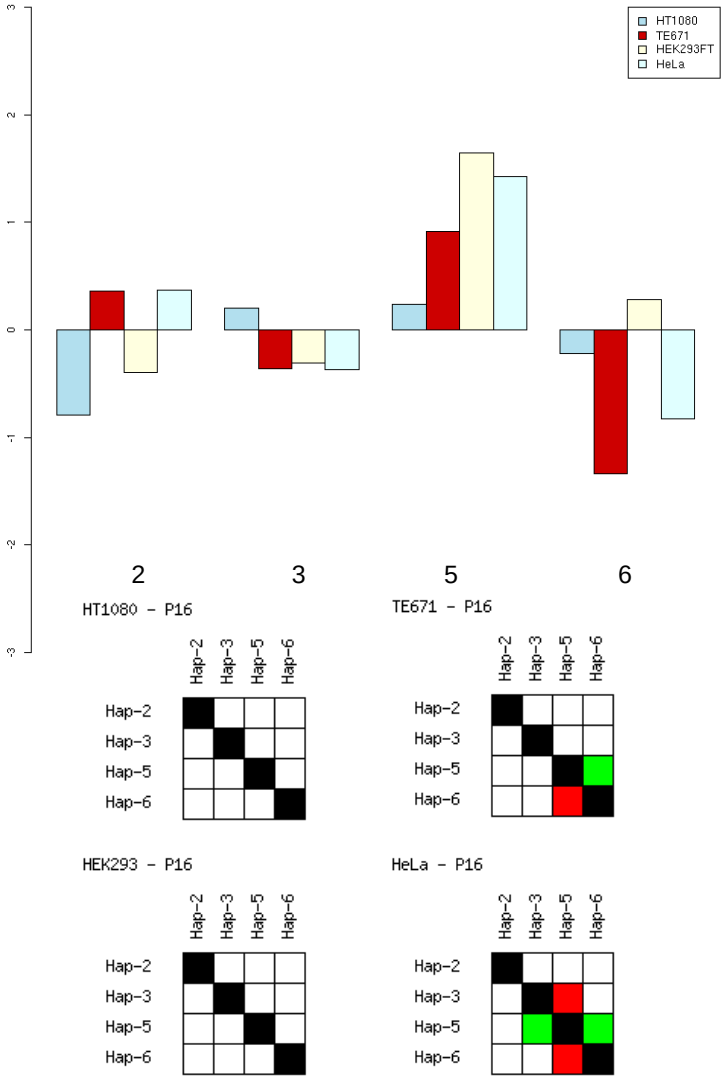
P12 – CDC45L



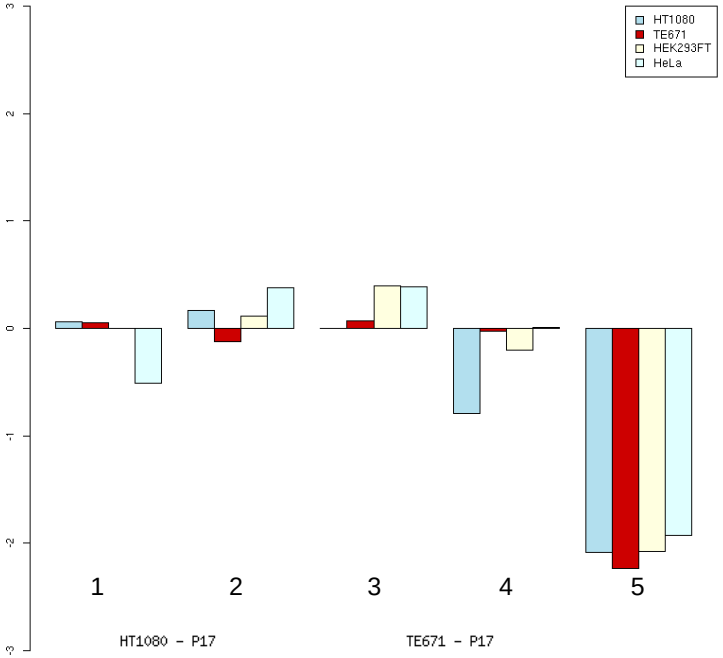
P15 – GNB1L

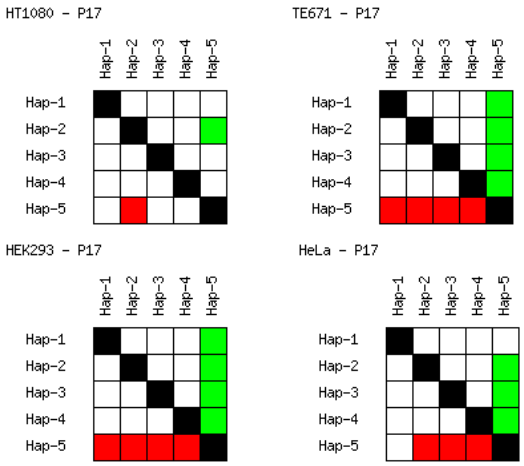


P16 – COMT

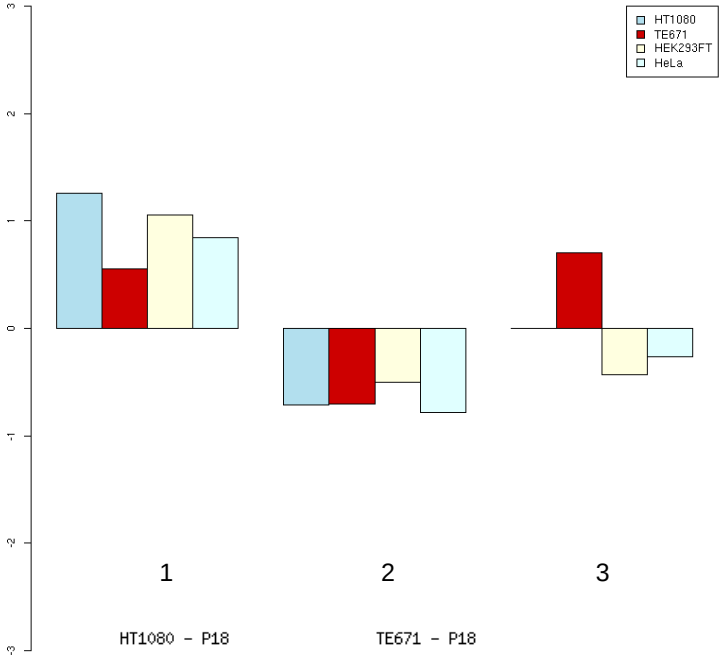


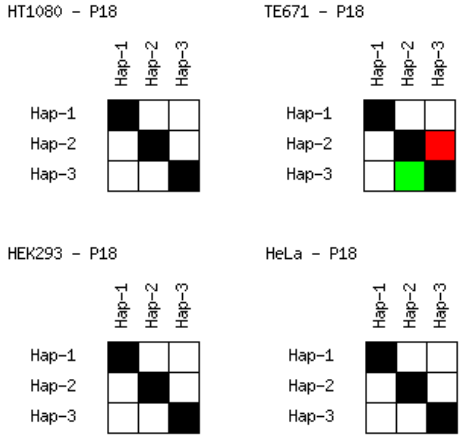
P17 – RANBP1



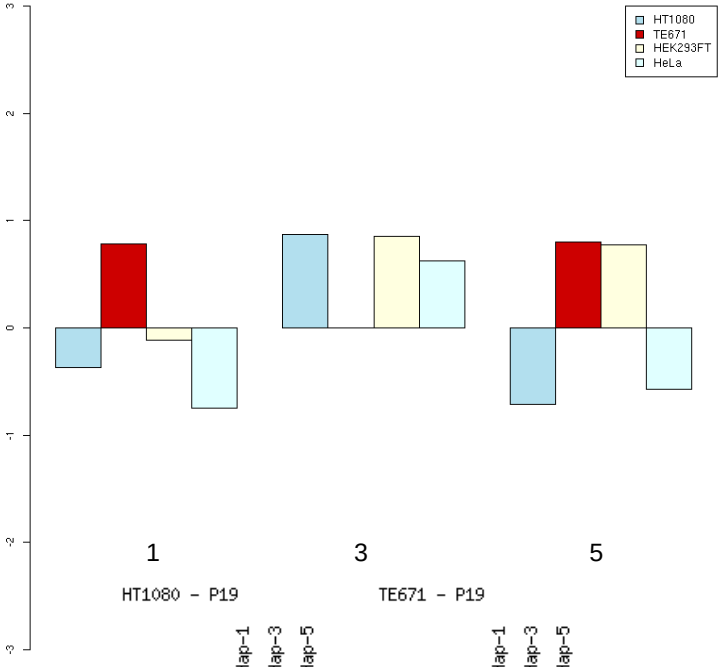


P18 – OTTHUMG00000030620

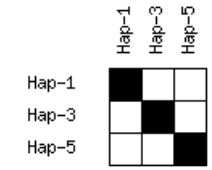




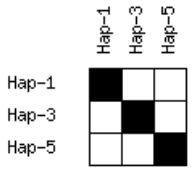
P19 – ZNF74



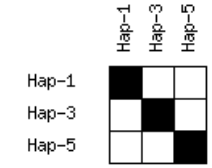
HT1080 – P19



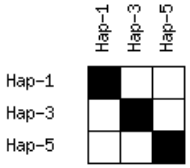
TE671 – P19



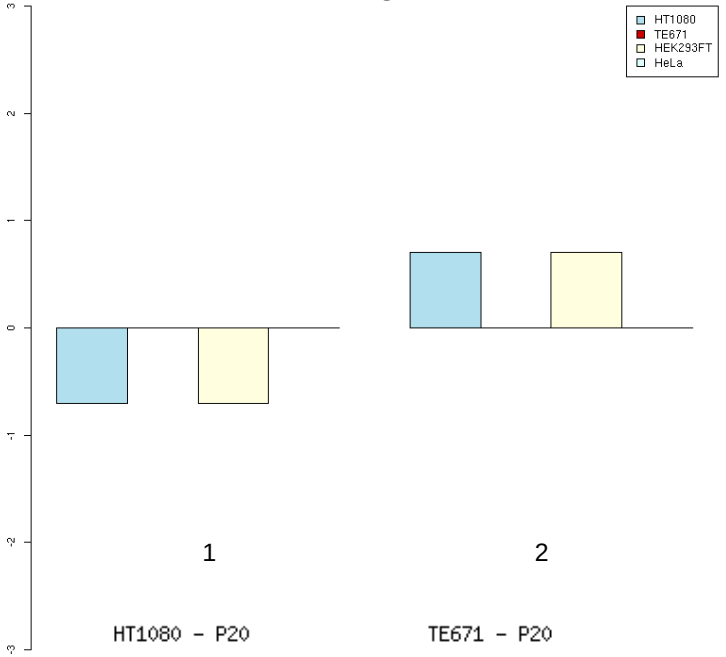
HEK293 – P19



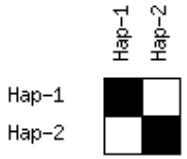
HeLa – P19



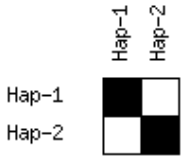
P20 – PCQAP



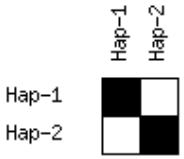
HT1080 – P20



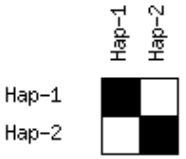
TE671 – P20



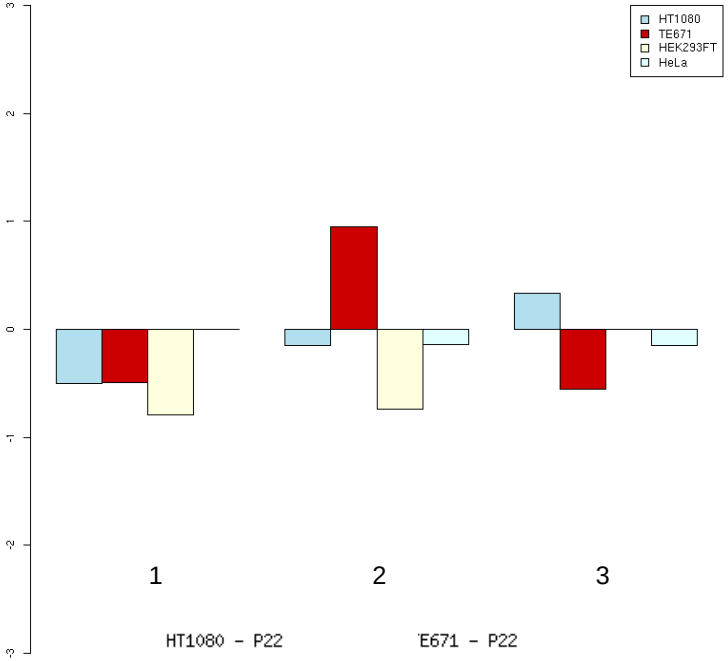
HEK293 – P20

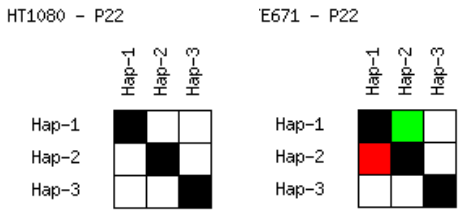


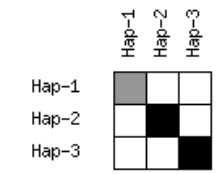
HeLa – P20

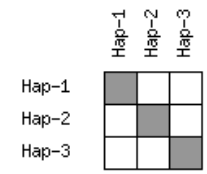


P22 – UBE2L3

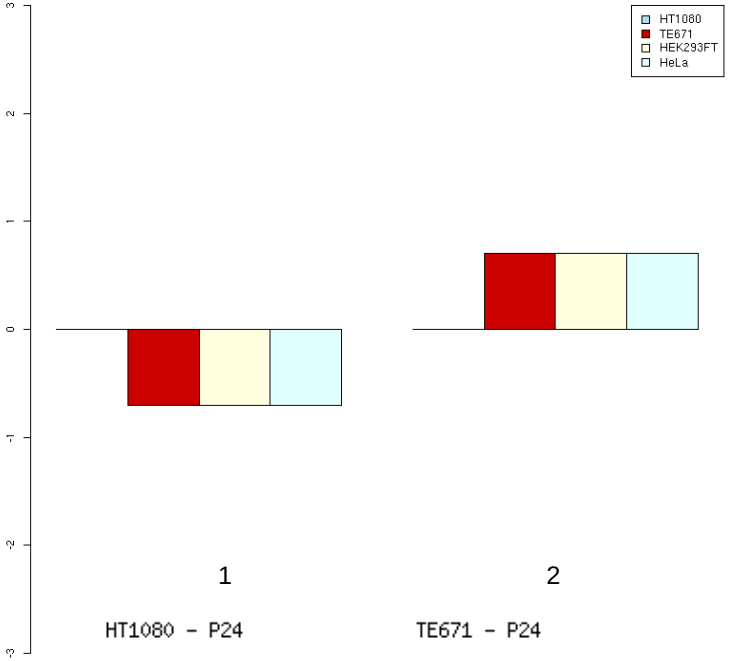


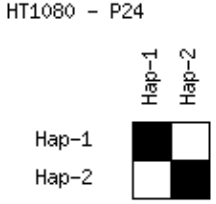


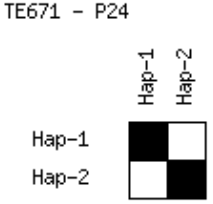


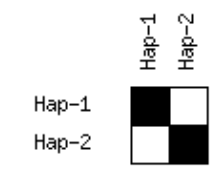


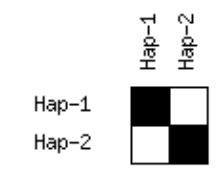
P24 – PPM1F



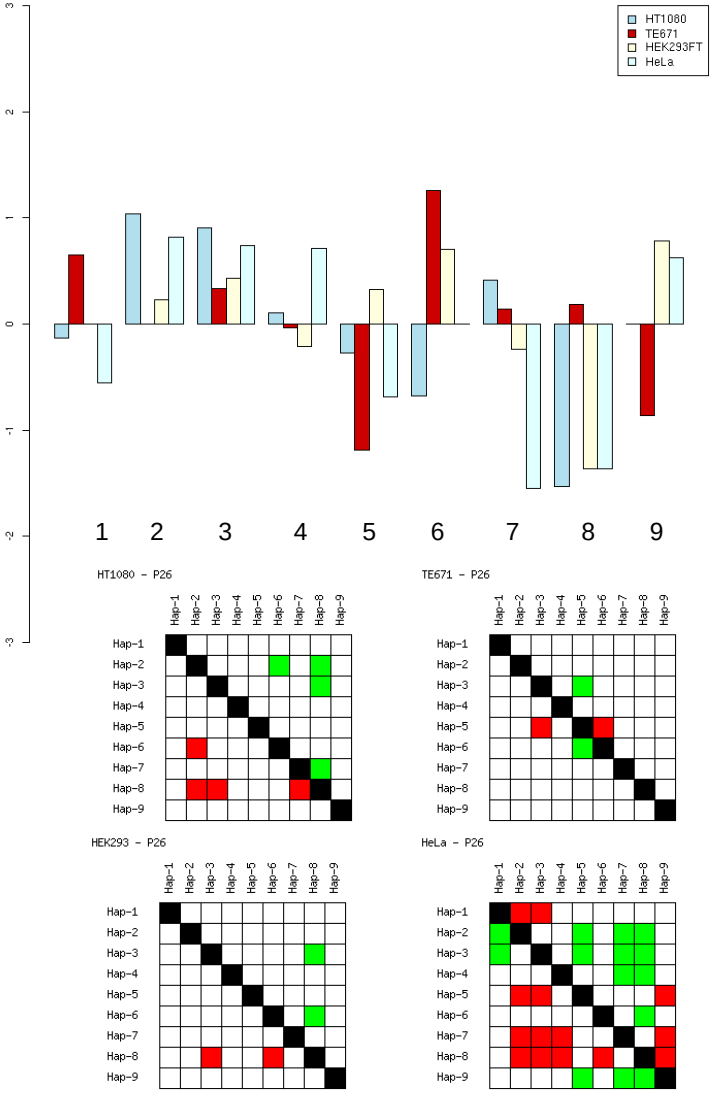




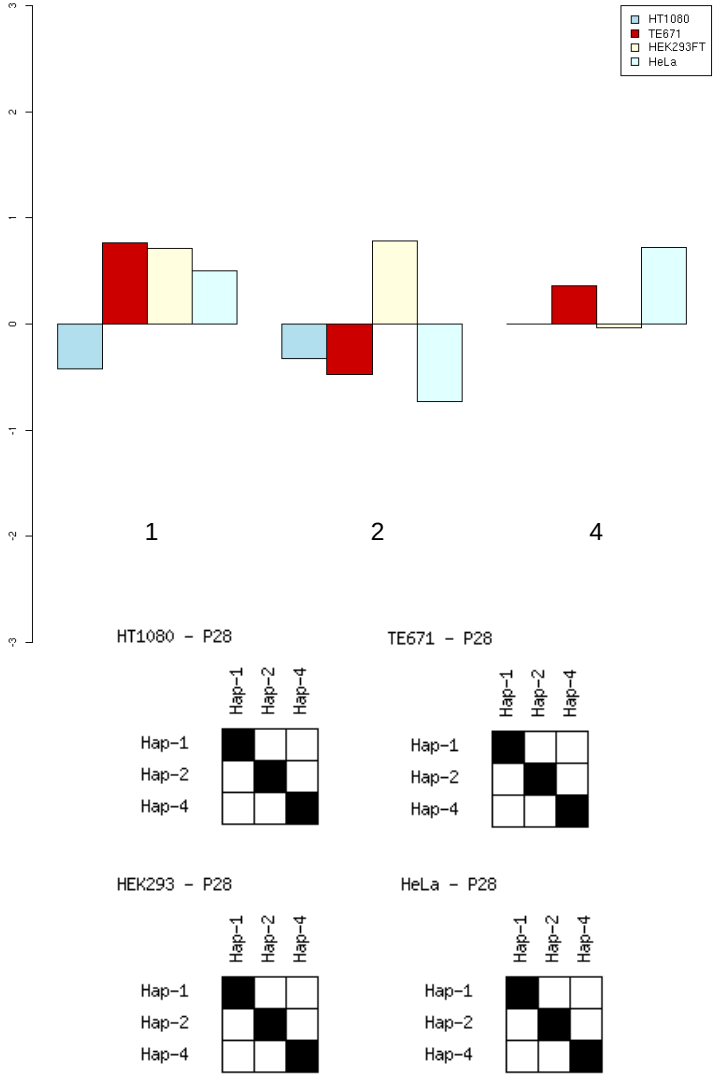




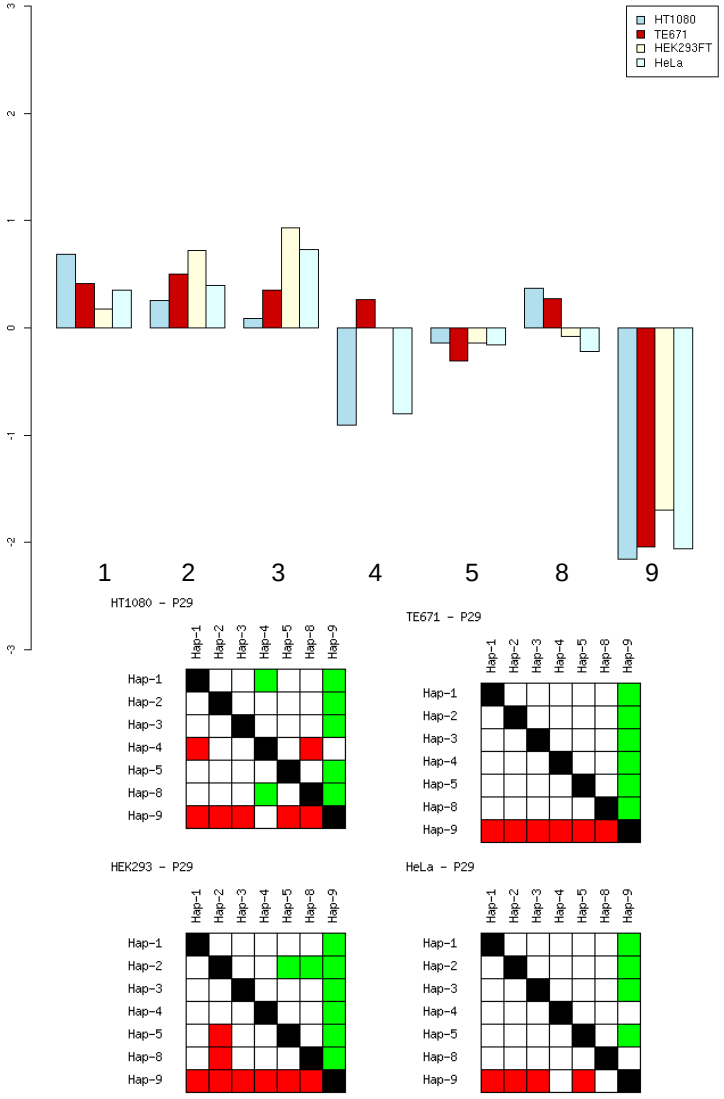
P26 – SUHW1



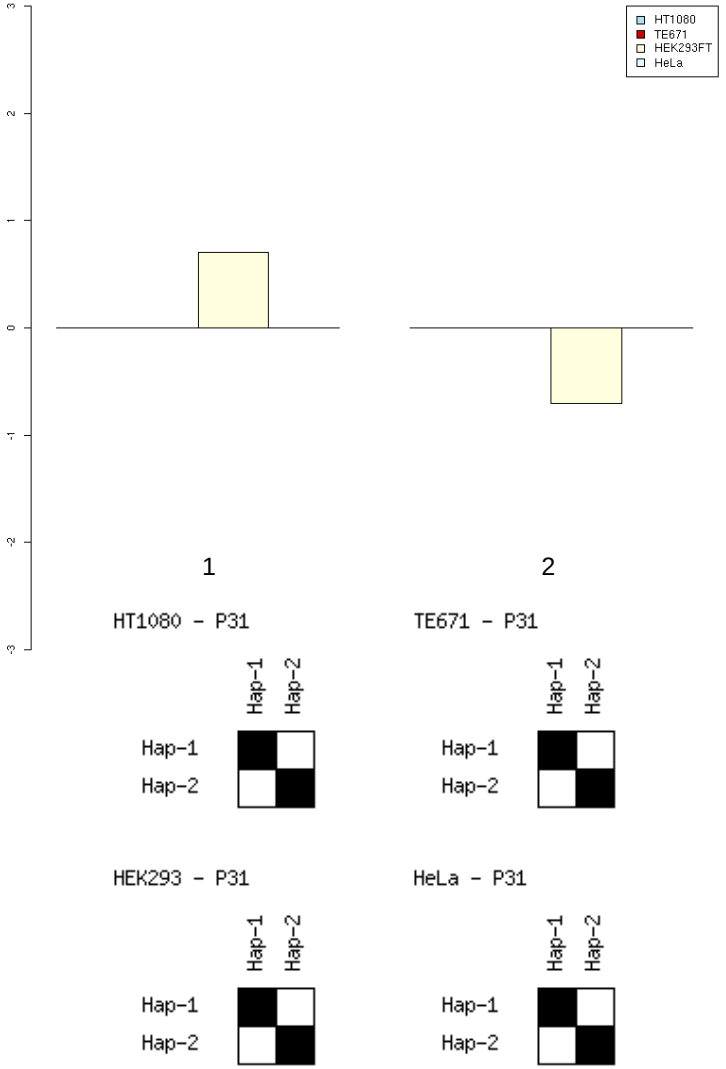
P28 – SMARCB1



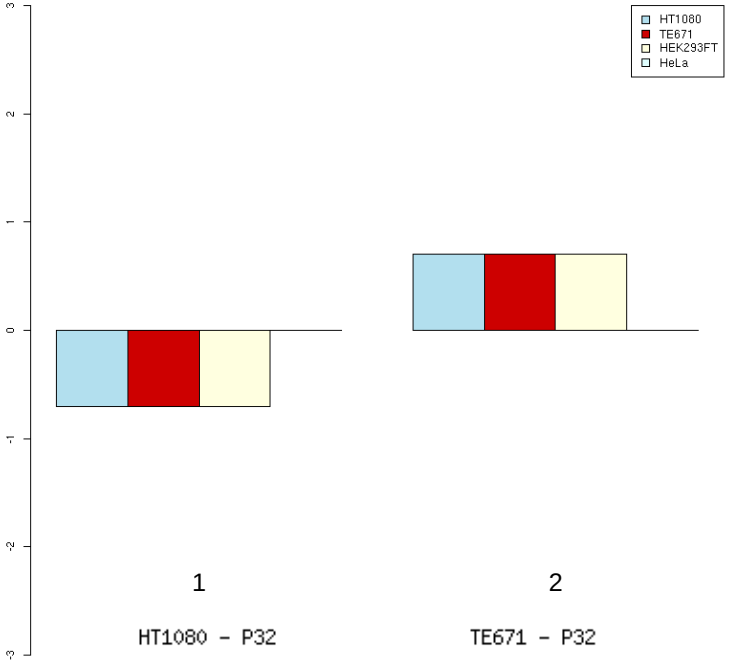
P29 – OTTHUM00000030257



P31 – SRR1L

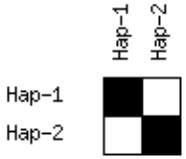


P32 – HPS4

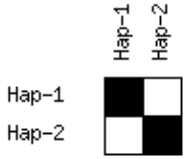


1 2

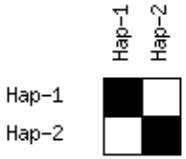
HT1080 – P32



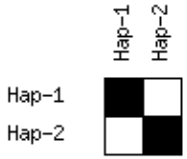
TE671 – P32



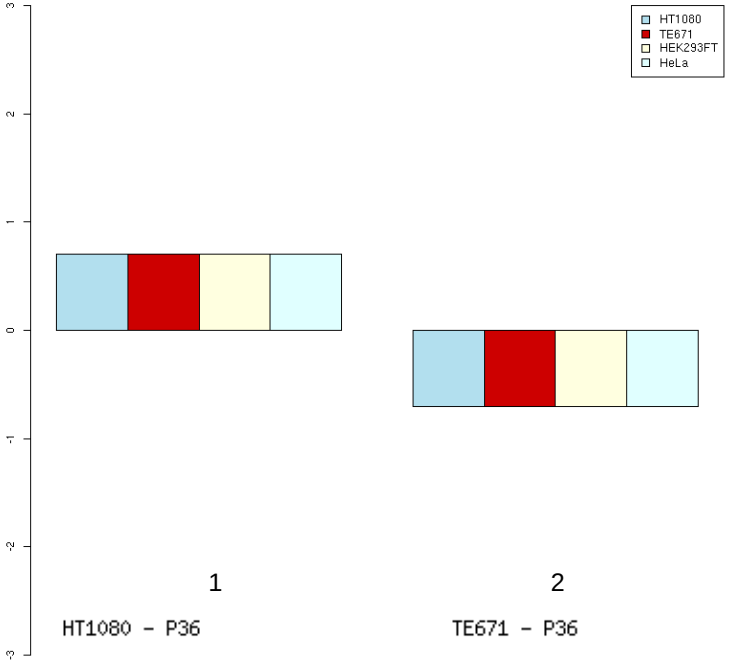
HEK293 – P32



HeLa – P32

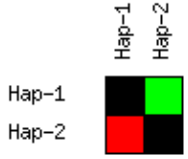


P36 – OTTHUM00000030143

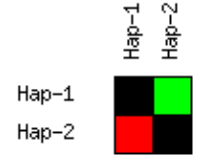


1 2

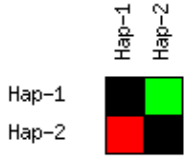
HT1080 – P36



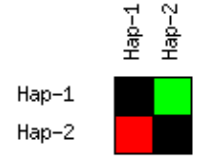
TE671 – P36



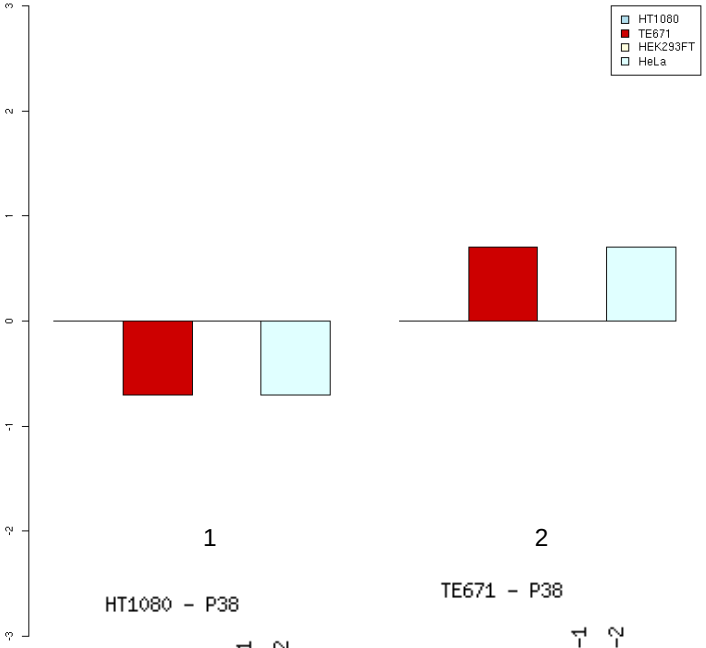
HEK293 – P36



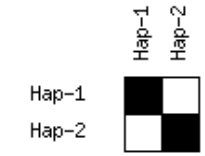
HeLa – P36



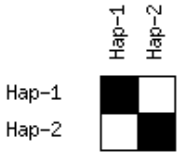
P38 – AP1B1



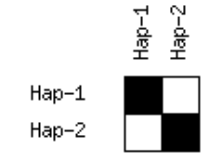
HT1080 – P38



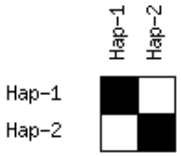
TE671 – P38



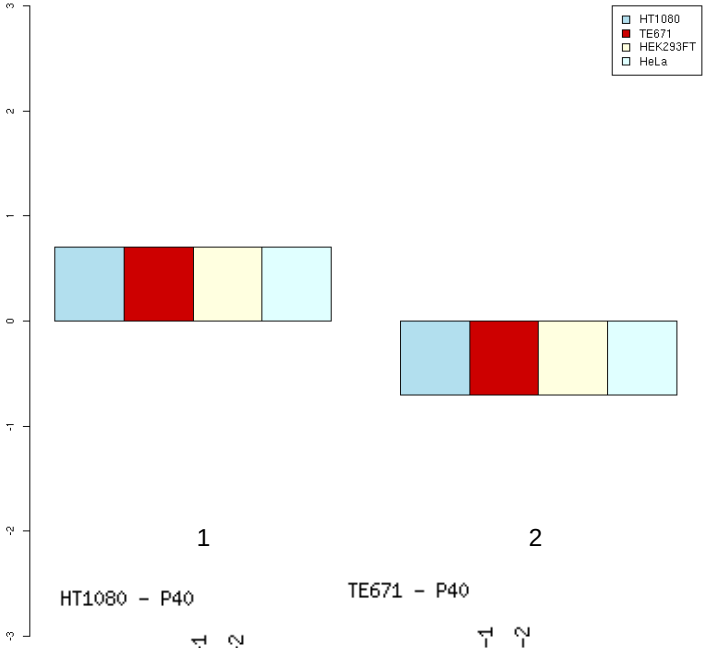
HEK293 – P38



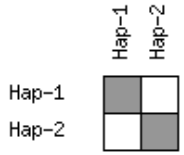
HeLa – P38



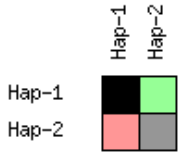
P40 – NIPSNAP1



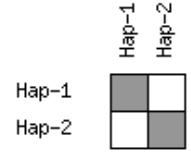
HT1080 – P40



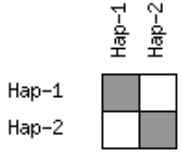
TE671 – P40



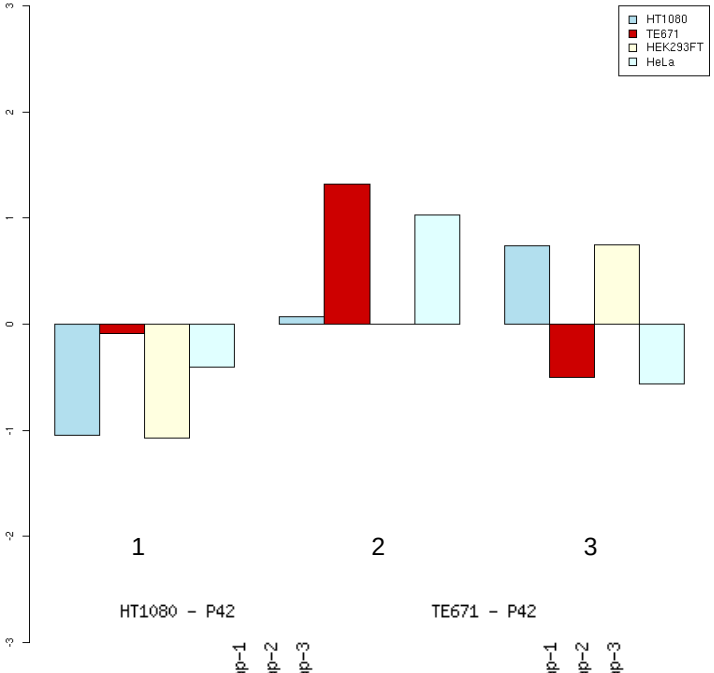
HEK293 – P40

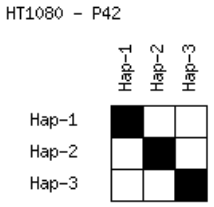


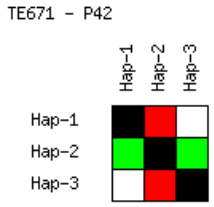
HeLa – P40

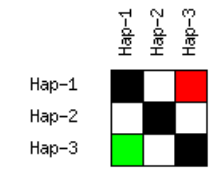


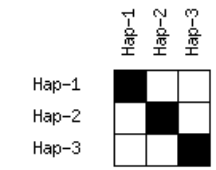
P42 – ZMAT5



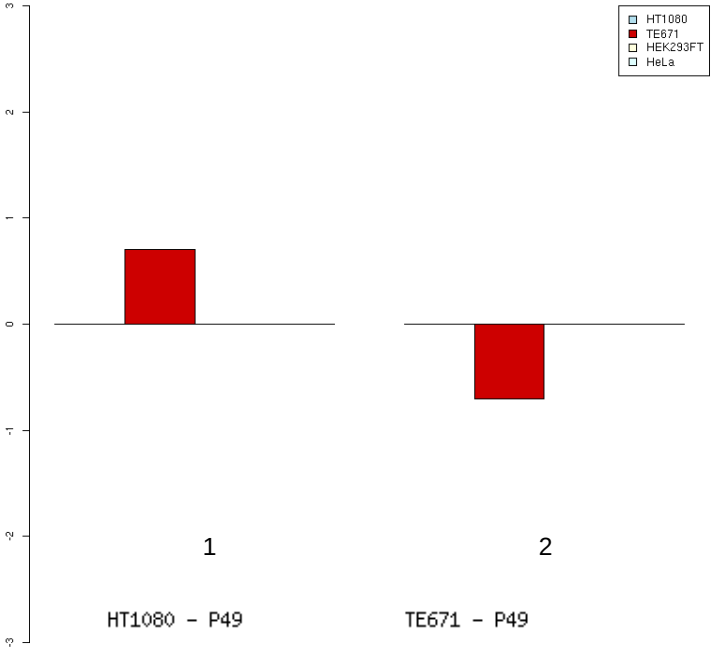


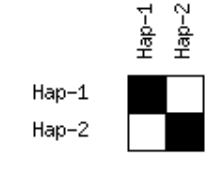


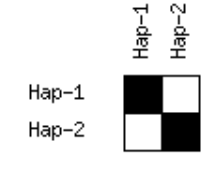




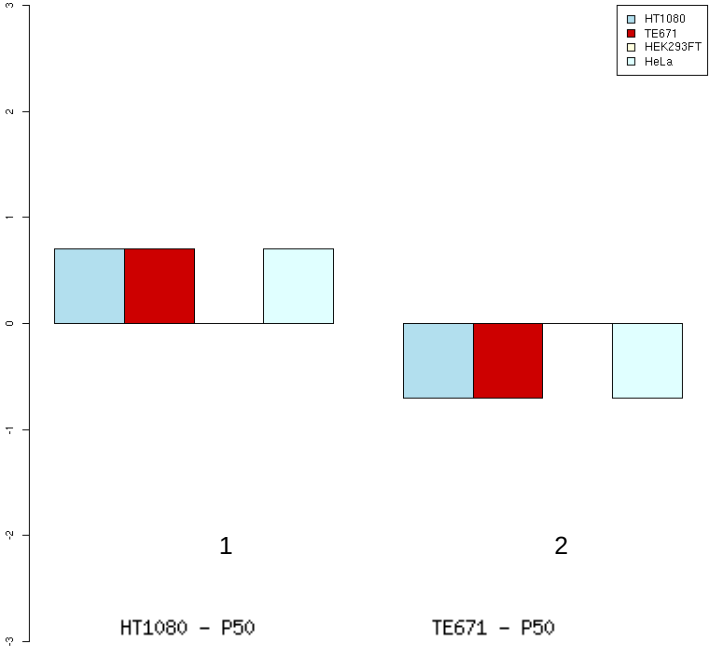
P49 – LIMK2



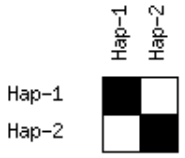




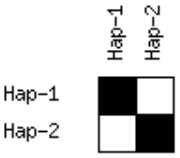
P50 – DEPDC5



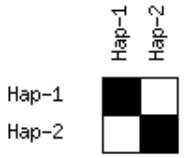
HT1080 – P50



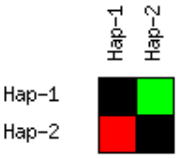
TE671 – P50



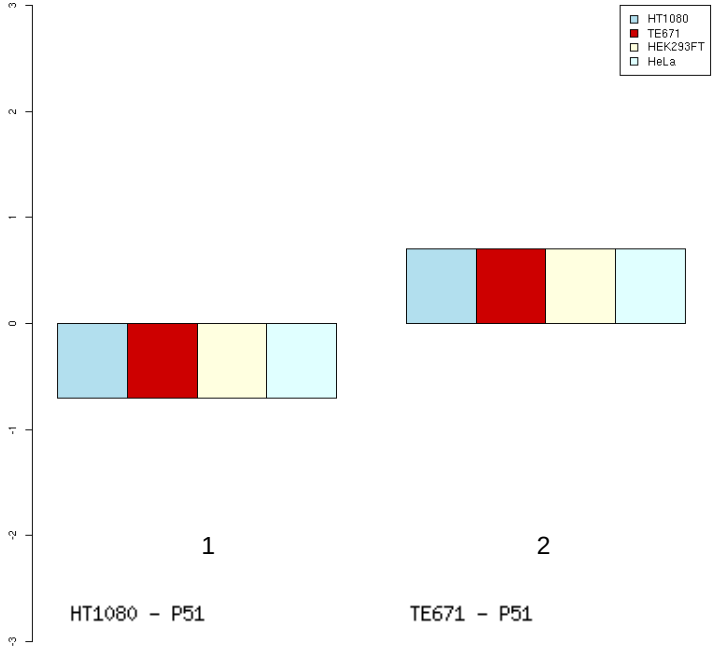
HEK293 – P50



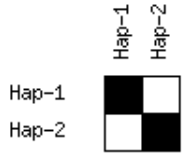
HeLa – P50



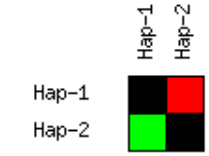
P51 – HSPC117



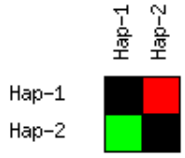
HT1080 – P51



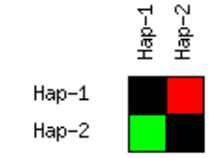
TE671 – P51



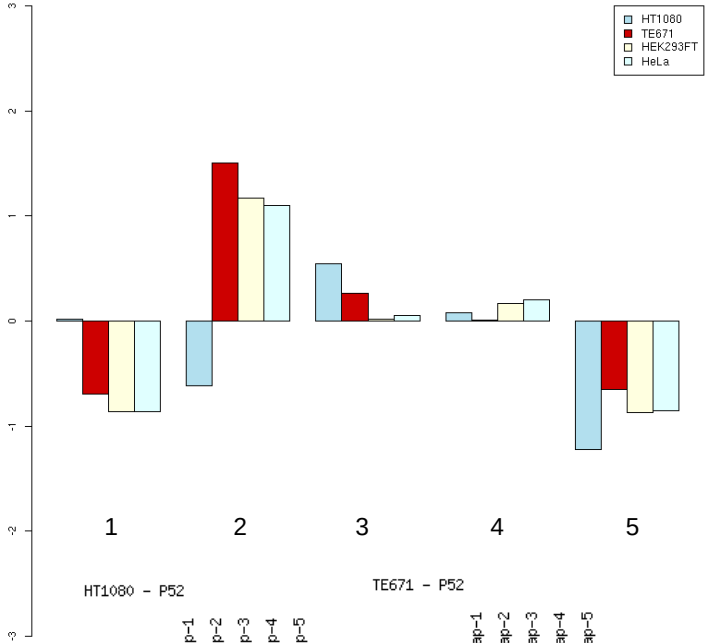
HEK293 – P51

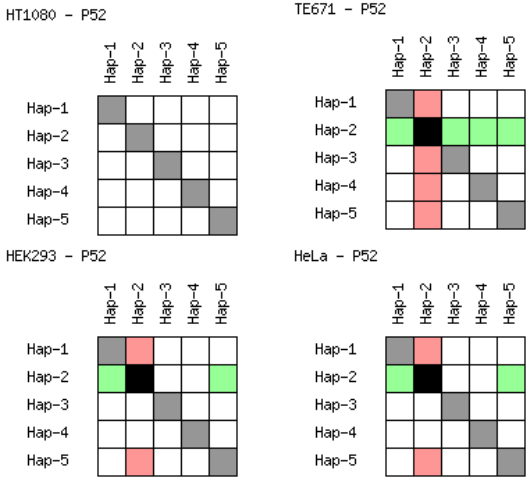


HeLa – P51

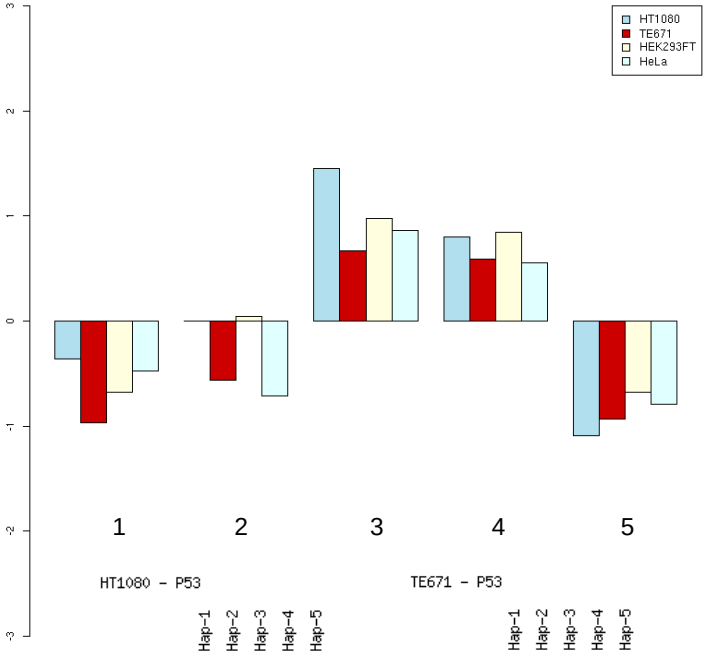


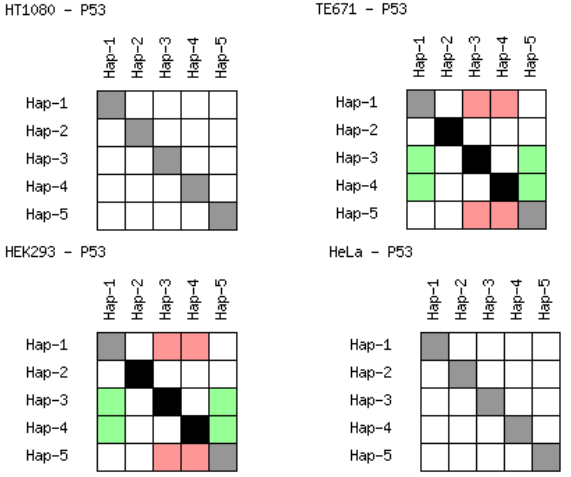
P52 – OTTHUM00000058273



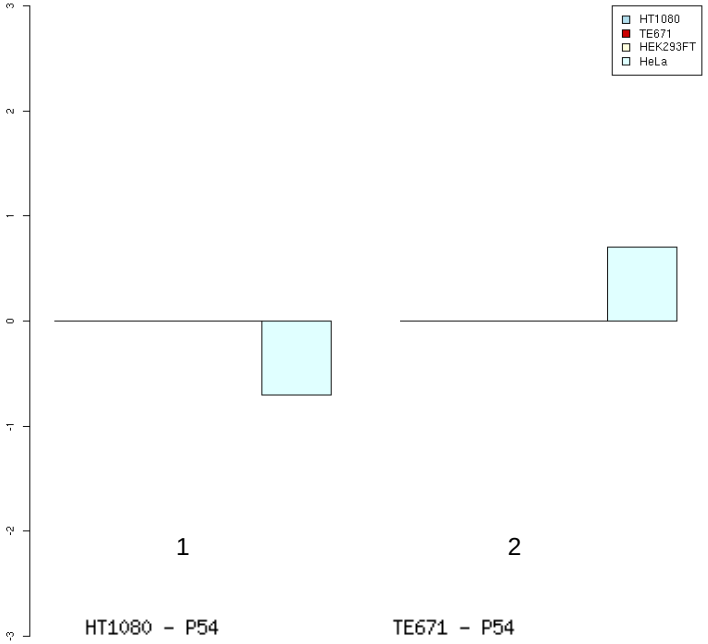


P53 – FBXO7

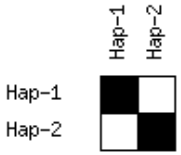




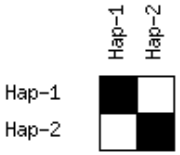
P54 – HMG2L1



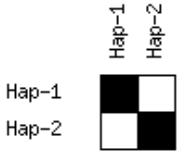
HT1080 – P54



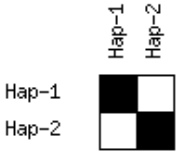
TE671 – P54



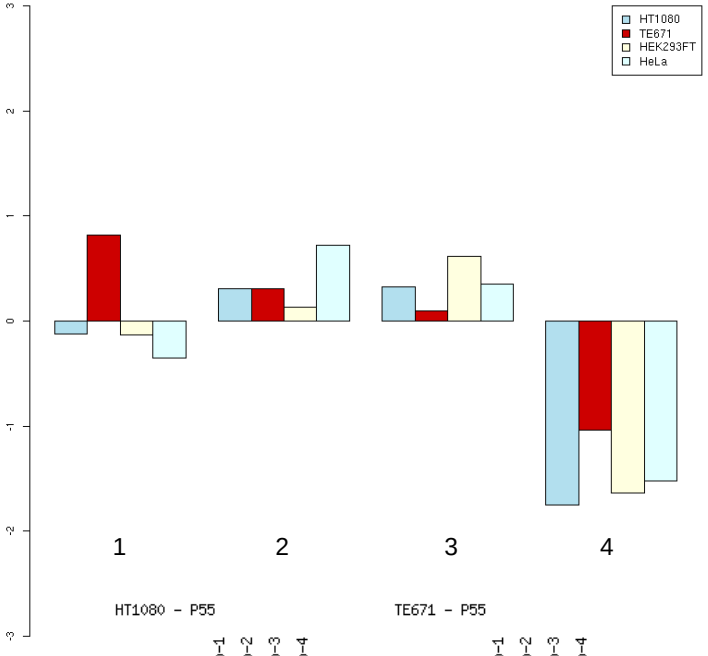
HEK293 – P54



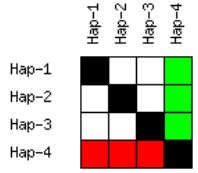
HeLa – P54



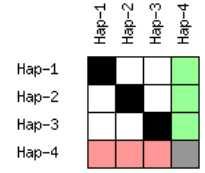
P55 – TOM1



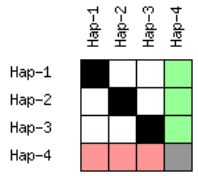
HT1080 – P55



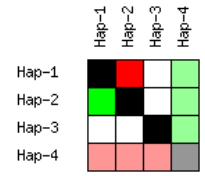
TE671 – P55



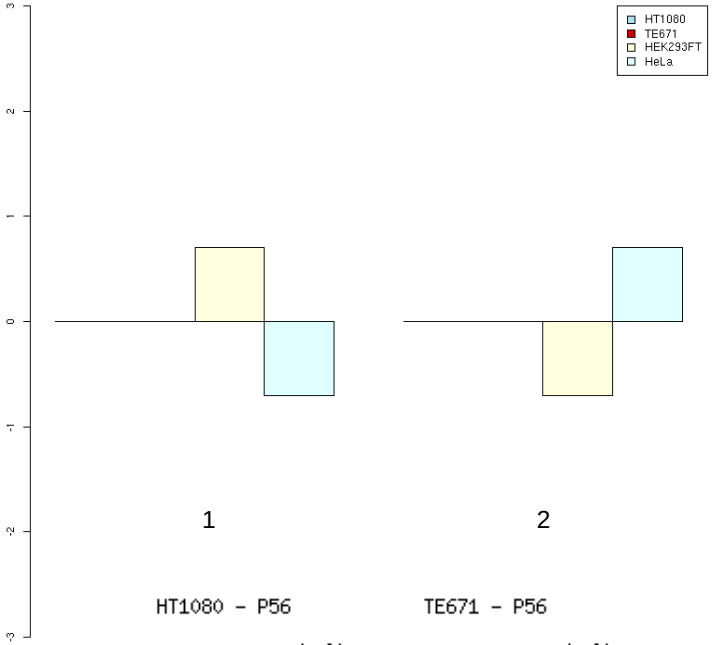
HEK293 – P55

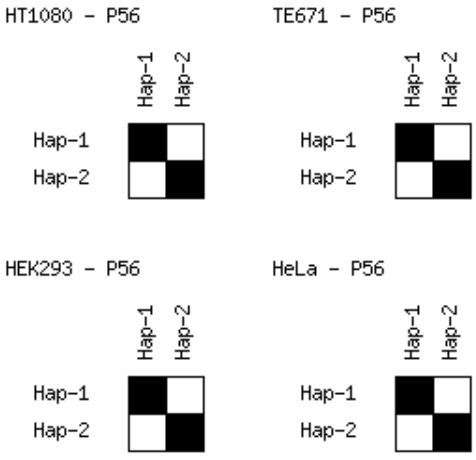


HeLa – P55

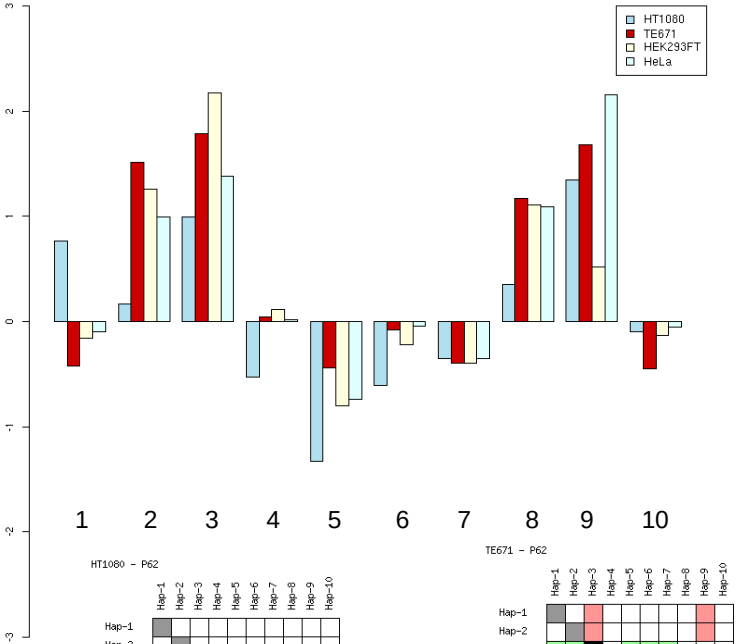


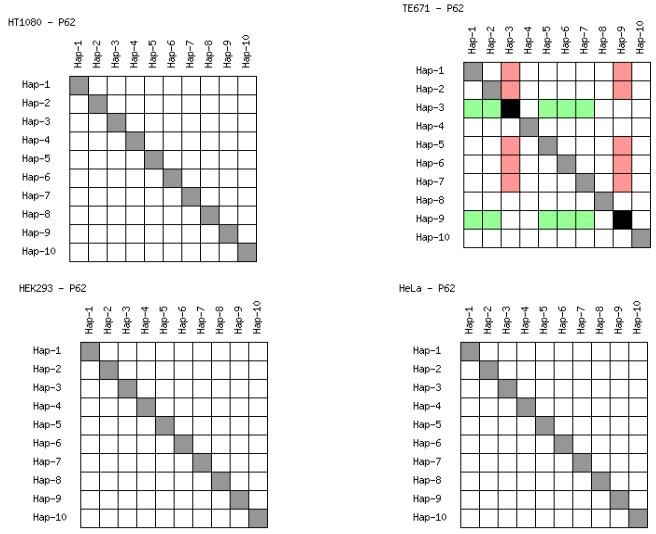
P56 – MYH1



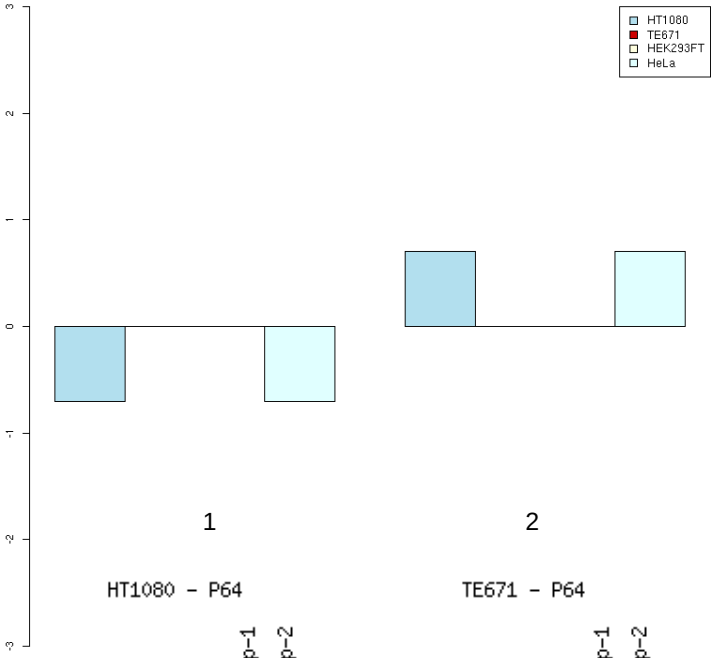


P62 – PSCD4

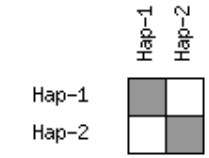




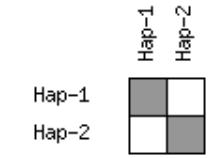
P64 – MFNG



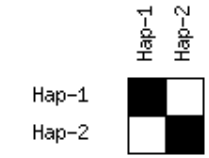
HT1080 – P64



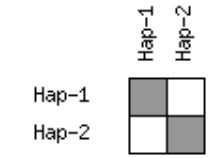
TE671 – P64



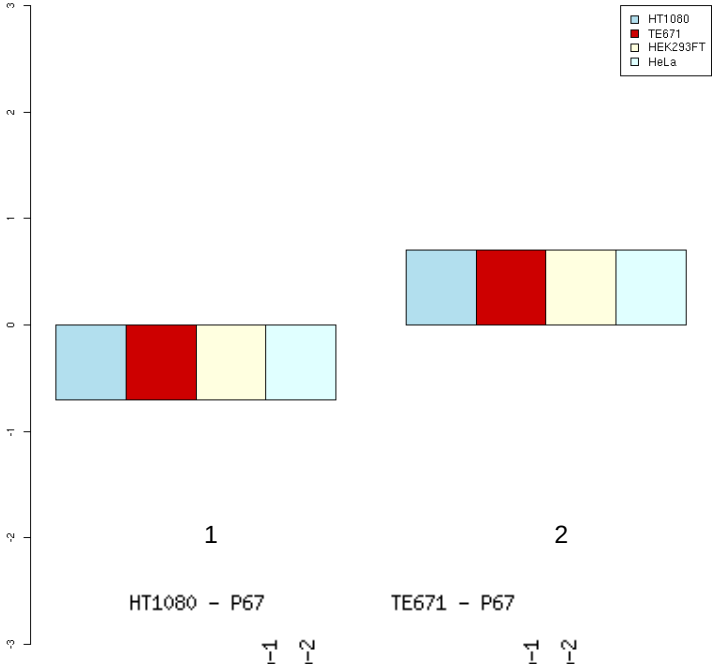
HEK293 – P64



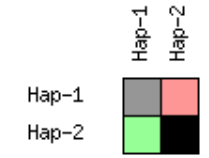
HeLa – P64



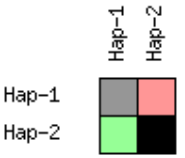
P67 – PRKCABP



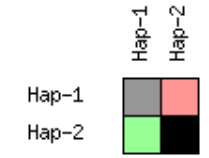
HT1080 – P67



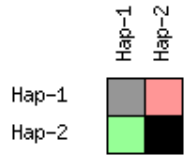
TE671 – P67



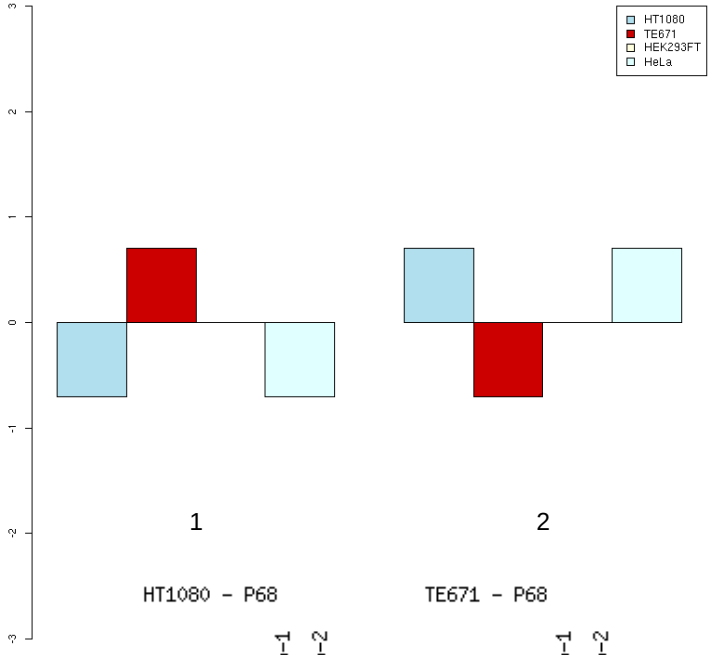
HEK293 – P67



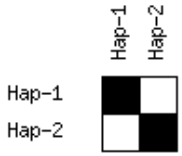
HeLa – P67



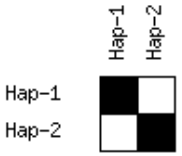
P68 – C22orf5



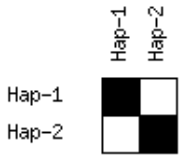
HT1080 – P68



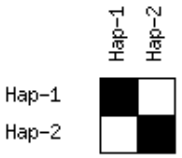
TE671 – P68



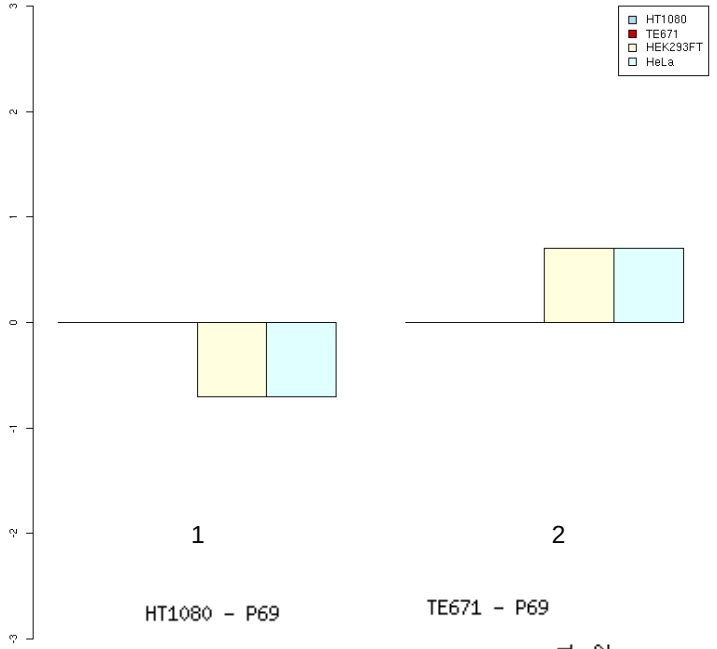
HEK293 – P68



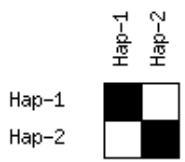
HeLa – P68



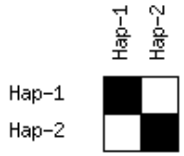
P69 – PGEA1



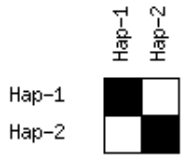
HT1080 – P69



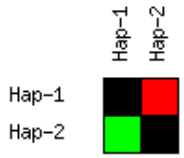
TE671 – P69



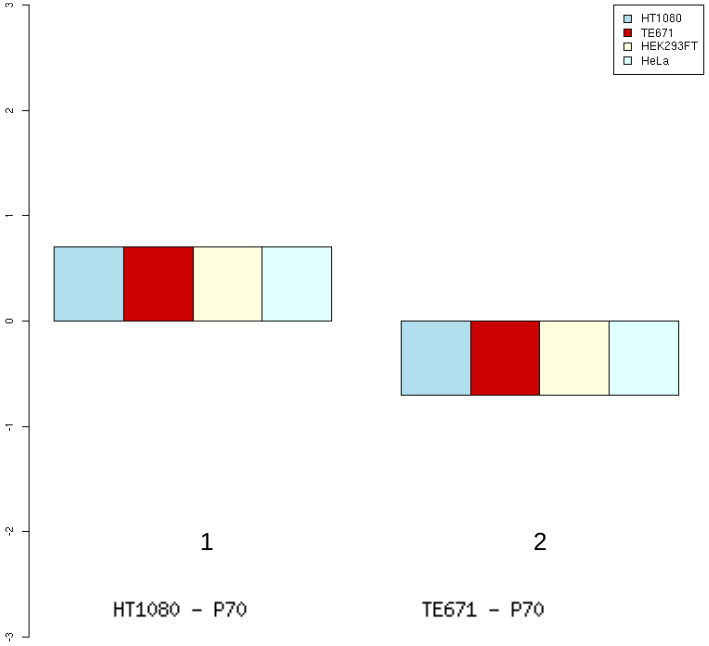
HEK293 – P69



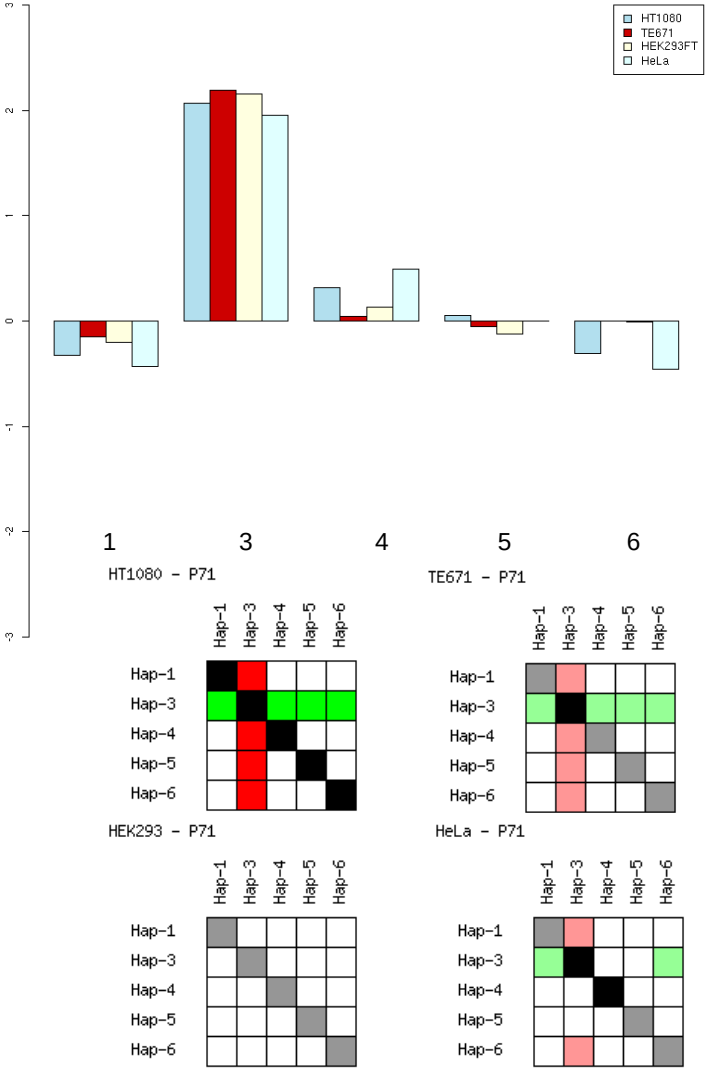
HeLa – P69



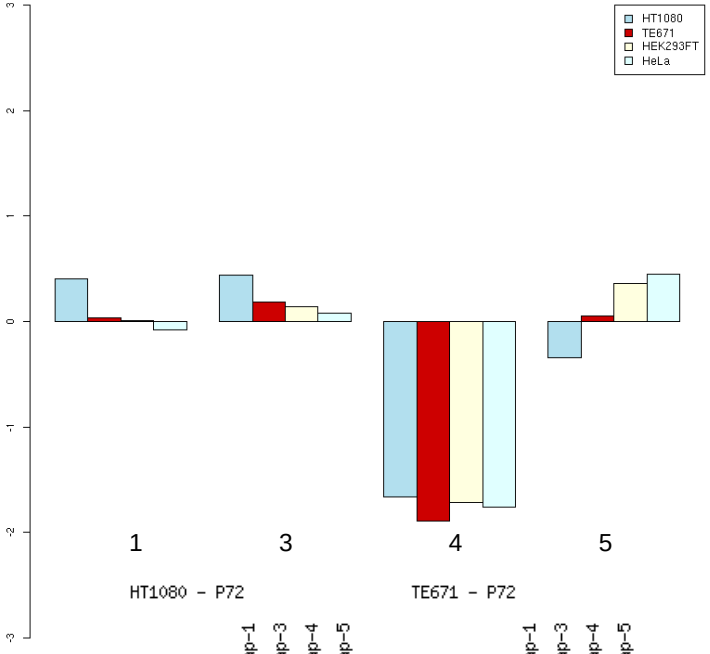
P70 – GTPBP1



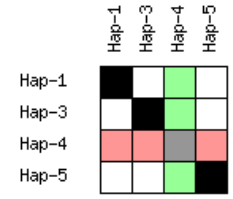
P71 – APOBEC3B



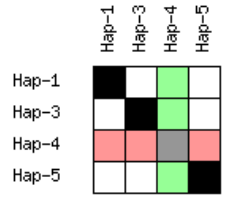
P72 – OTTHUM00000030194



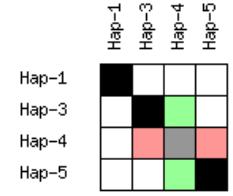
HT1080 – P72



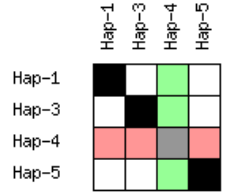
TE671 – P72



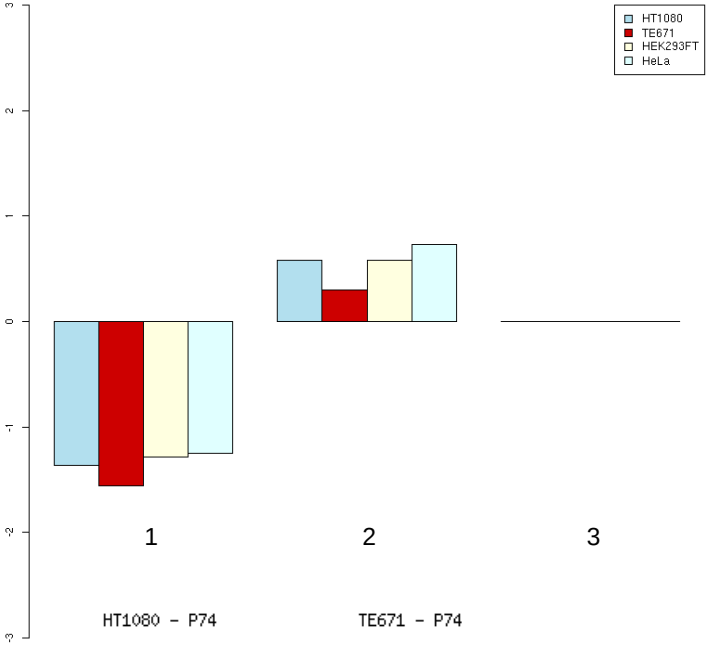
HEK293 – P72



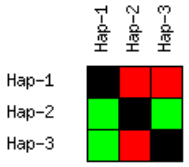
HeLa – P72



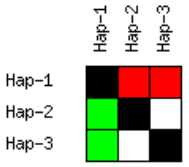
P74 – PHF5A



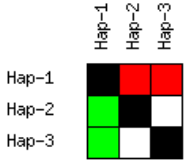
HT1080 – P74



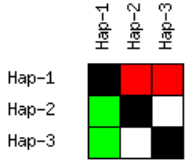
TE671 – P74



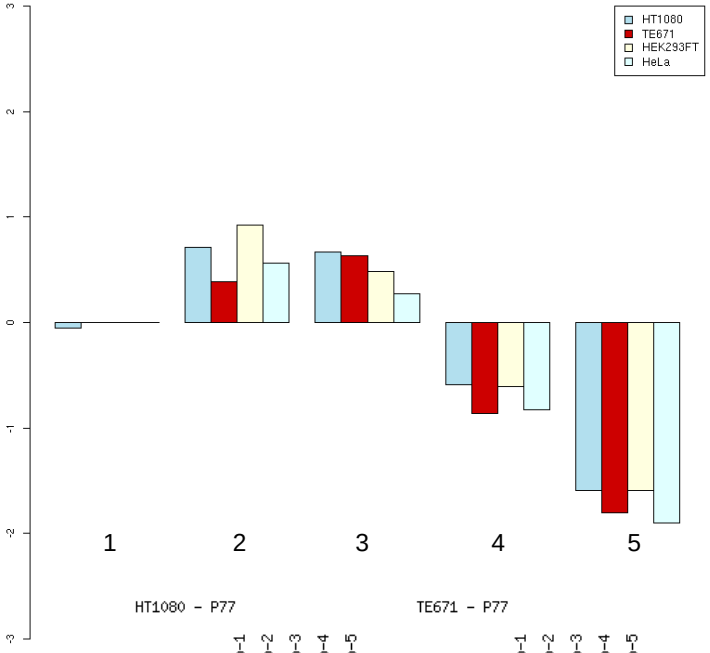
HEK293 – P74



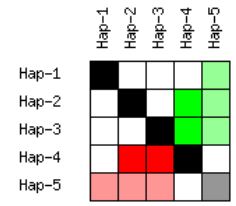
HeLa – P74



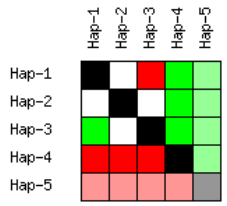
P77 – OTTHUM00000030087



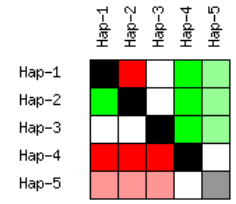
HT1080 – P77



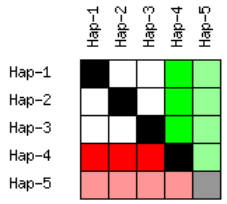
TE671 – P77



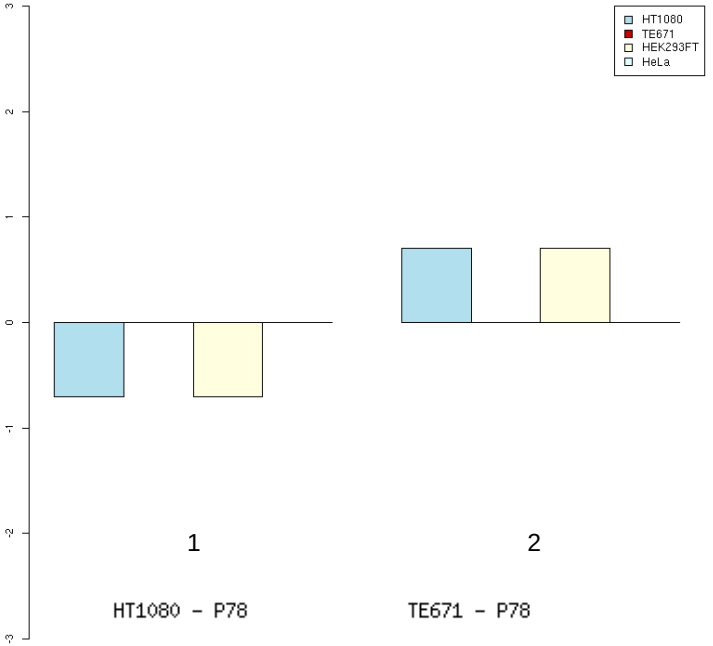
HEK293 – P77



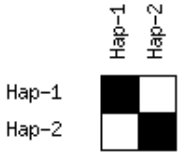
HeLa – P77



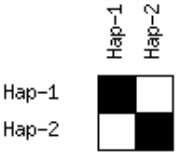
P78 – SREBF2



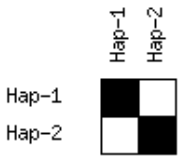
HT1080 – P78



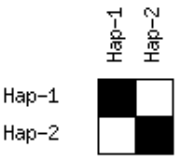
TE671 – P78



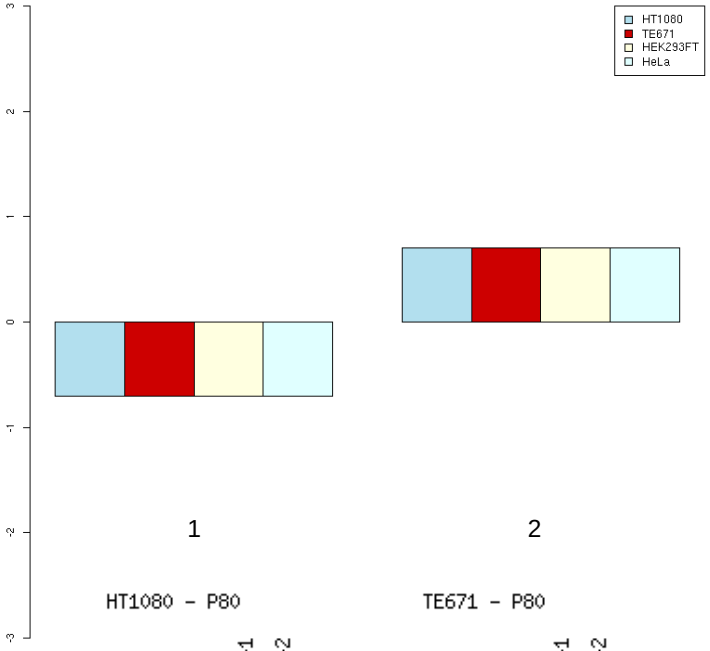
HEK293 – P78



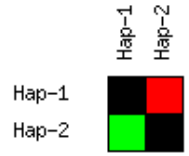
HeLa – P78



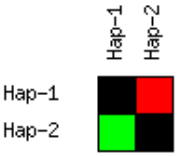
P80 – OTTHUM00000030498



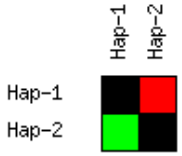
HT1080 – P80



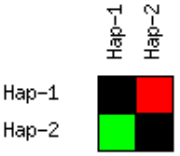
TE671 – P80



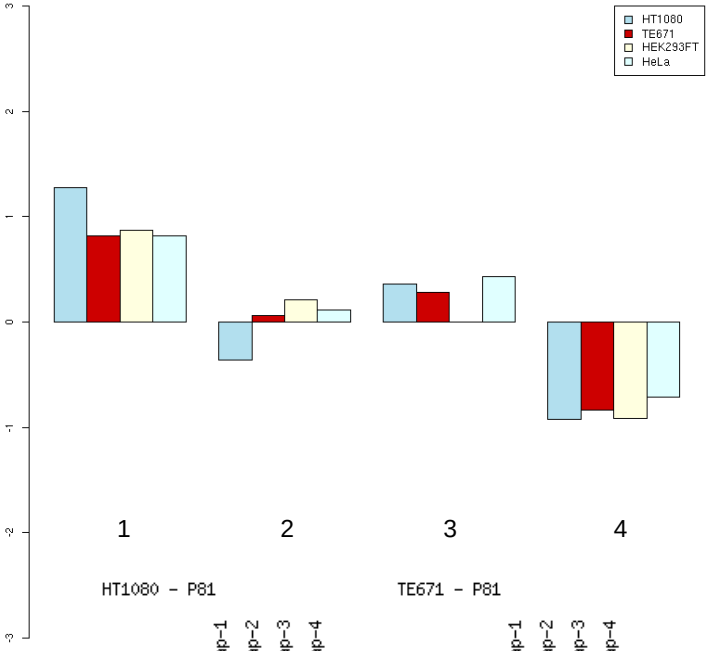
HEK293 – P80



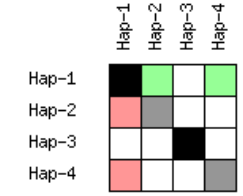
HeLa – P80



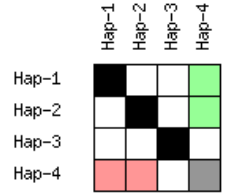
P81 – NAGA



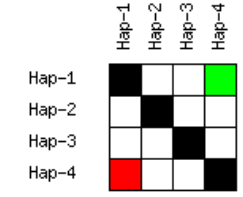
HT1080 – P81



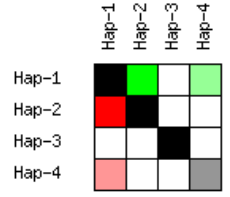
TE671 – P81



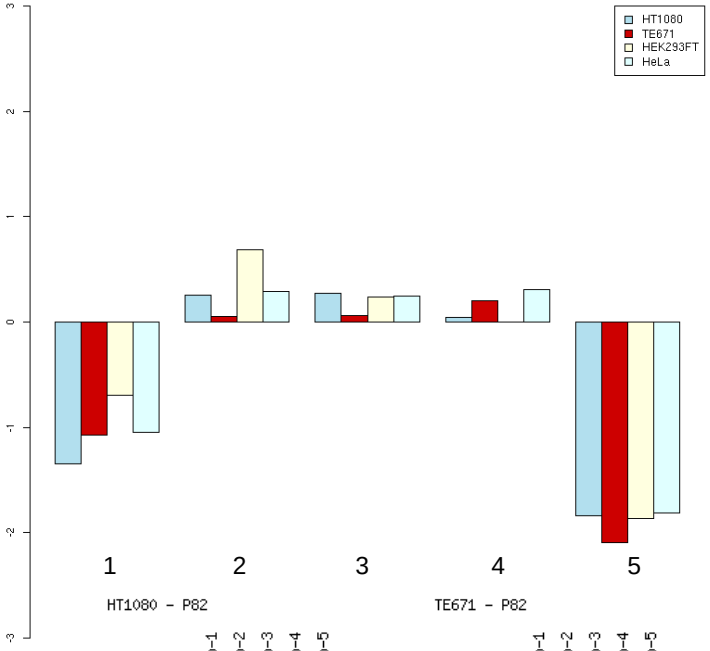
HEK293 – P81



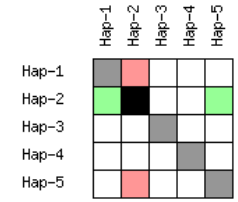
HeLa – P81



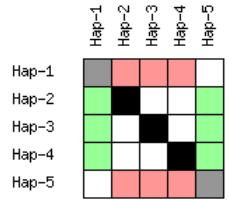
P82 – OTTHUM00000030175



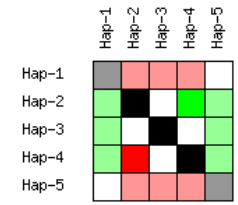
HT1080 – P82



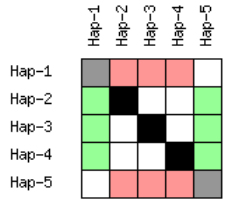
TE671 – P82



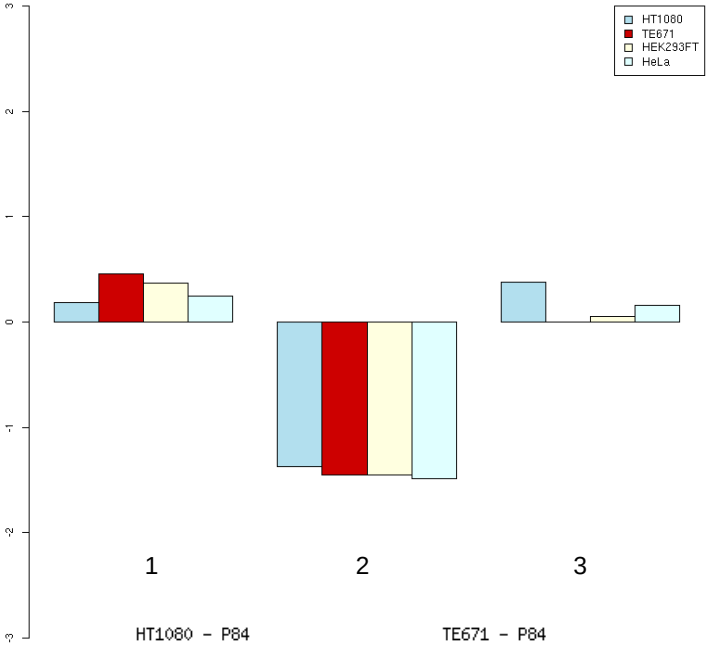
HEK293 – P82



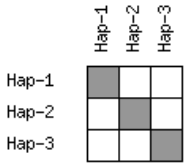
HeLa – P82



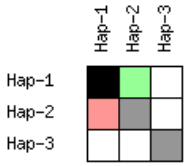
P84 – SERHL



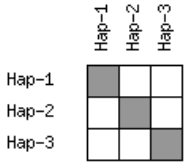
HT1080 – P84



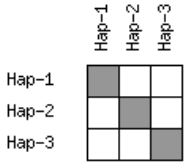
TE671 – P84



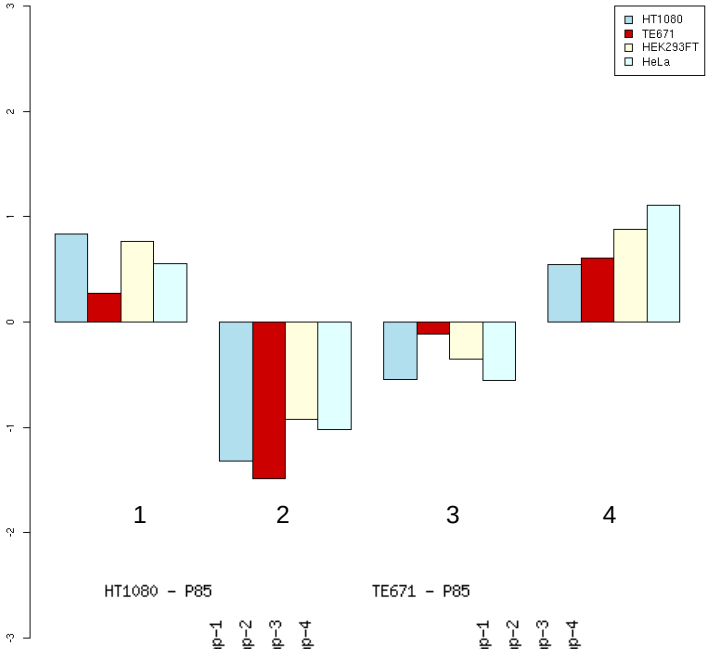
HEK293 – P84

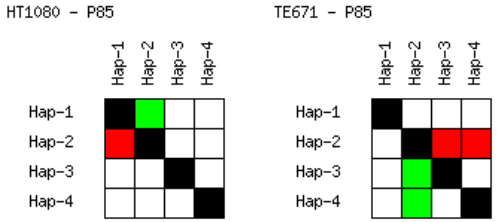


HeLa – P84

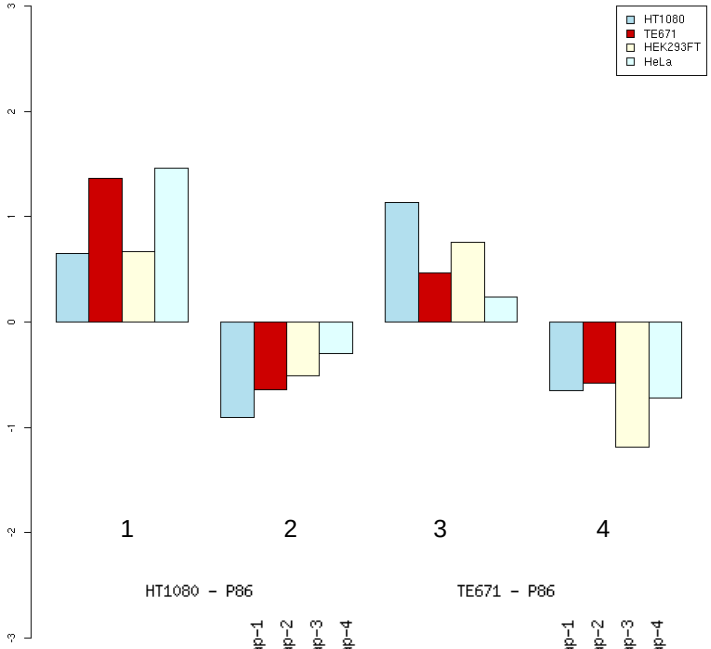


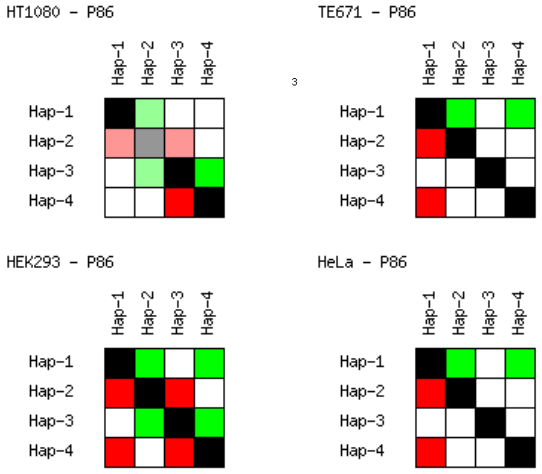
P85 – POLDIP3



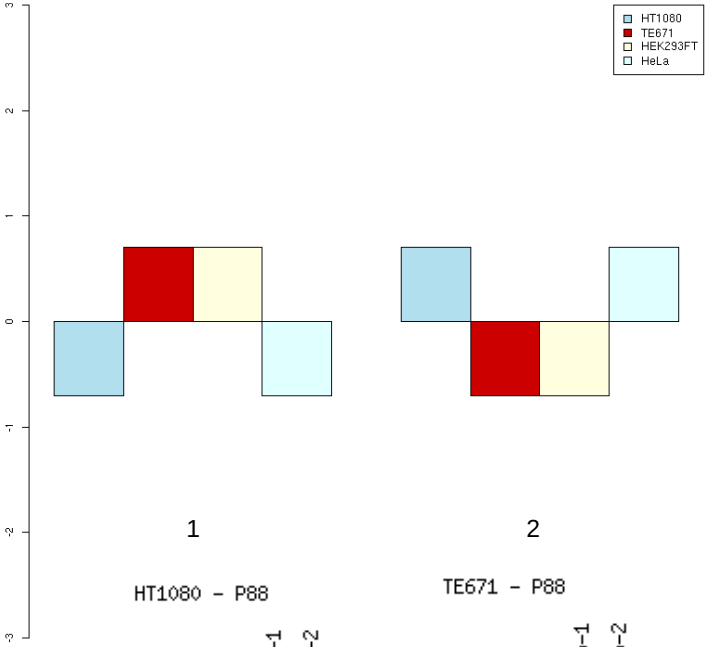


P86 – OTTHUM00000030962

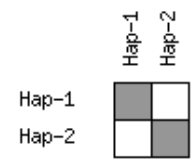




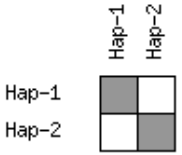
P88 – MPPED1



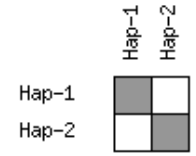
HT1080 – P88



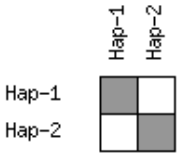
TE671 – P88



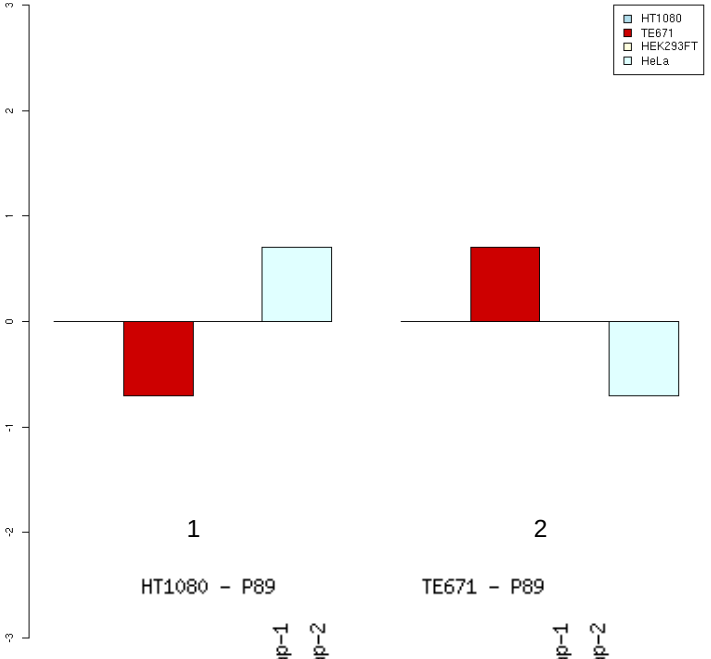
HEK293 – P88



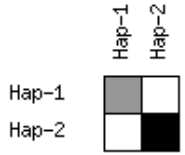
HeLa – P88



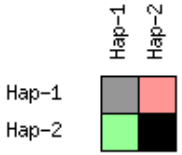
P89 – PNPLA5



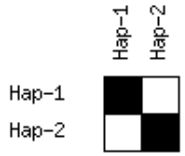
HT1080 – P89



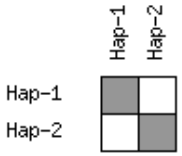
TE671 – P89



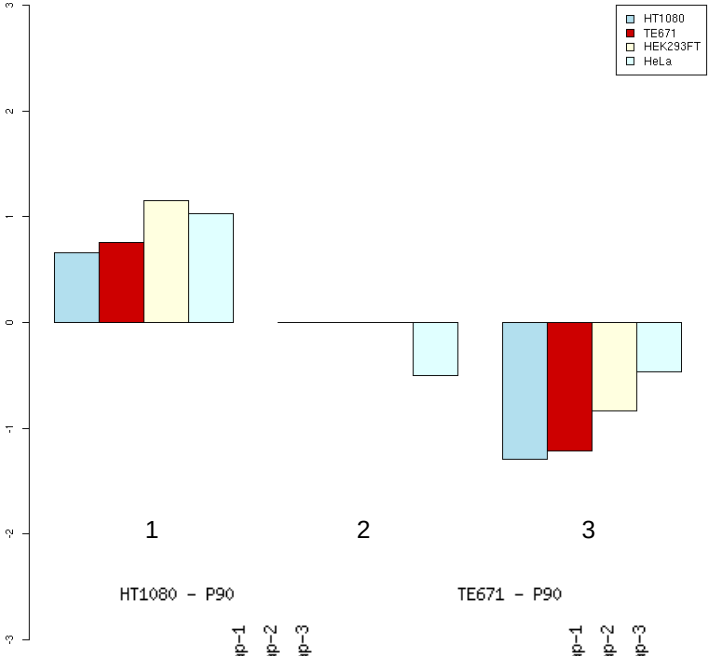
HEK293 – P89



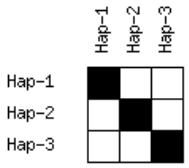
HeLa – P89



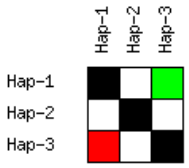
P90 – SAMM50



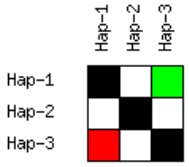
HT1080 – P90



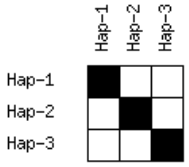
TE671 – P90



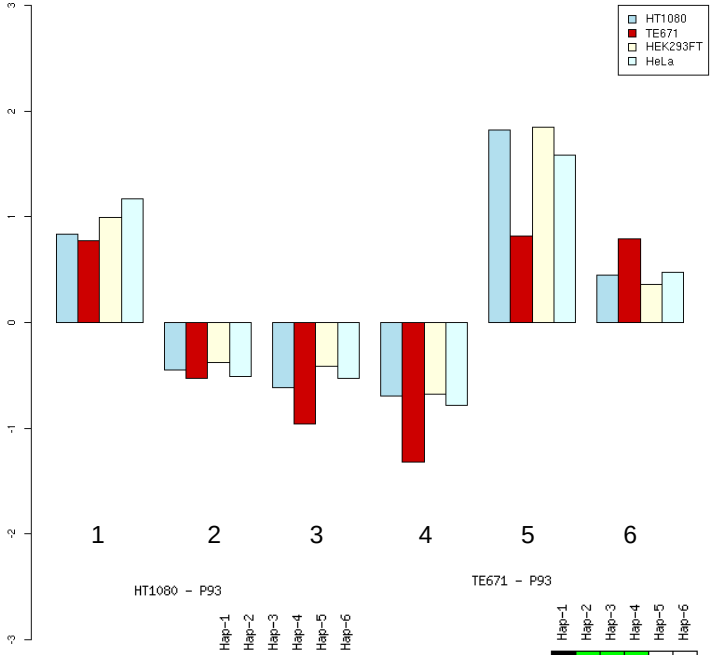
HEK293 – P90



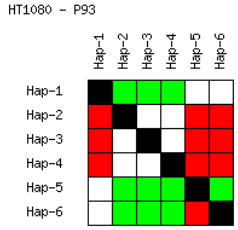
HeLa – P90



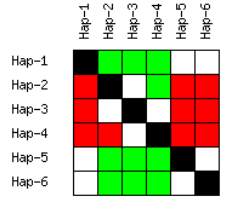
P93 – NUP50



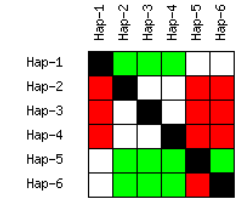
HT1080 – P93



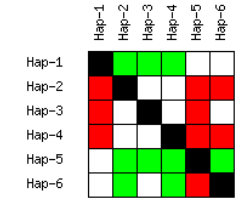
TE671 – P93



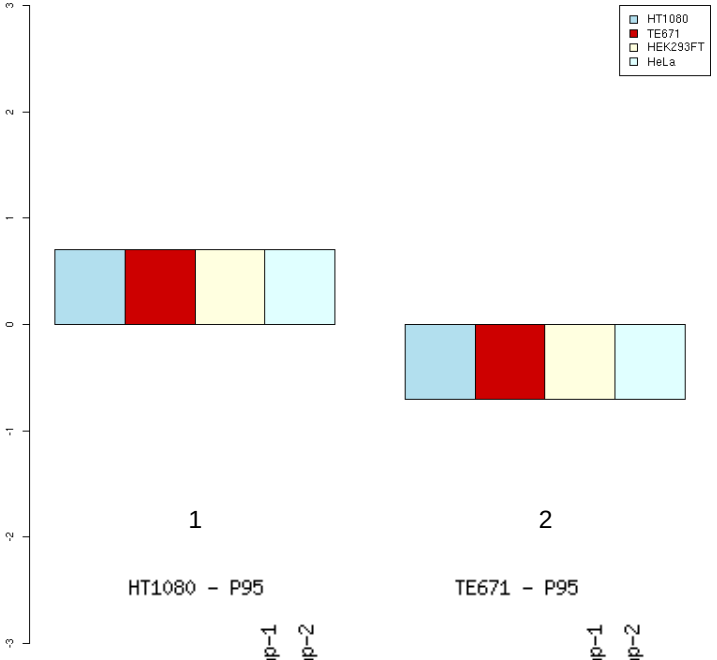
HEK293 – P93



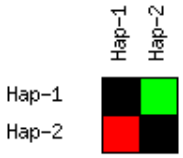
HeLa – P93



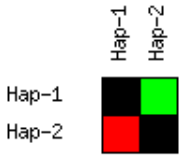
P95 – C22orf8



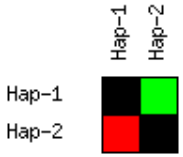
HT1080 – P95



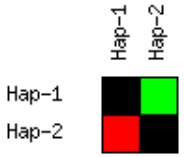
TE671 – P95



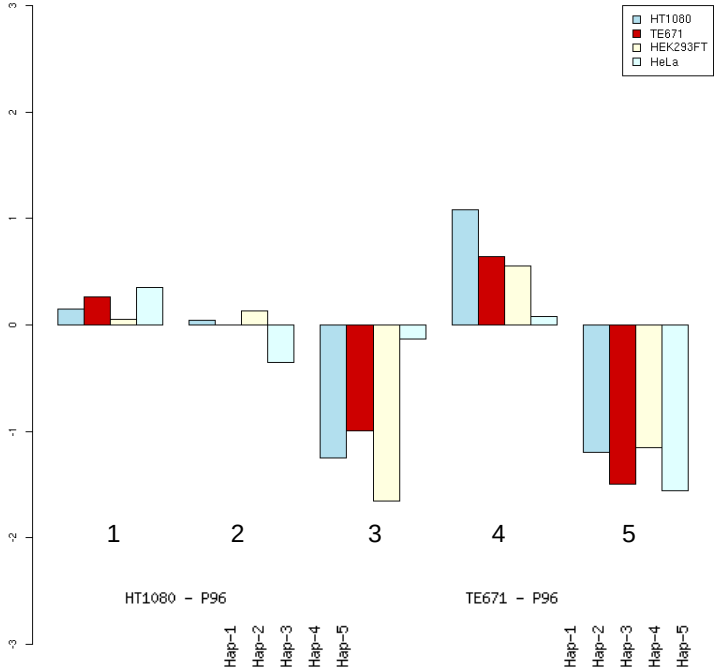
HEK293 – P95



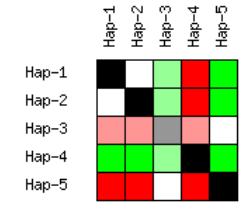
HeLa – P95



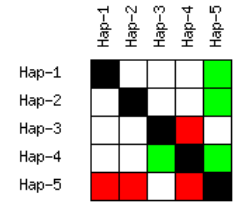
P96 – RIBC2



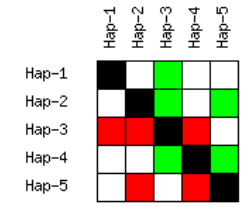
HT1080 – P96



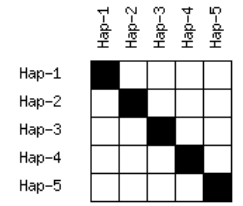
TE671 – P96



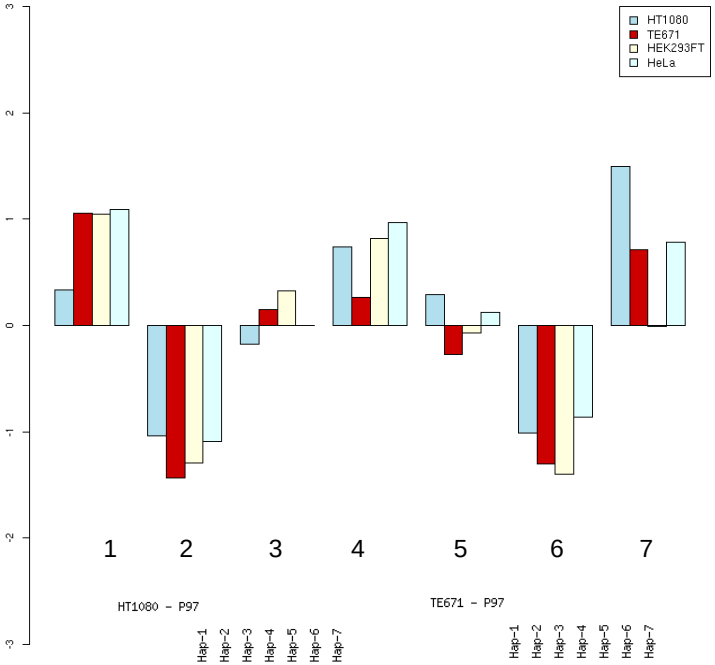
HEK293 – P96

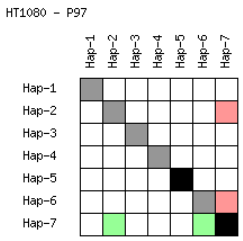


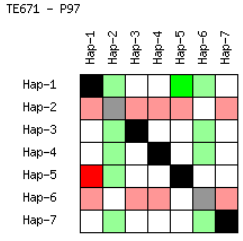
HeLa – P96

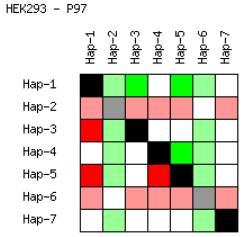


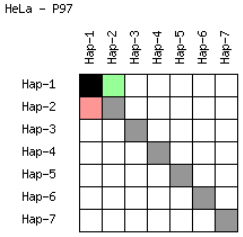
P97 – SMC1L2



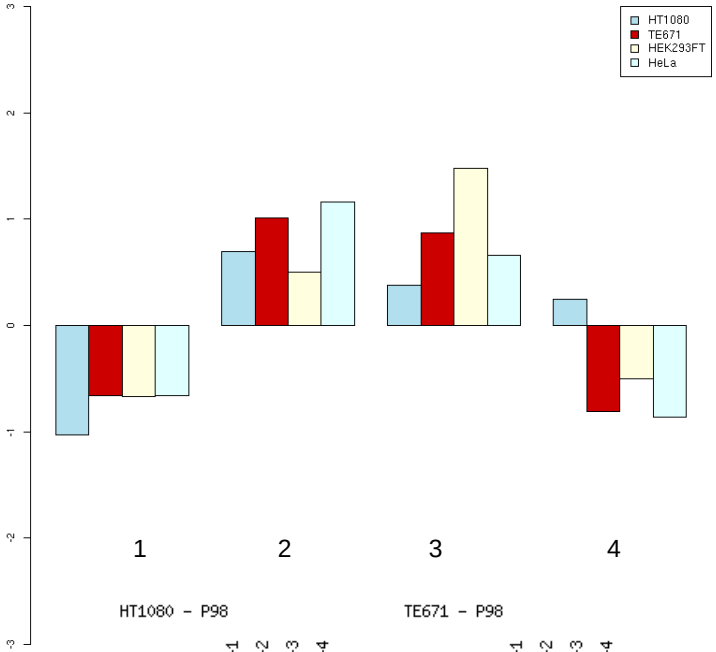


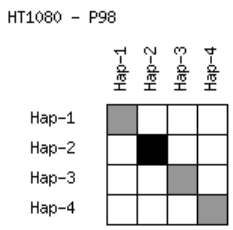


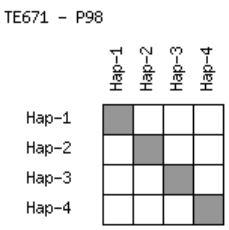


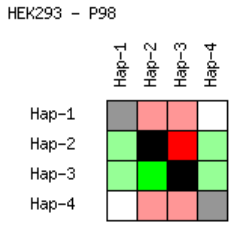


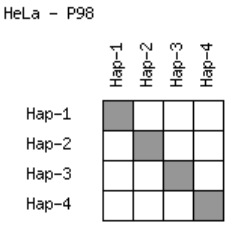
P98 – OTTHUM00000030109



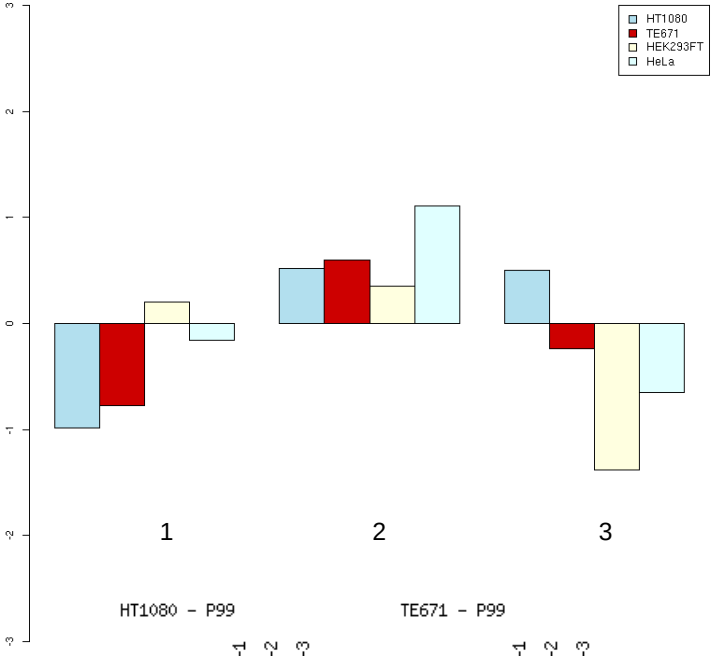




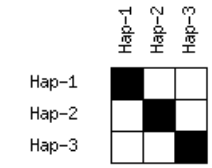




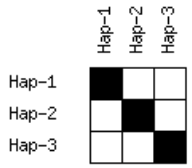
P99 – OTTHUM00000030672



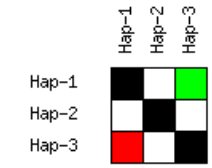
HT1080 – P99



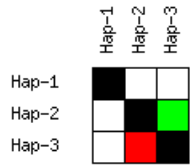
TE671 – P99



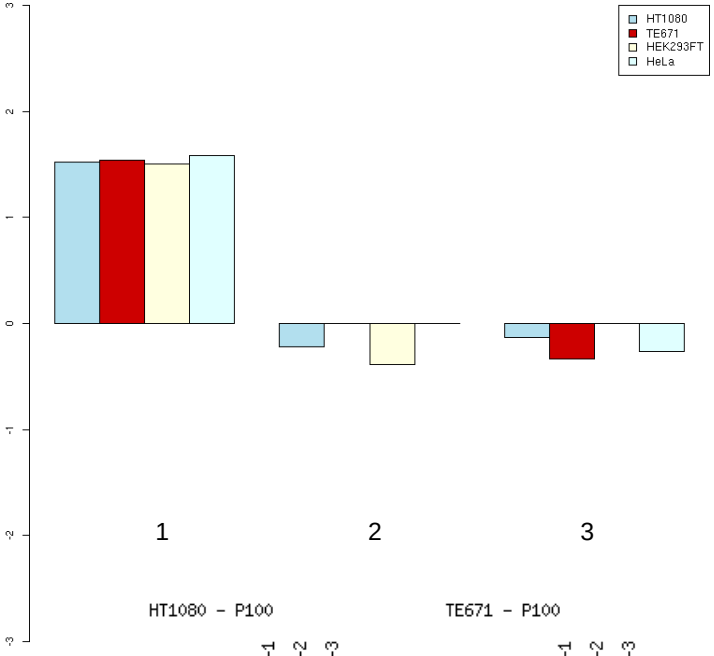
HEK293 – P99



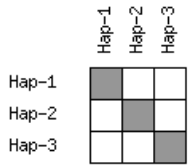
HeLa – P99



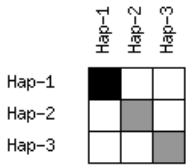
P100 – PKDREJ



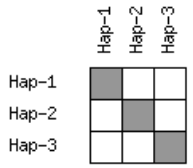
HT1080 – P100



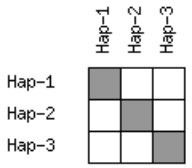
TE671 – P100



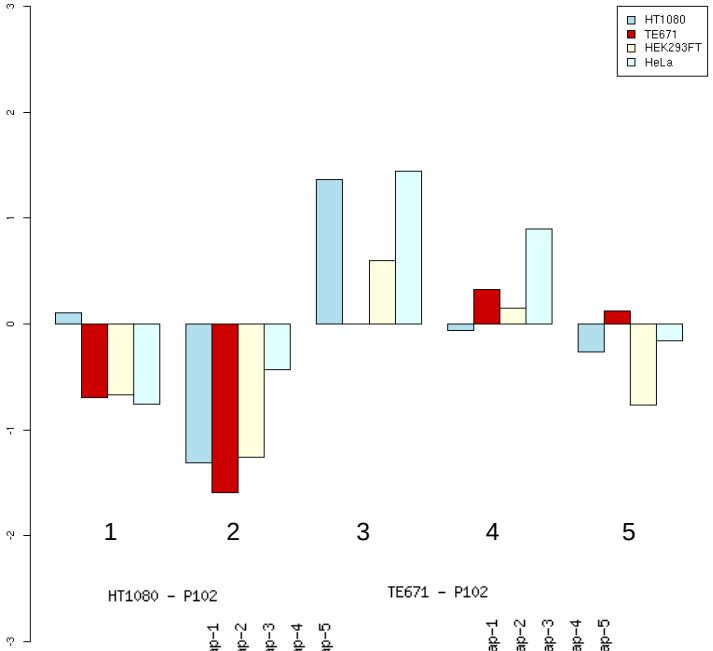
HEK293 – P100

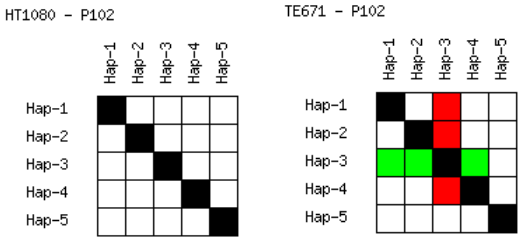


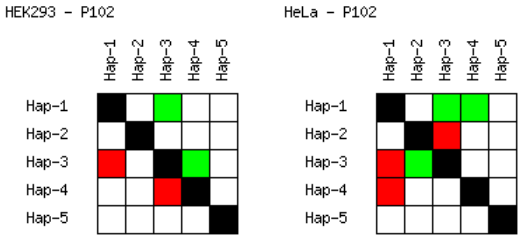
HeLa – P100



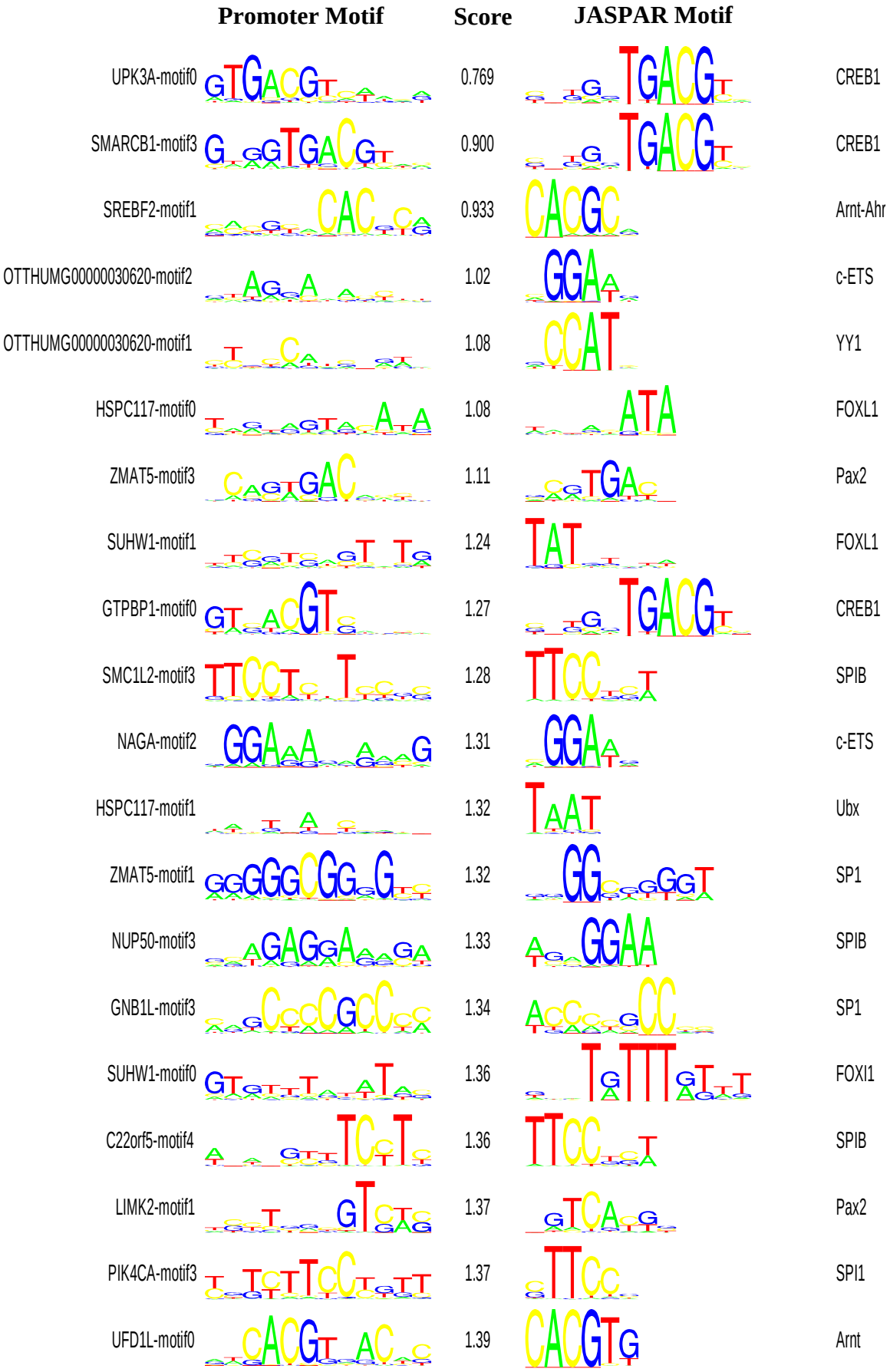
P102 – TBC1D22A







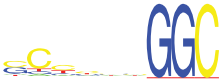
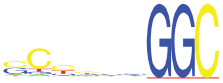
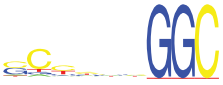
Appendix F – *De novo* generated motifs matching JASPAR



Promoter Motif		Score	JASPAR Motif	
PPM1F-motif3		1.39		ABI4
OTTHUMG00000030620-motif0		1.41		En1
ZNF74-motif3		1.41		SPI1
NCF4-motif2		1.43		Klf4
TBX1-motif2		1.44		SPI1
FBXO7-motif4		1.44		SPIB
OTTHUMG00000030620-motif3		1.45		RUSH1-alfa
MFNG-motif4		1.47		SPI1
HSPC117-motif4		1.49		FOXL1
NEFH-motif0		1.50		Arnt-Ahr
NAGA-motif4		1.50		Pax2
OTTHUMG00000030620-motif4		1.51		c-ETS
PGEA1-motif3		1.51		SPI1
RR22_HUMAN-motif3		1.51		MNB1A
RR22_HUMAN-motif2		1.53		c-ETS
OTTHUMG00000030175-motif4		1.53		c-ETS
VPREB1-motif2		1.54		SP1
PSCD4-motif1		1.55		Klf4
COMT-motif0		1.56		SP1
CSF2RB-motif0		1.56		HMG-1
MN1-motif1		1.56		Myf

Promoter Motif		Score	JASPAR Motif	
OTTHUMG00000030194-motif3		1.56		Myf
PSCD4-motif0		1.58		c-ETS
SAMM50-motif1		1.58		MafB
COMT-motif1		1.59		SPI1
OTTHUMG00000030194-motif1		1.61		c-ETS
SUHW1-motif4		1.61		RUSH1-alfa
PSCD4-motif2		1.62		NHLH1
OTTHUMG00000030175-motif0		1.62		NHLH1
SUHW1-motif2		1.63		FOXL1
AP1B1-motif1		1.63		ZNF42_1-4
NUP50-motif4		1.65		FOXL1
PNPLA5-motif3		1.66		ABI4
SAMM50-motif3		1.66		ID1
CRYBB3-motif4		1.67		Pax2
CLDN5-motif0		1.67		RUSH1-alfa
ZMAT5-motif2		1.68		c-ETS
UBE2L3-motif1		1.69		ABI4
OTTHUMG00000030194-motif4		1.71		SP1
OTTHUMG00000030384-motif3		1.71		Arnt-Ahr
CSF2RB-motif4		1.72		ZNF42_1-4
HORMAD2-motif0		1.73		c-ETS

Promoter Motif		Score	JASPAR Motif	
CSF2RB-motif2		1.73		ZNF42_5-13
GNB1L-motif4		1.73		SPI1
NCF4-motif4		1.74		RUSH1-alfa
SRR1L-motif4		1.74		RUSH1-alfa
MN1-motif4		1.74		c-ETS
OTTHUMG00000030205-motif0		1.75		ABI4
PCQAP-motif4		1.75		c-ETS
COMT-motif2		1.78		ABI4
LIMK2-motif2		1.78		SP1
C22orf8-motif2		1.78		SPIB
OTTHUMG00000030257-motif1		1.79		HAND1-TCF3
PPM1F-motif4		1.80		ZNF42_1-4
APOBEC3B-motif2		1.81		ABI4
AP1B1-motif0		1.82		Macho-1
PARVG-motif3		1.82		SP1
UFD1L-motif1		1.82		ABI4
LIMK2-motif4		1.83		c-ETS
GTPBP1-motif3		1.83		ABI4
OTTHUMG00000030172-motif1		1.83		MafB
NUP50-motif2		1.83		SP1

Promoter Motif		Score	JASPAR Motif	
OTTHUMG00000030087-motif2		1.84		IRF1
NCF4-motif3		1.85		Pax2
GTPBP1-motif2		1.85		RUSH1-alfa
TBC1D22A-motif0		1.86		SP1
GNB1L-motif0		1.87		MafB
GALR3-motif3		1.88		Arnt-Ahr
RR22_HUMAN-motif0		1.89		HMG-1
OTTHUMG00000030194-motif0		1.89		ABI4
CSF2RB-motif3		1.90		Pax2
NIPSNAP1-motif1		1.91		Myf
HMG2L1-motif4		1.91		TFAP2A
PCQAP-motif2		1.91		ABI4
GNB1L-motif1		1.92		c-ETS
AP1B1-motif4		1.92		RUSH1-alfa
GALR3-motif2		1.92		TFAP2A
OTTHUMG00000030087-motif1		1.92		TFAP2A
OTTHUMG00000030205-motif3		1.92		Klf4
C22orf5-motif2		1.93		TFAP2A
DGCR14-motif1		1.93		RUSH1-alfa
DEPDC5-motif2		1.94		c-ETS
C22orf8-motif3		1.94		Pax2

Promoter Motif		Score	JASPAR Motif	
PARVG-motif4		1.95		ABI4
SUHW1-motif3		1.95		c-ETS
OTTHUMG00000030143-motif1		1.95		ABI4
RR22_HUMAN-motif1		1.95		Myf
SMARCB1-motif4		1.96		NHLH1
PNPLA5-motif1		1.97		Klf4
PIK4CA-motif4		1.98		RUSH1-alfa
PHF5A-motif3		1.98		SPI1
RIBC2-motif1		2.00		SP1