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.

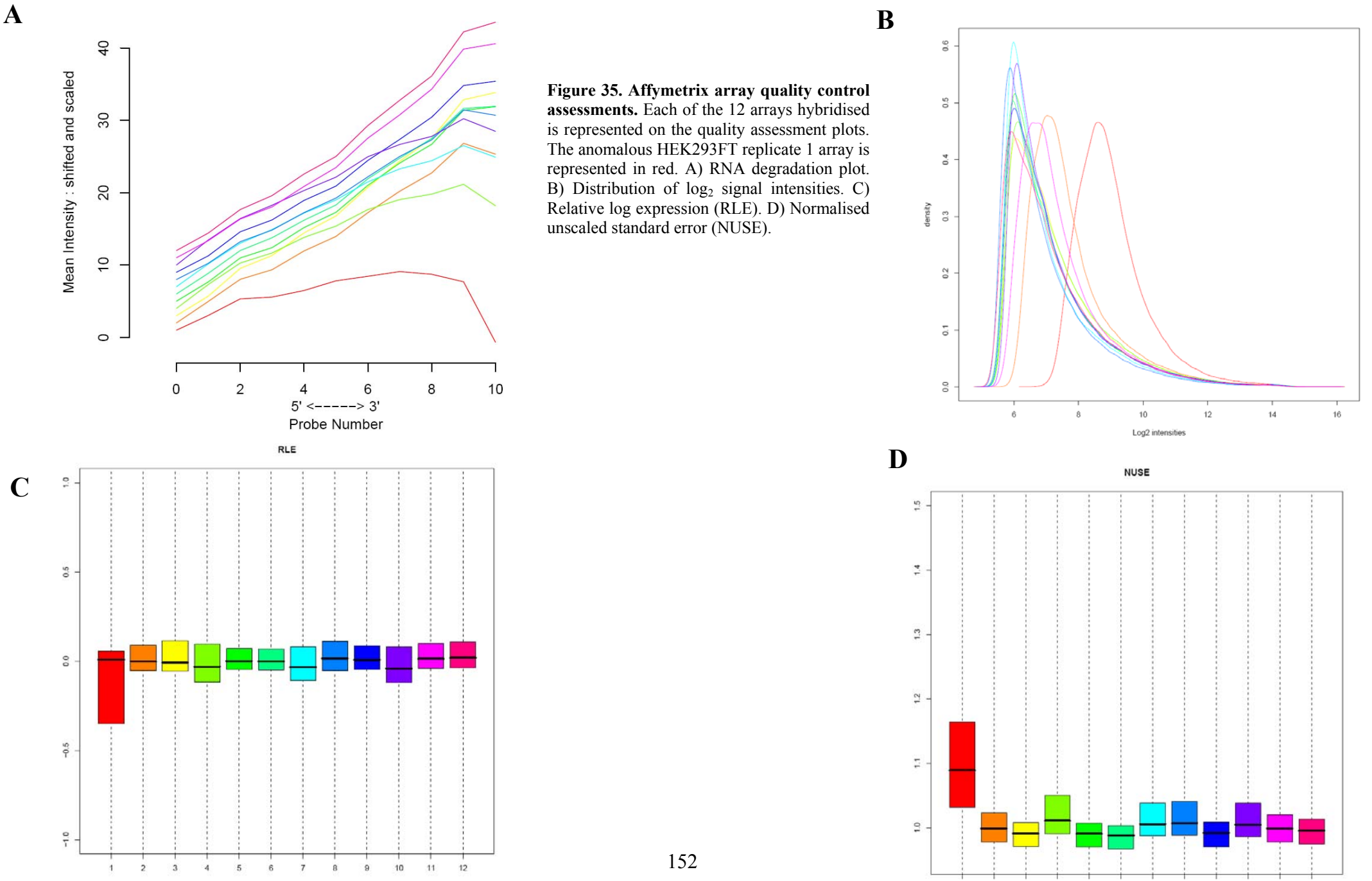


Figure 35. Affymetrix array quality control assessments. Each of the 12 arrays hybridised is represented on the quality assessment plots. The anomalous HEK293FT replicate 1 array is represented in red. A) RNA degradation plot. B) Distribution of $\log_2$ signal intensities. C) Relative log expression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

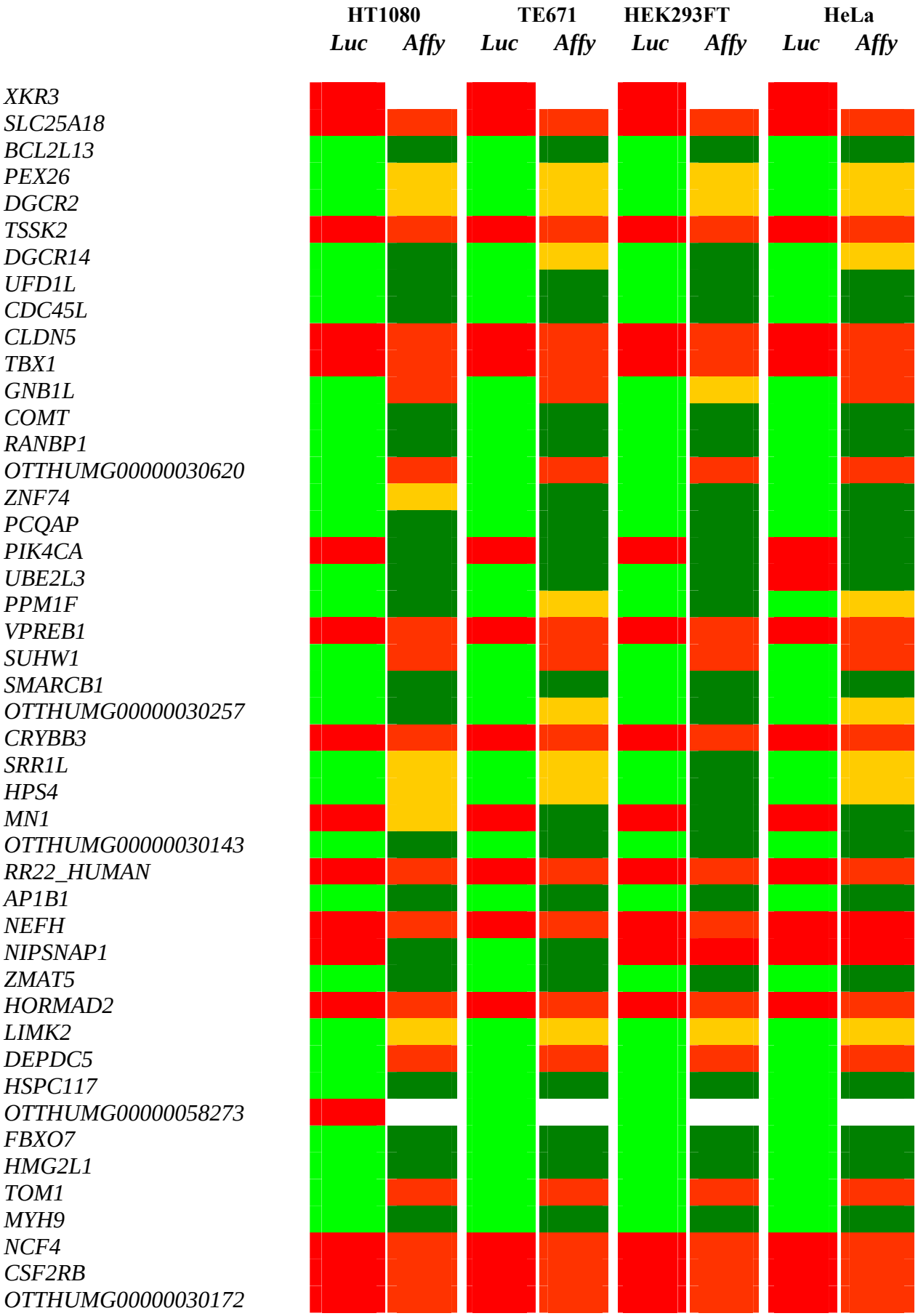
5.2.4 Comparison of endogenous gene expression with cloned promoter activity

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

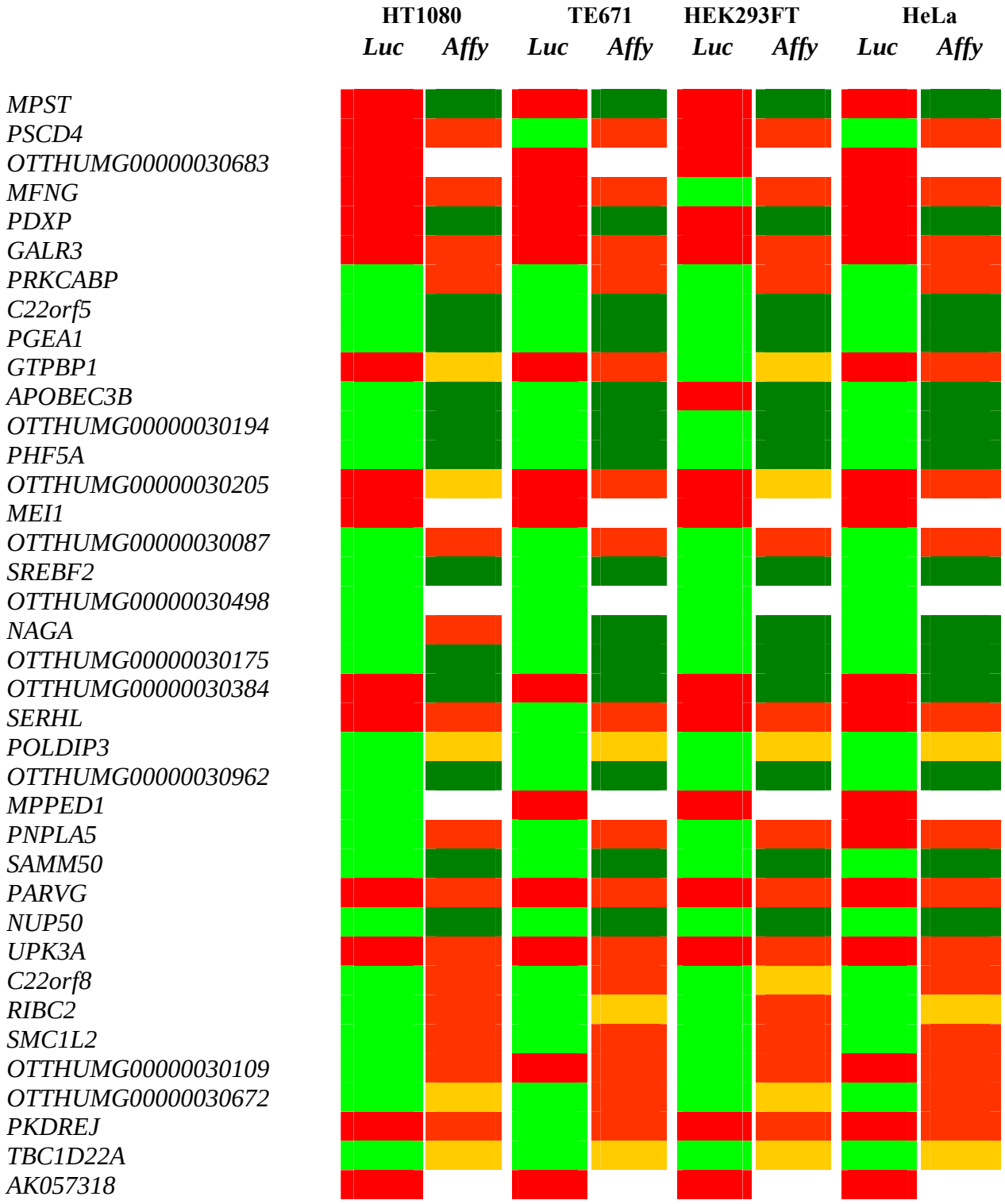


Table 12. Concordance of promoter activity and gene expression for tested promoters in 4 cell lines. Active and inactive promoters in each cell line are designated by green and red shading respectively. The consensus gene expression call is shown next to the promoter activity information in a slightly 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.

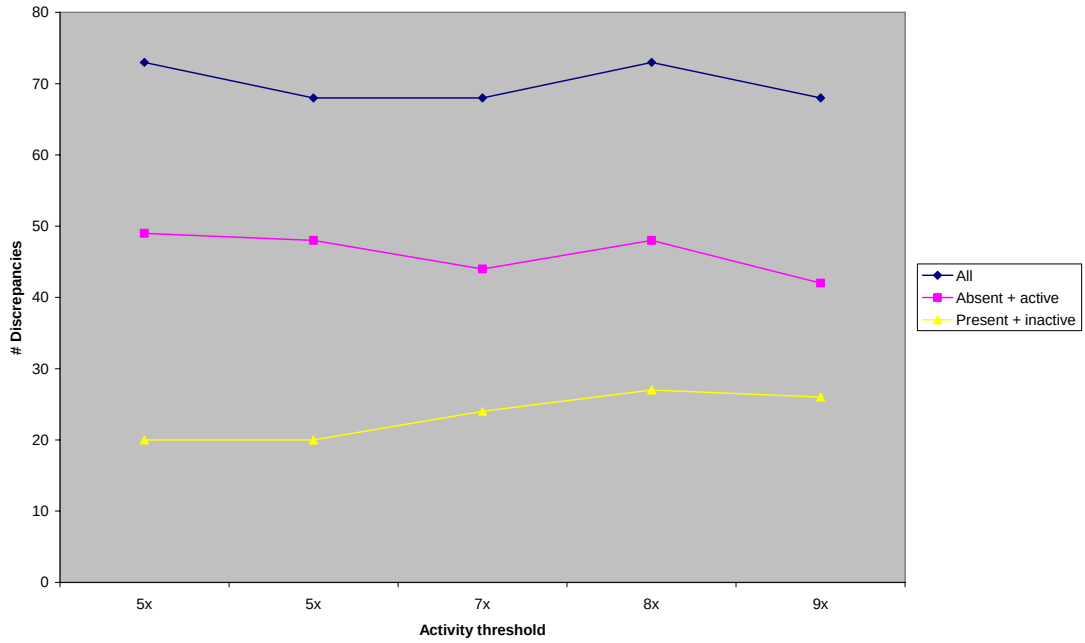


Figure 36. Relationship of promoter activity threshold to the number and type of mismatched calls between the luciferase and expression data. The number of instances where the presence, marginal or absent calls matched what would be expected from the promoter activity (y axis) was examined as a function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

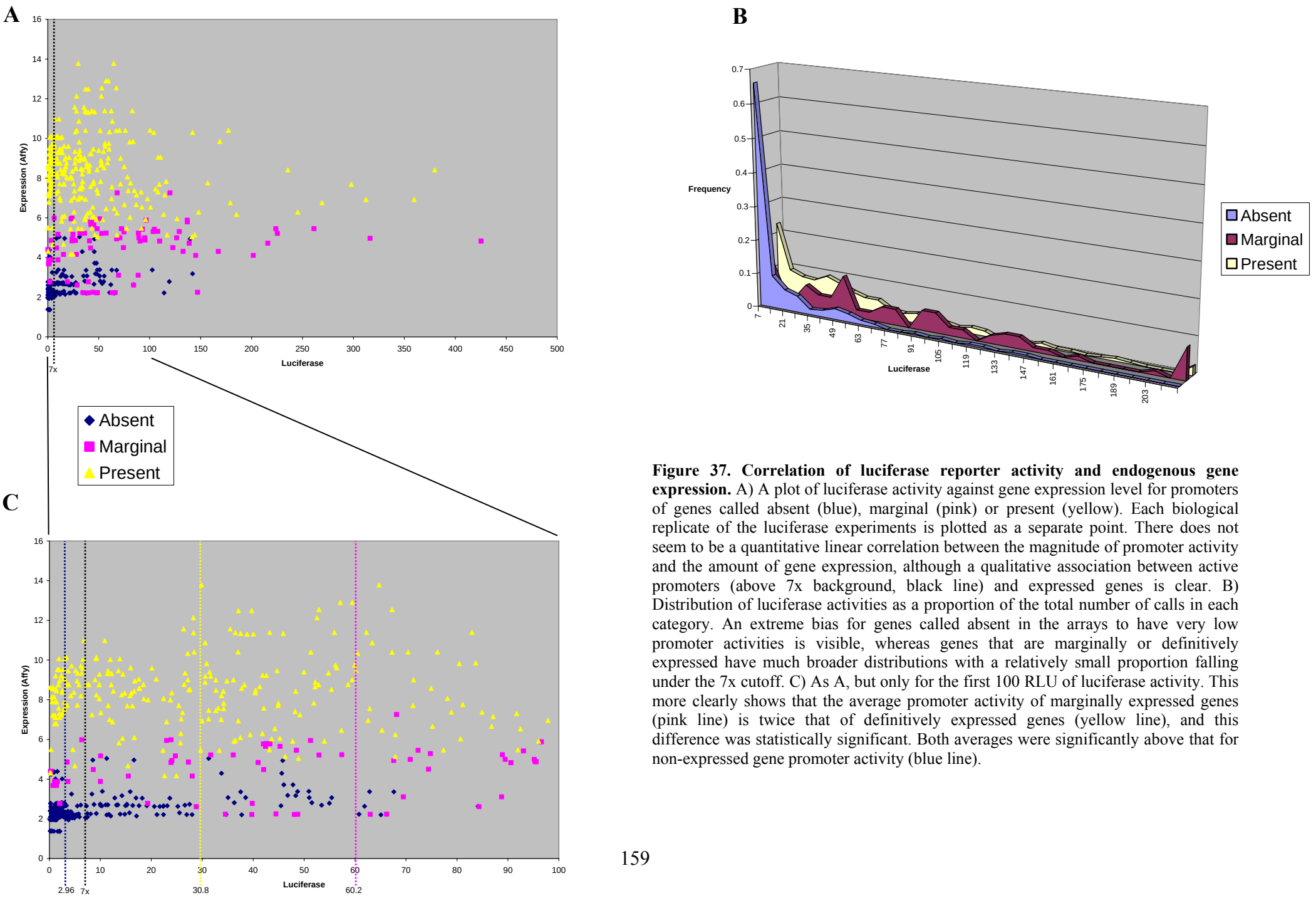


Figure 37. Correlation of luciferase reporter activity and endogenous gene expression. A) A plot of luciferase activity against gene expression level for promoters of genes called absent (blue), marginal (pink) or present (yellow). Each biological replicate of the luciferase experiments is plotted as 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.

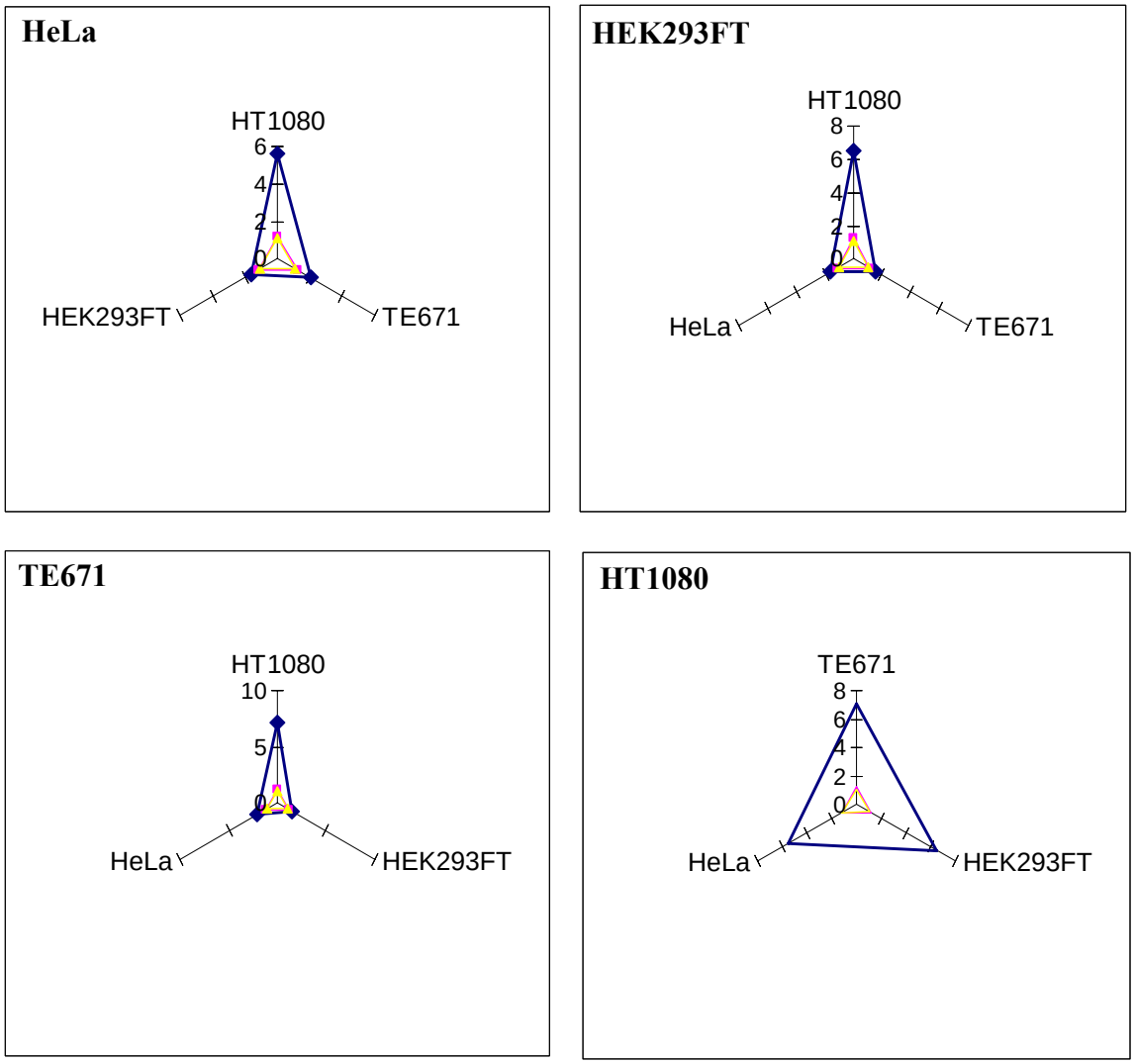


Figure 38. Distances between the 4 cell lines according to the overall activity / expression profiles of cloned promoter constructs (blue), transcription factors (pink) and cloned promoter genes (yellow). Each of the four panels compares one cell line (in bold) to the three others, showing how close 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.

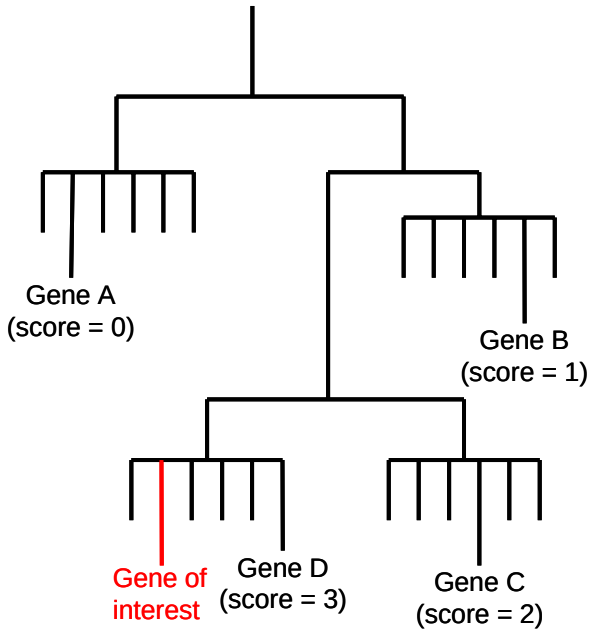


Figure 39. Simplified tree showing the scoring system used to identify co-regulated genes. For a particular gene of interest, the co-expression of each other gene on the array is calculated as the number of partitions on the tree for which the two genes segregate together. 4 genes are highlighted on 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).

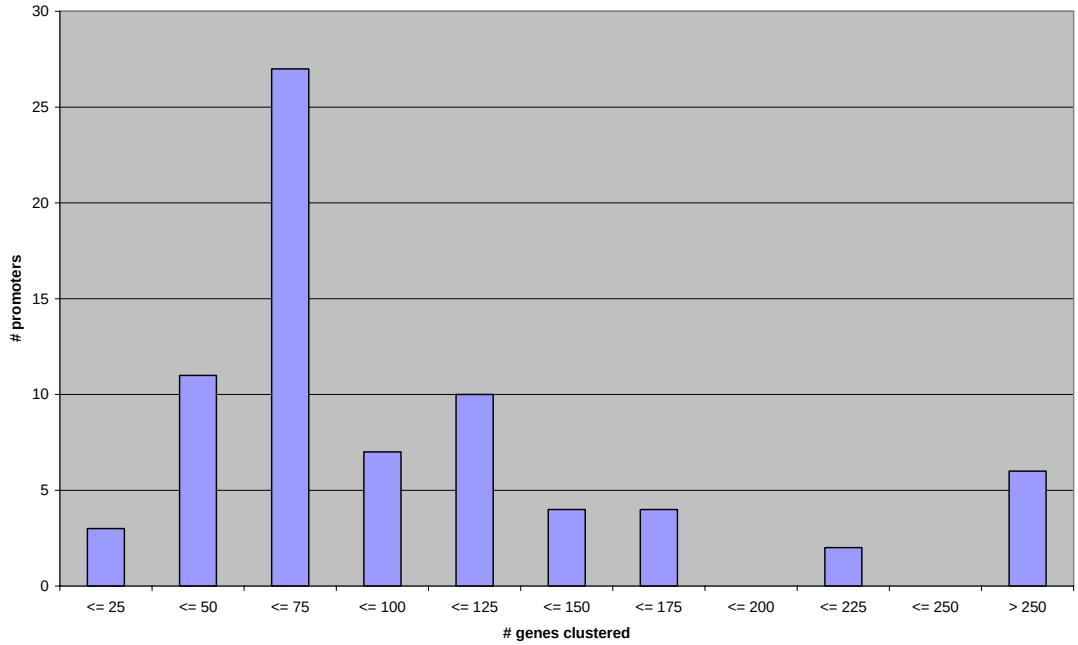


Figure 40. Number of genes clustered with the cloned promoters. The majority of cloned promoter genes clustered with between 50 and 125 other genes.

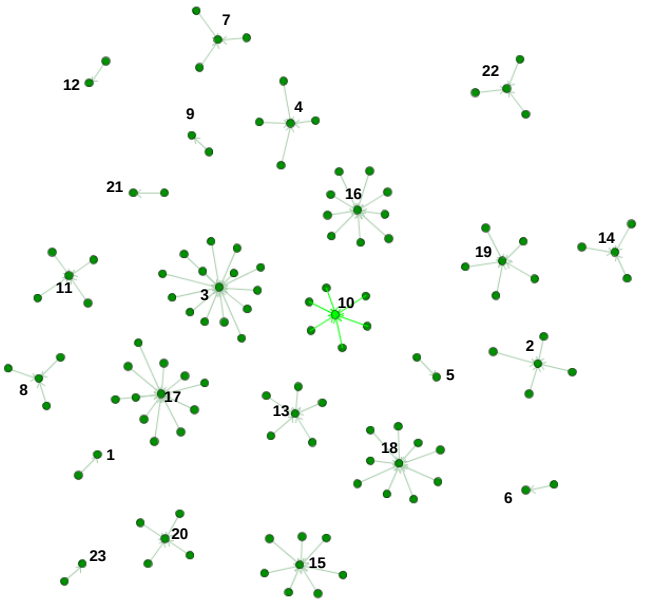
The sequences between 600 base pairs upstream and 100 base pairs downstream of the TSS's of the genes in each cluster were extracted from Ensembl using BioMart. The program nestedMICA (Down and Hubbard 2005) was used to look for motifs within each cluster separately. nestedMICA functions by analysing 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

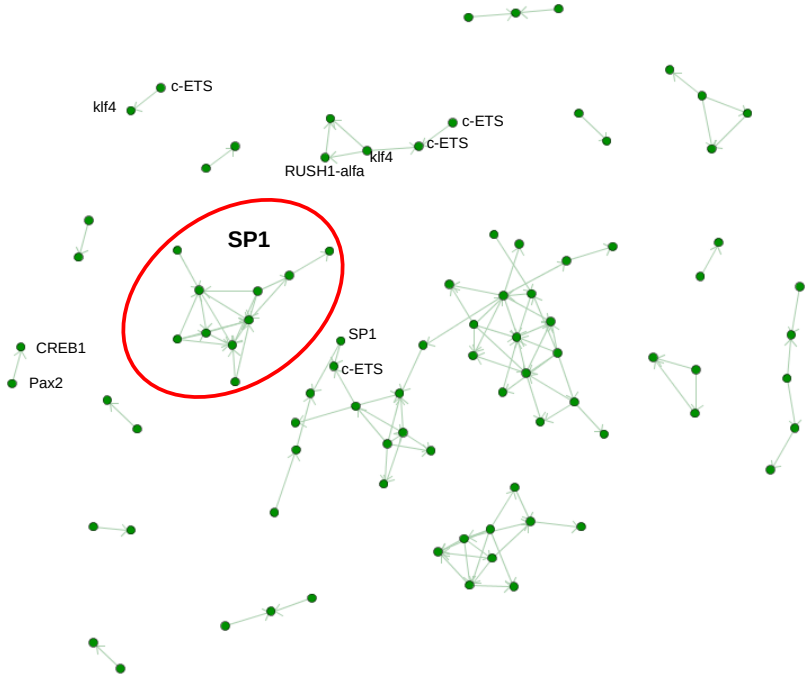


Figure 41. Comparative analysis of the motifs discovered in the clusters of co-regulated genes. Motifs are shown as green nodes joined by lines showing similarity matches. A) Motifs matching TFBS weight matrices in JASPAR with a threshold of 2 or under. 23 different weight matrices were matched to at 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.