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.

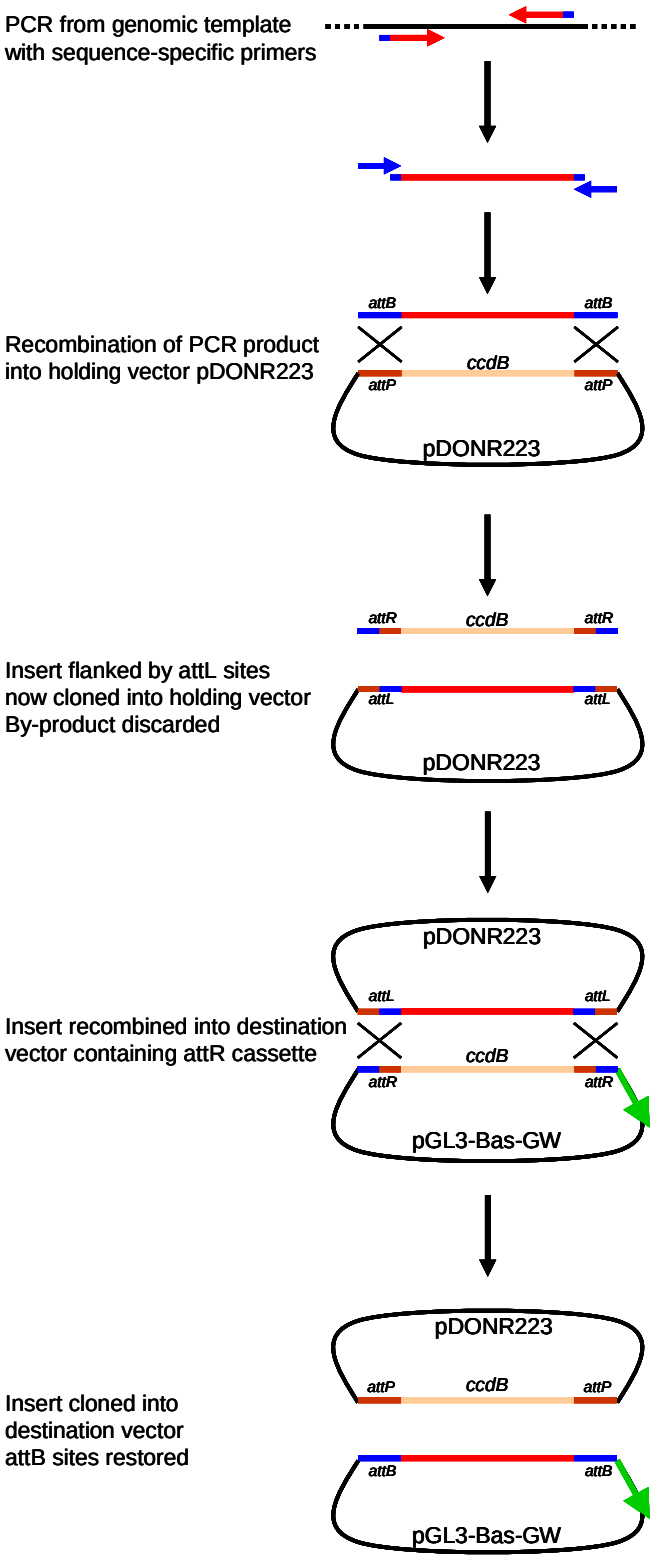


Figure 17. Cloning a PCR fragment using the Gateway cloning technology by Invitrogen.

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.

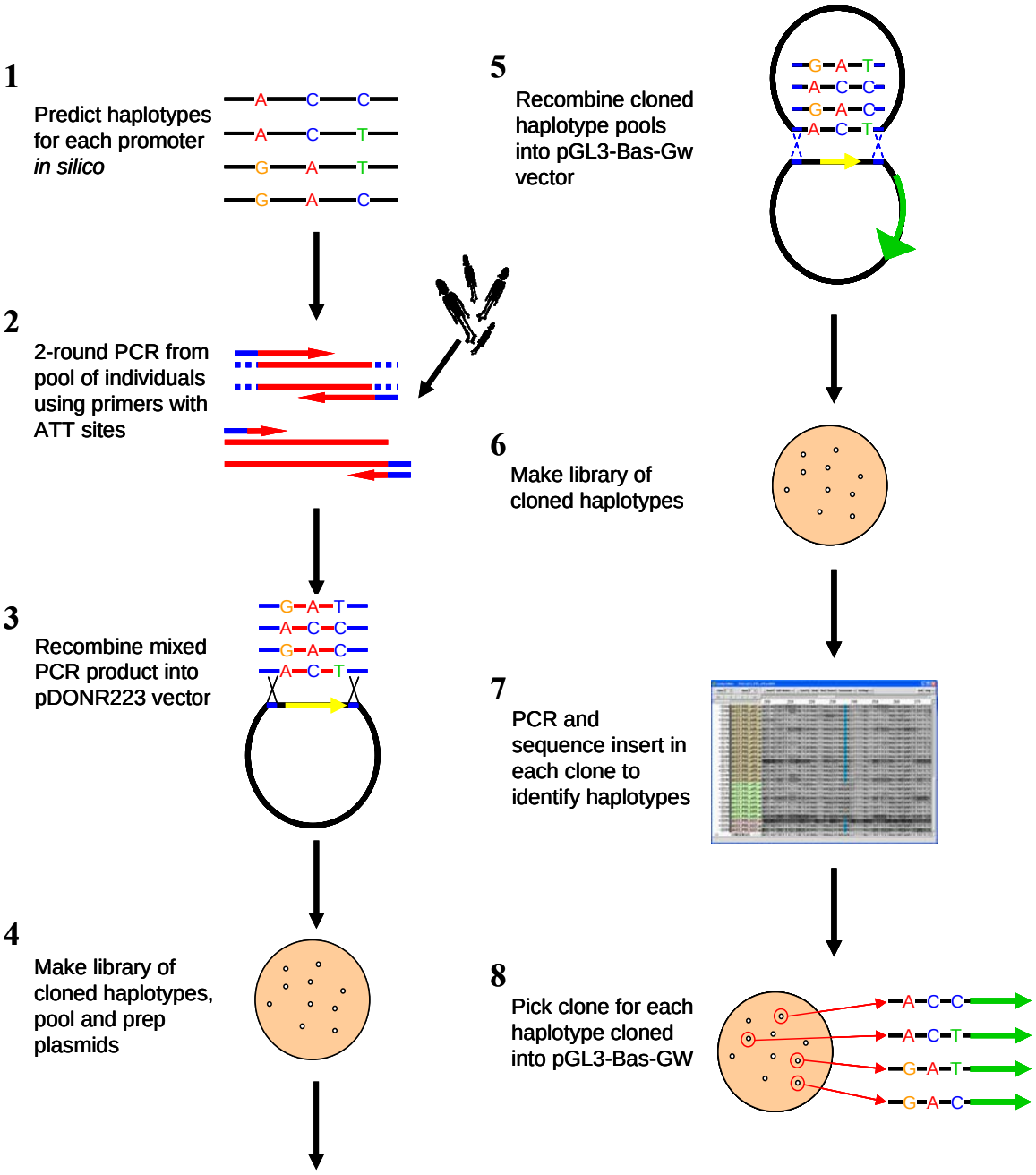


Figure 18. High-throughput strategy for cloning promoter haplotypes into luciferase reporter vectors using Gateway technology.

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.

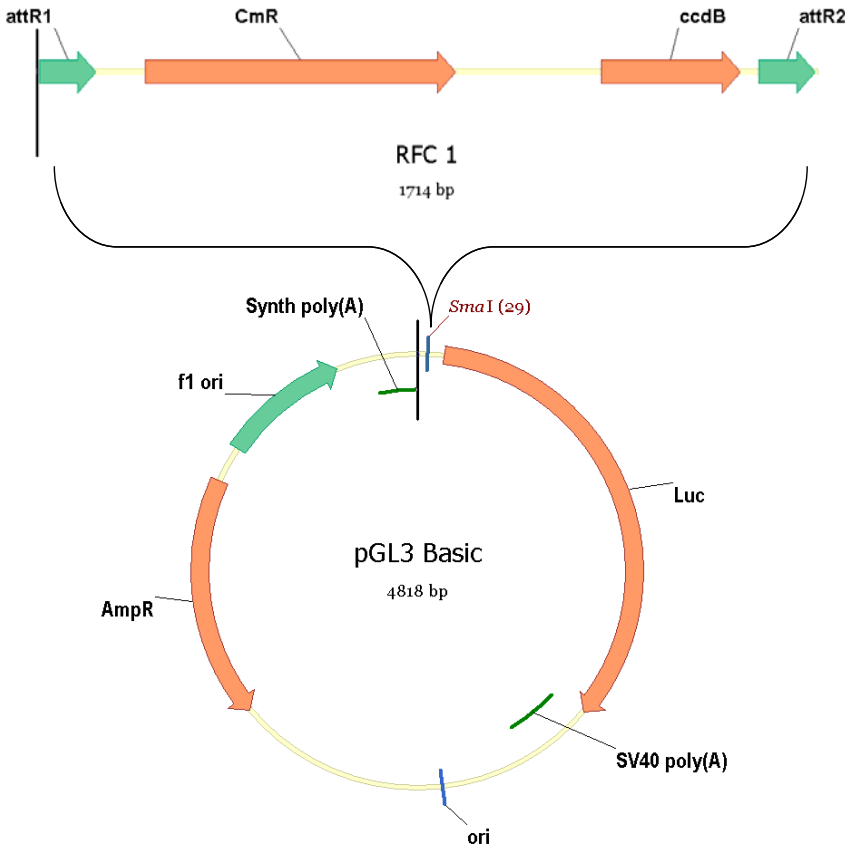


Figure 19. Modification of the pGL3 Basic reporter vector by the insertion of a Gateway acceptor cassette into the multi-cloning site (MCS). The MCS contains a *Sma*I restriction site that leaves blunt ends when digested.

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).

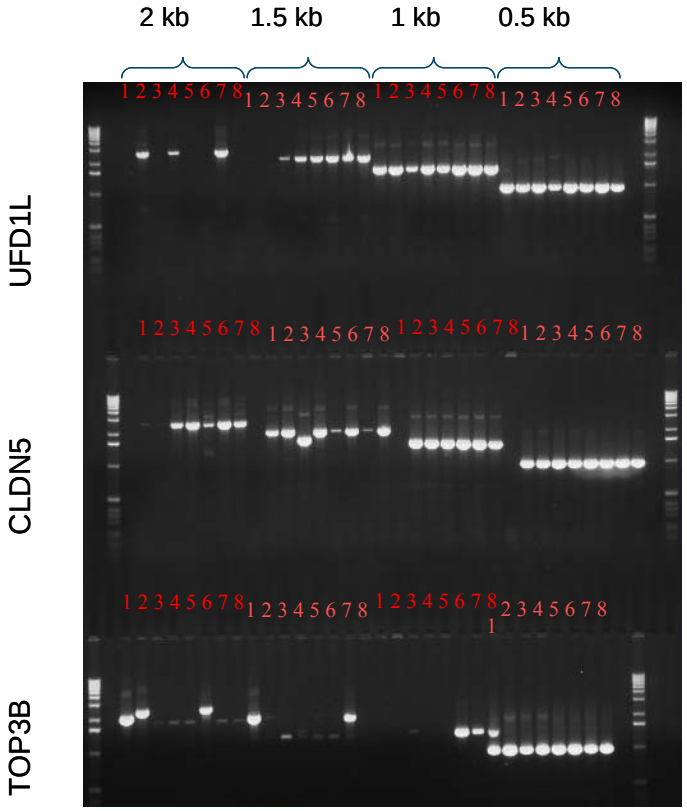


Figure 20. Cloning of four different promoter fragment sizes from the UF1DL, CLDN5 and TOP3B promoters using the high-throughput Gateway method. The fragment sizes tested extended for approximately 2 kb, 1.5 kb, 1 kb and 0.5kb upstream of the annotated TSS. The fragments were amplified for cloning by 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.

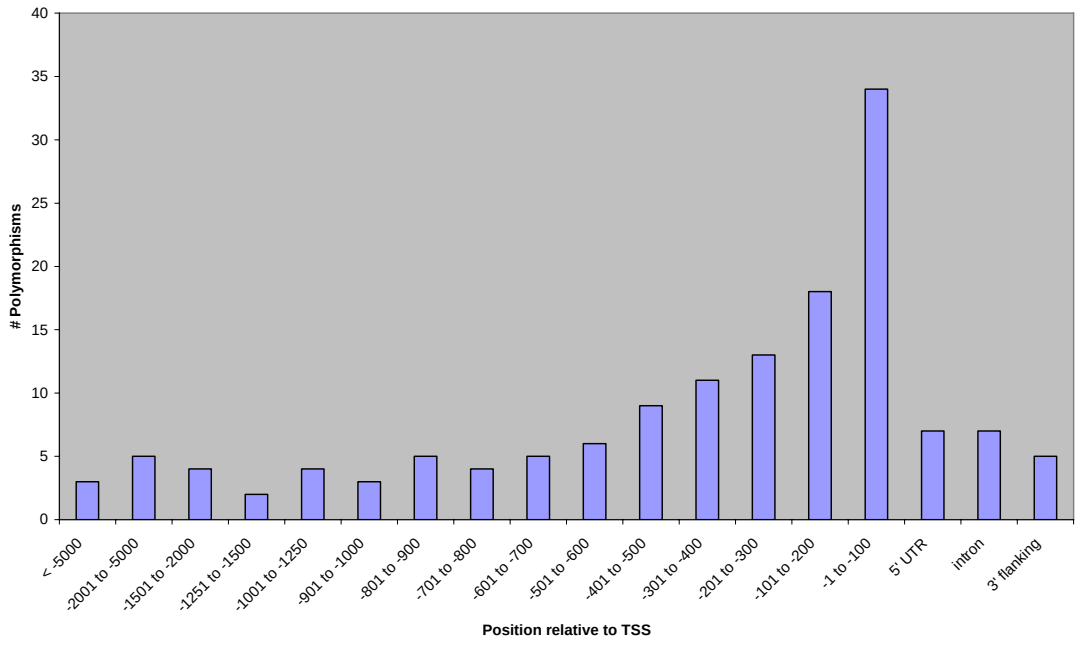


Figure 21. Profile of the numbers of experimentally verified promoter polymorphisms present in the literature. Data were taken from the supplementary material of Rockman and Wray 2002.

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.

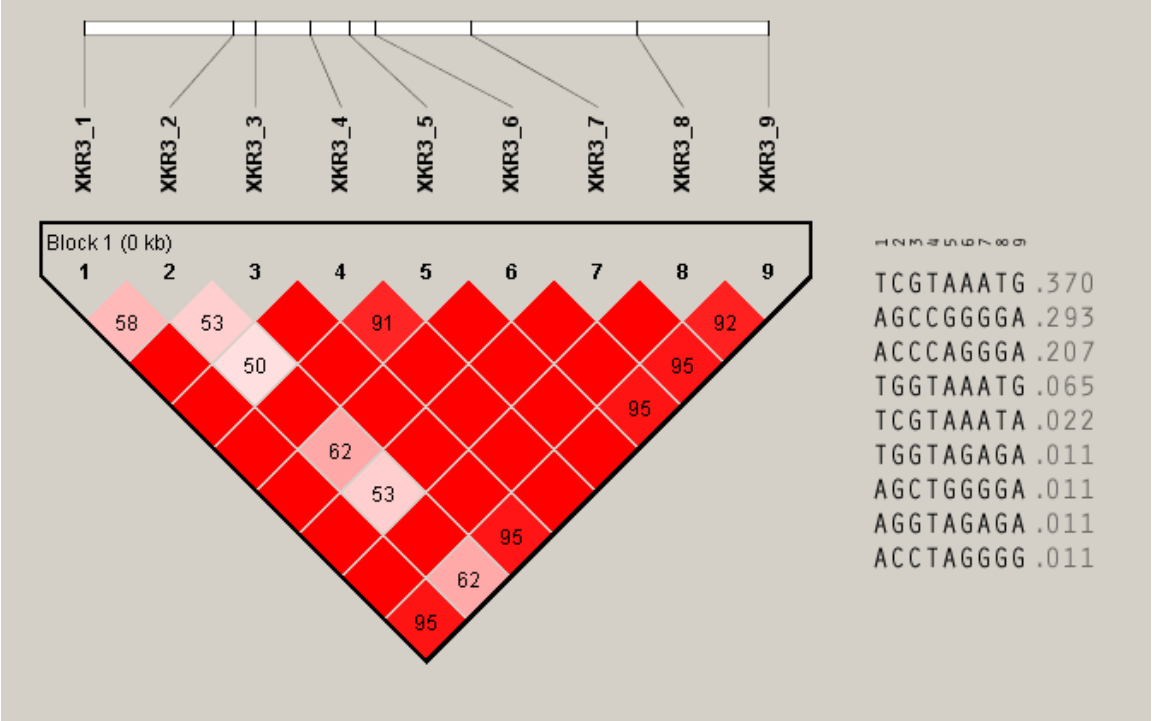


Figure 22. Prediction of haplotypes in the XKR3 promoter using the genotypes produced in the re-sequencing. This was the promoter with the highest number of SNPs in the fragment targeted for cloning, with 9 SNPs and 9 haplotypes. The coloured boxes between each SNP pair are a measure of the degree of 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.

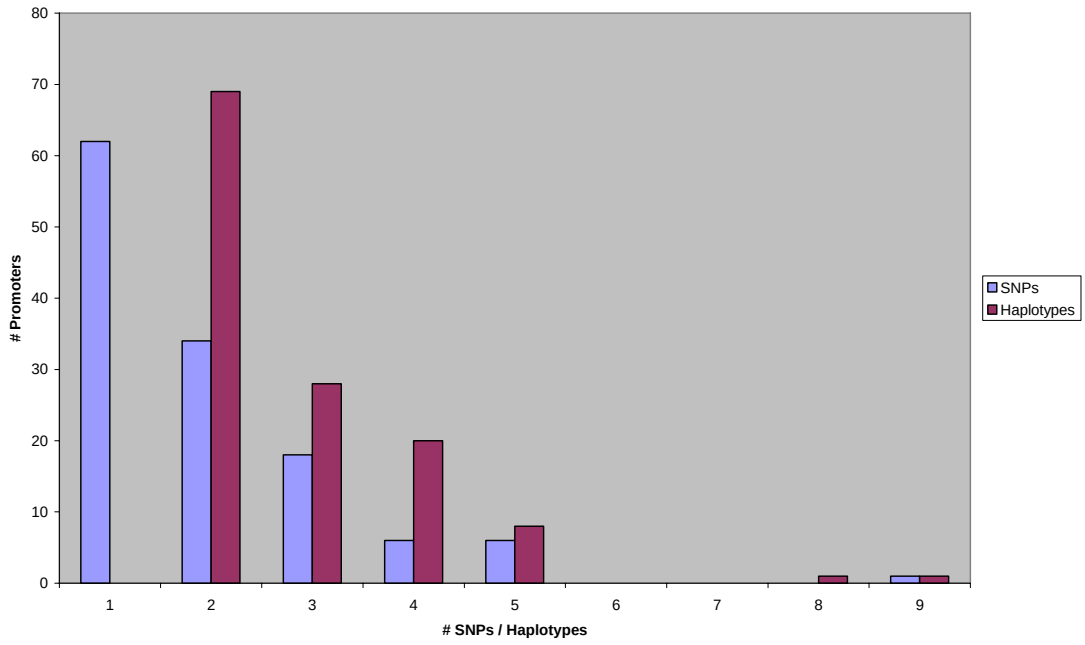


Figure 23. The numbers of SNPs and haplotypes present in the promoter fragments targeted for cloning and functional testing.

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.

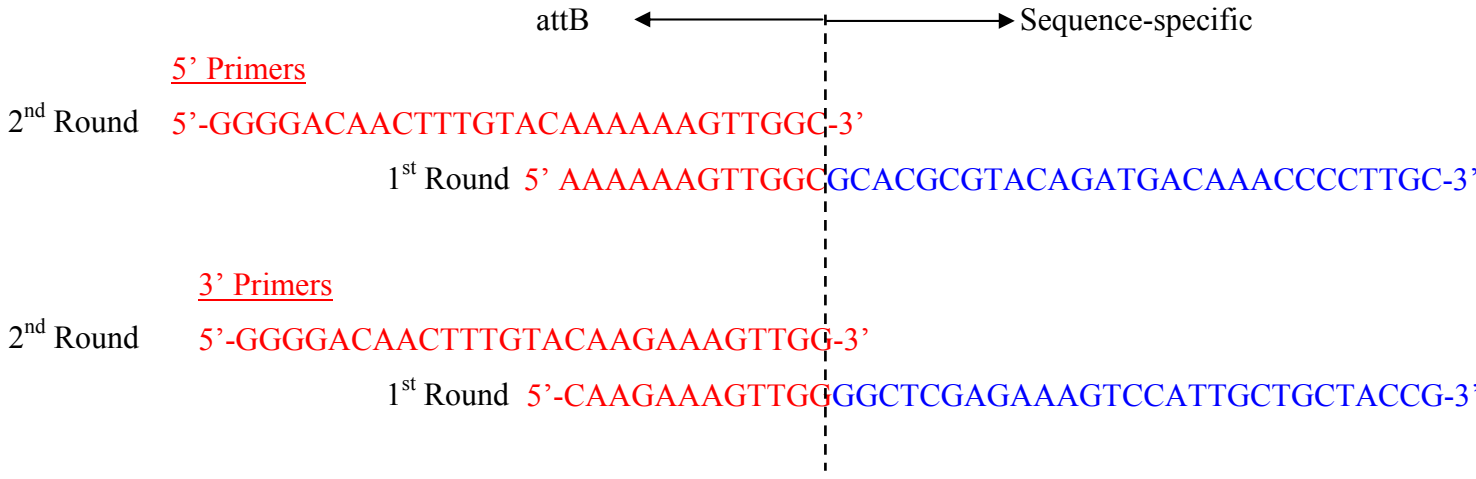


Figure 24. Primer design strategy for inserting attB sites upstream of promoter fragments by 2-round PCR. The first round primers contained a ~20-mer sequence-specific 3' section (blue), followed by linkers at the 5' end (red). The second round primers were universal and designed to anneal to the linker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

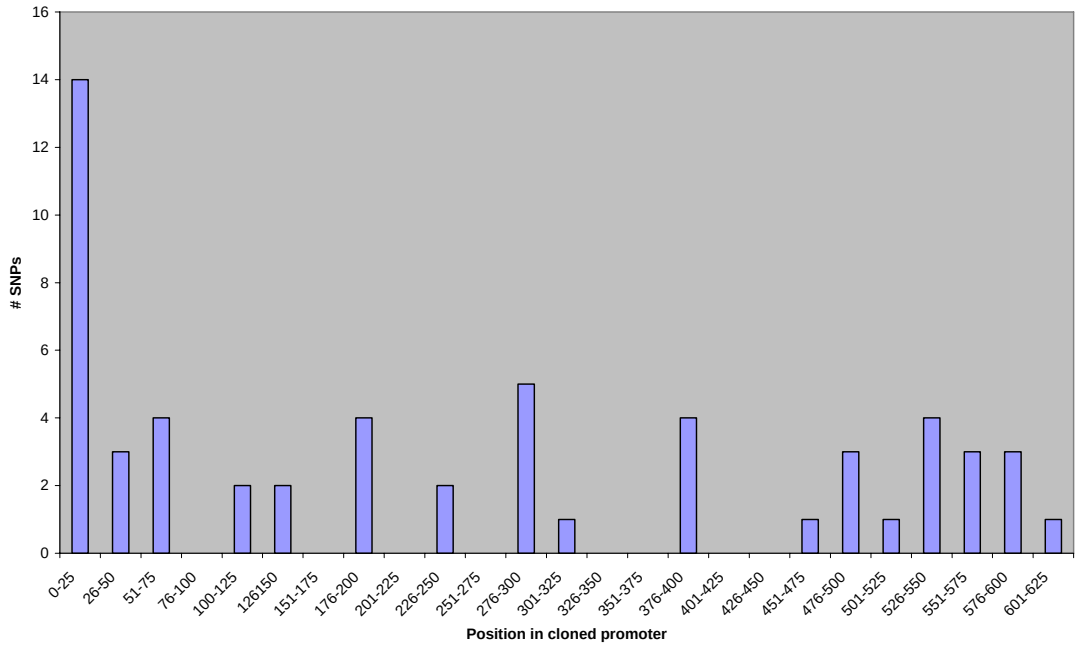


Figure 25. Distribution of SNPs in cloned promoter fragments that were not found in the re-sequencing data.

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

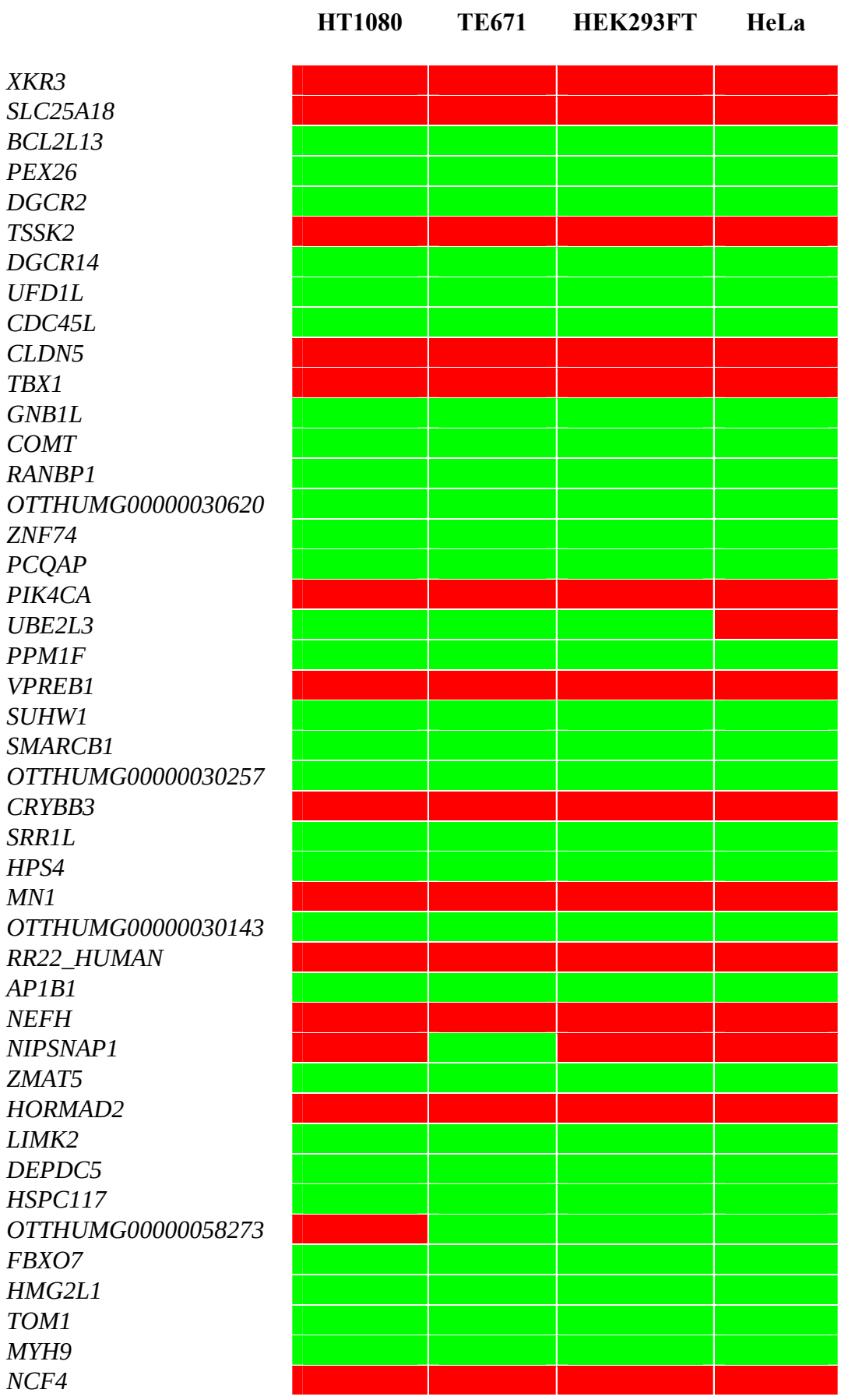
ensure success, or preferably a genotyping assay could be designed to confirm the genotype of the SNPs in each cell with less chance of being affected by surrounding DNA that is less tractable to sequencing. However, due to time constraints this was not possible.

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

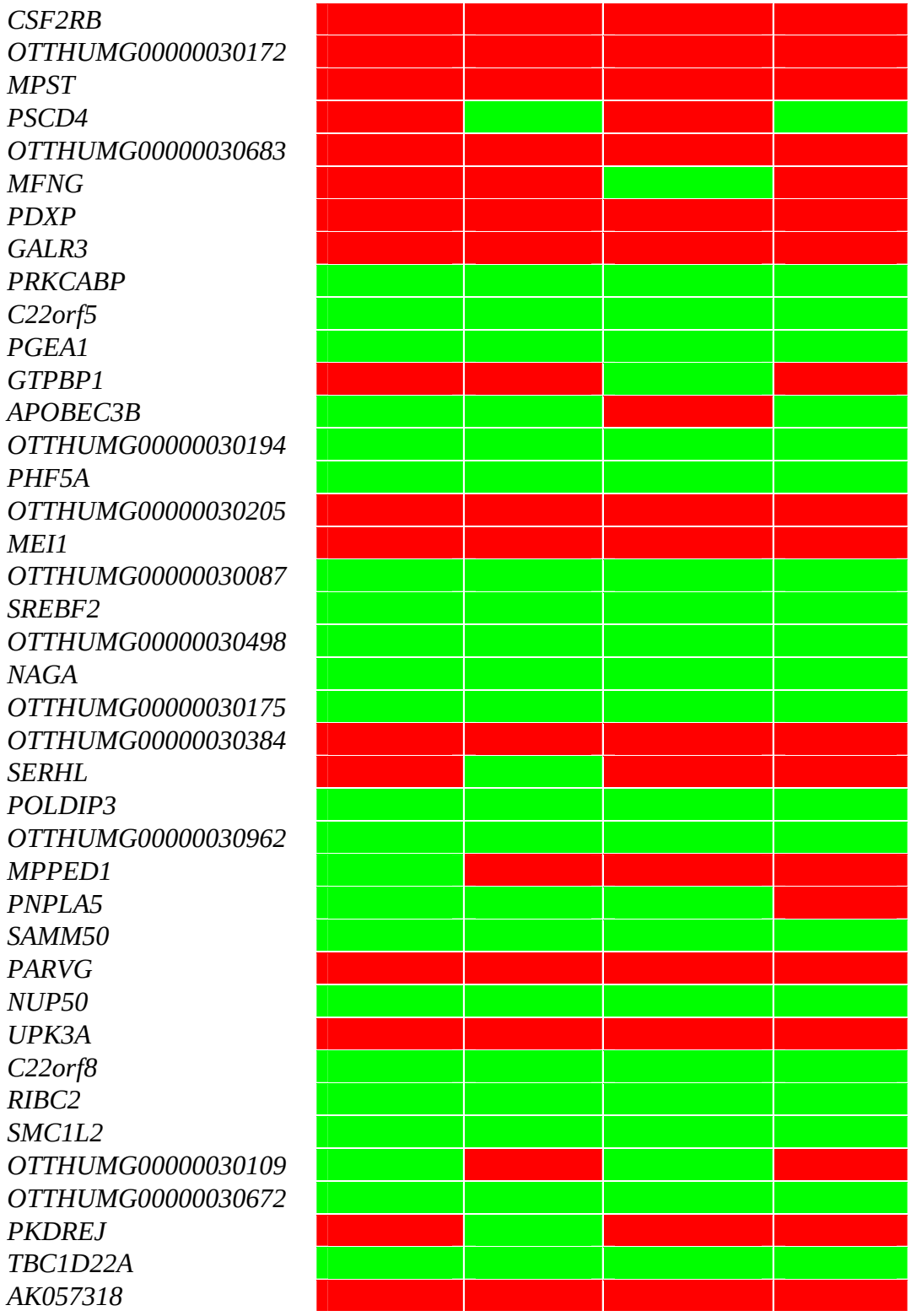


Table 9. Promoters active in each cell line. A promoter was defined as active if at least one haplotype gave a signal at least 7 times that of the promoter-less control plasmid. Promoters in green passed this threshold, while those in red were not active. Promoters are listed in the order of their occurrence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

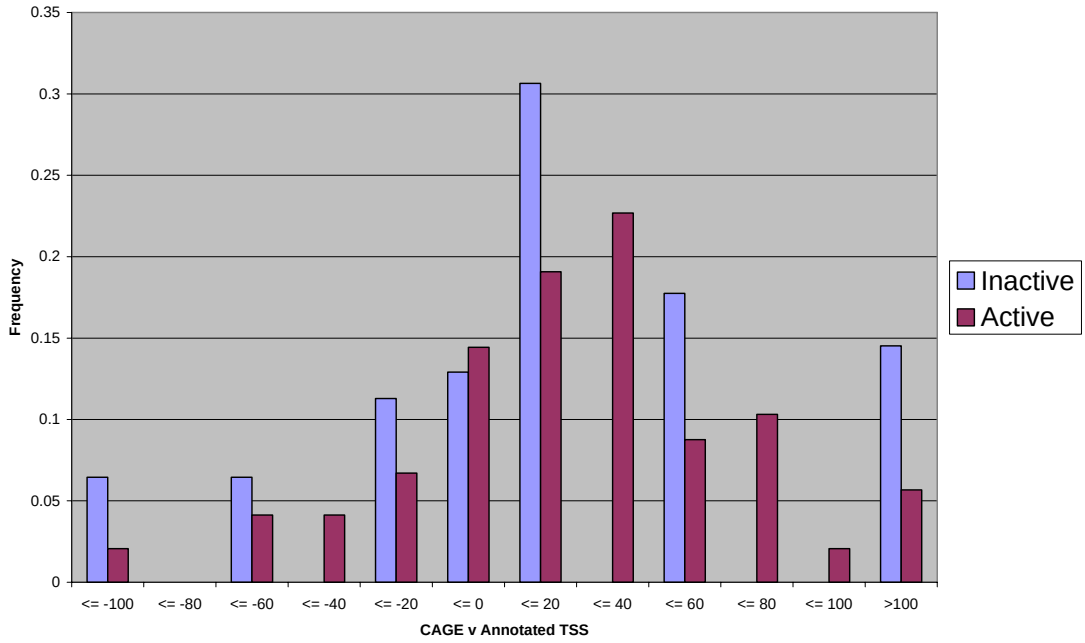


Figure 26. Correlation between accuracy of TSS annotation and cloned promoter activity. The x axis is the position of the CAGE-verified TSS relative to the annotated TSS.

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.

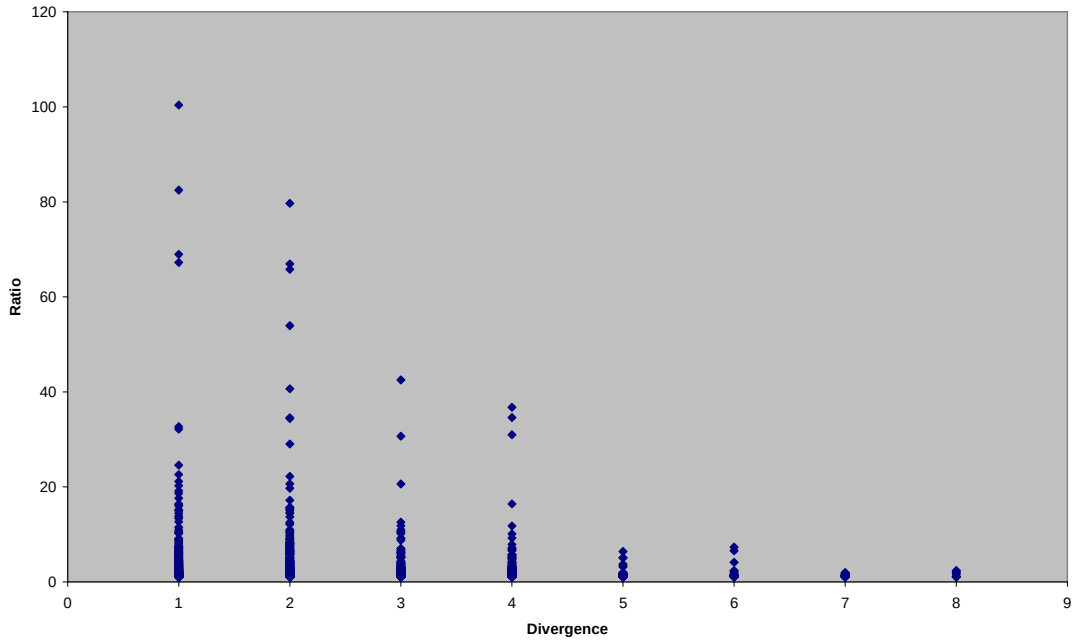


Figure 27. The effect of sequence divergence between promoter haplotypes on the degree of difference in promoter activity. Each point is the ratio of the activity levels of a pair of haplotypes from the same promoter, and all possible haplotype pairs where at least one haplotype is active are plotted.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

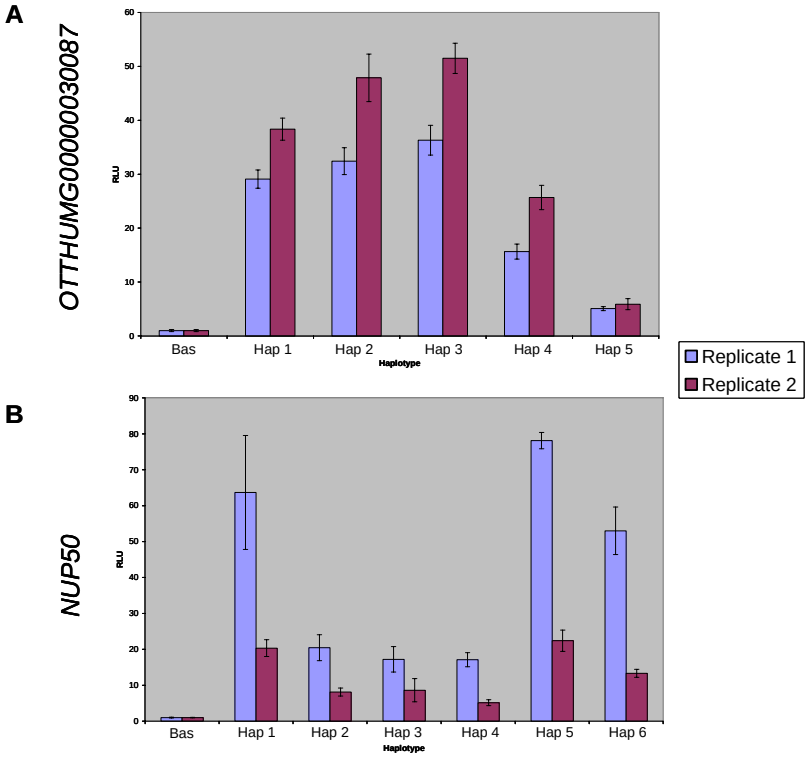


Figure 28. Conservation of promoter activity patterns but not magnitude of luciferase expression in 2 promoters. A) *OTTHUMG00000030087* luciferase results in TE671. B) *NUP50* luciferase results in HeLa. Error bars represent the standard deviation of 4 technical replicates. Promoter activities are plotted 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.

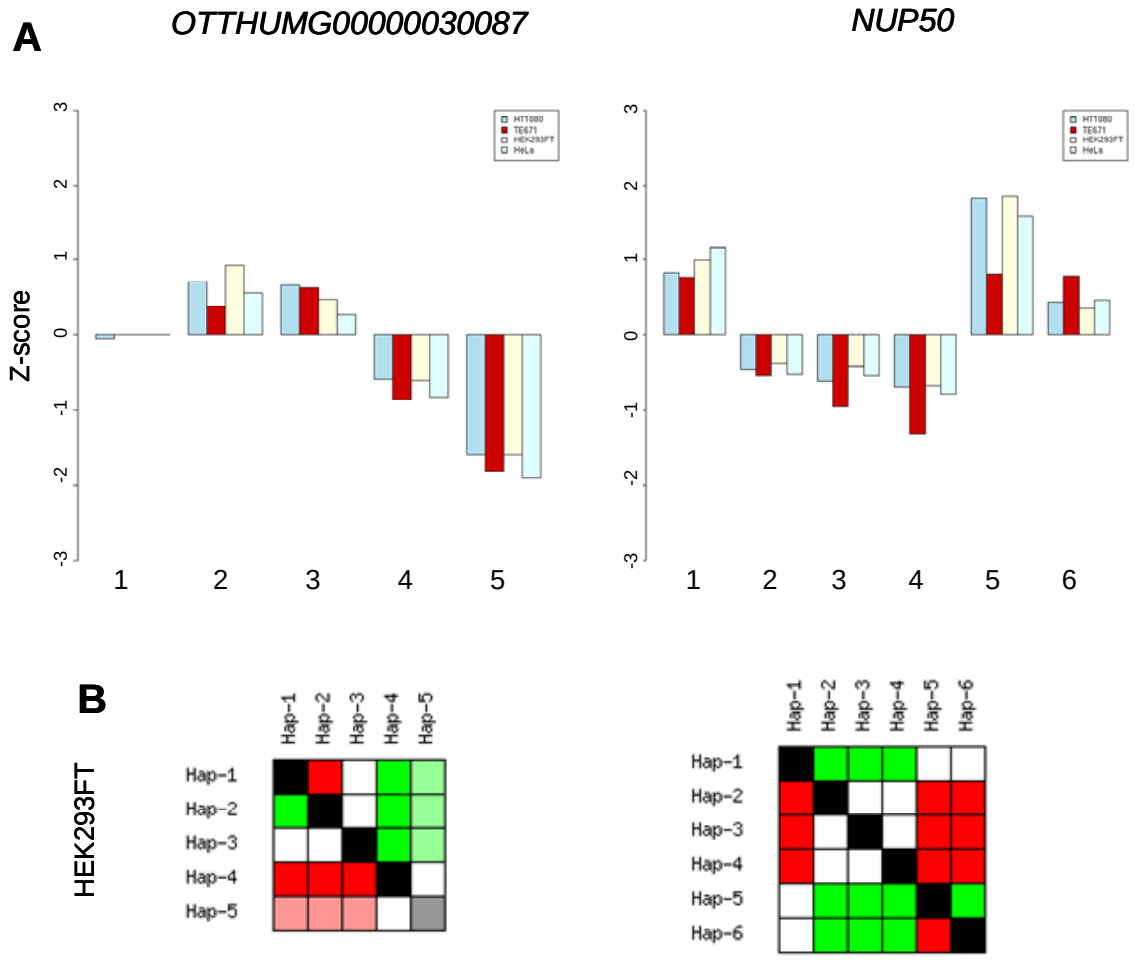


Figure 29. Visualisation of the luciferase reporter results for the *OTTHUMG00000030087* and *NUP50* promoters shown in Figure 28. A) Plot of the Z scores for each of the 4 cell lines. In this visualisation, the Z score represents deviation from the median, and the score itself is the median of the values 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

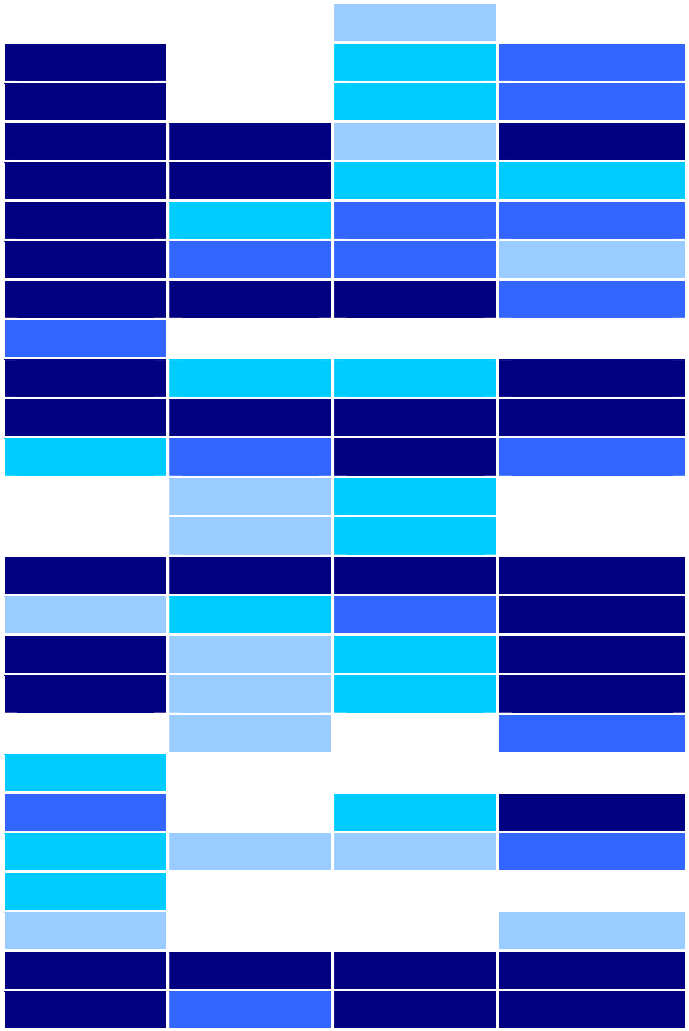
to a lack of statistical significance due to variability between technical replicates. Interestingly, these results using 4 cell lines have not revealed more cell type-specific variation than previous studies using 2 of the same cell lines used here (Buckland et al. 2005). It is also striking that where 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.

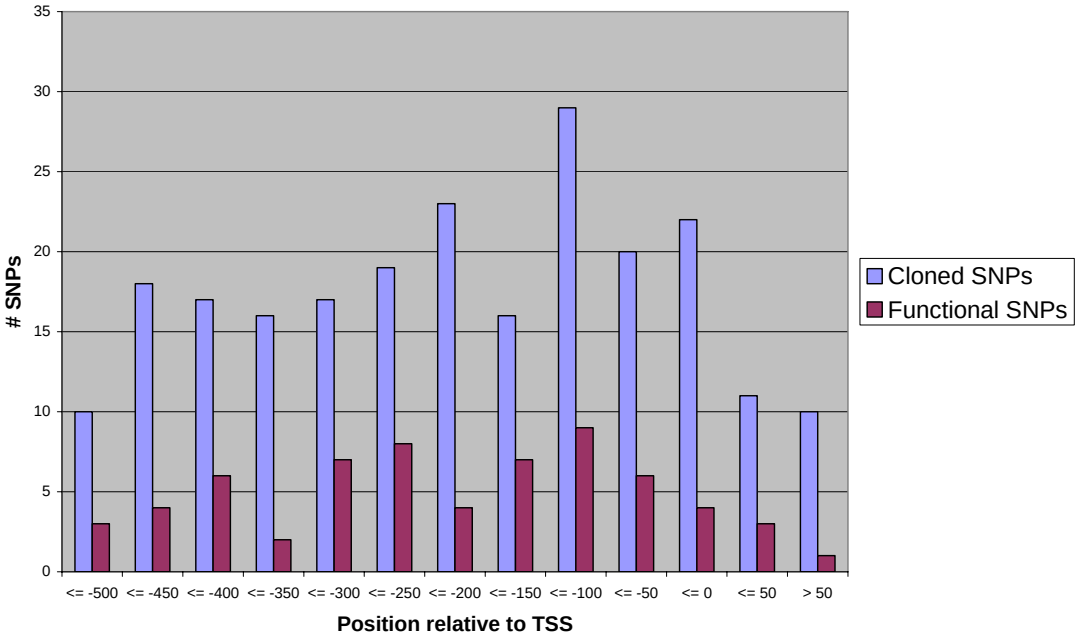


Figure 30. Distribution of cloned promoter polymorphisms and functional promoter polymorphisms as a function of distance from the TSS. No clear differences are visible between the distributions of functional relative to cloned polymorphisms overall. There is a marked drop in the number of both classes 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.

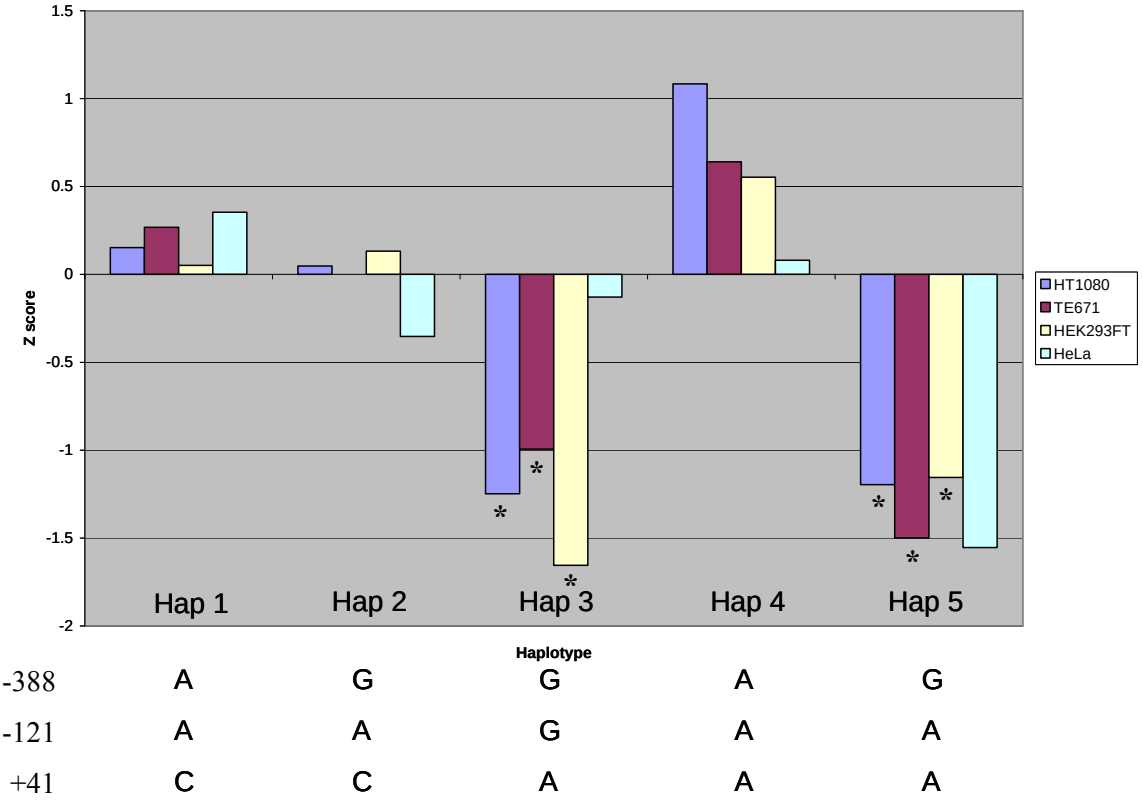


Figure 31. Z score plot of haplotype activities in the *RIBC2* promoter. Haplotypes 3 and 5, both containing a G at position -388 and an A at position +41 are significantly lower than other haplotypes in HT1080, TE671 and HEK293FT (marked by asterisks). There were no reproducibly significant differences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

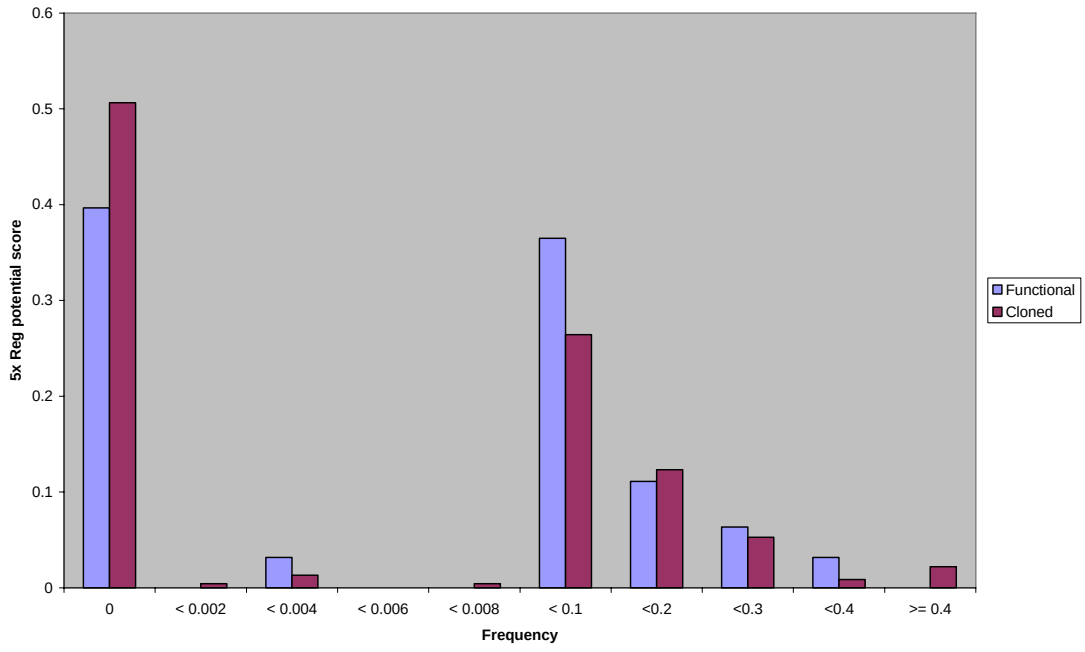


Figure 32. Frequencies of regulatory potential score for functional polymorphisms and promoter polymorphisms overall.

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.

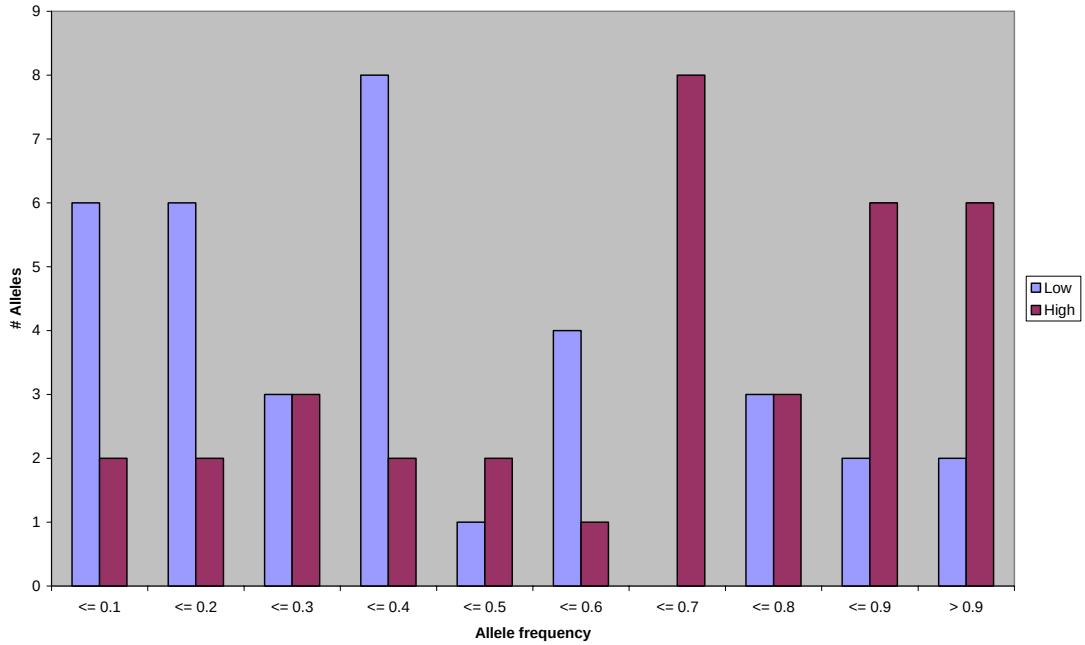


Figure 33. Allele frequencies of the low- and high-activity alleles in the functional SNPs for which frequency information was available

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.

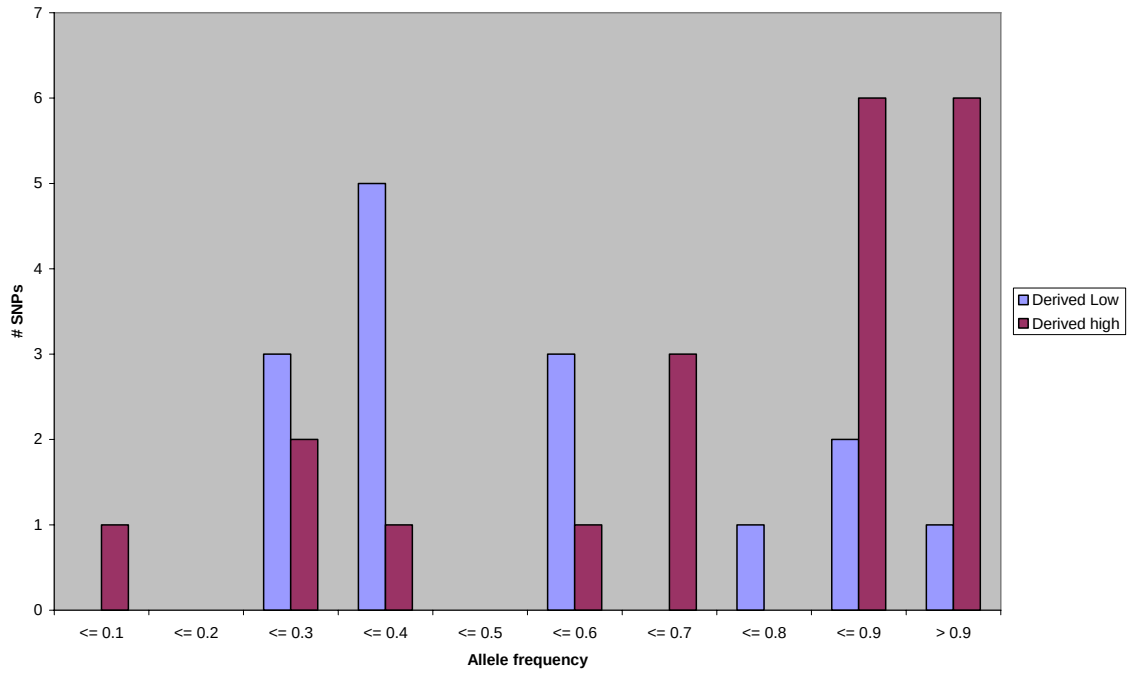


Figure 34. Allele frequencies of low- and high-activity derived alleles. The distribution of derived high-expression alleles (upregulatory mutations) is skewed towards high frequencies relative to derived low-experssion alleles (downregulatory mutations).

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.