6 Discussion and Future Work

6.1 Discussion

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

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

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

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

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

There is also evidence that non-binding site-dependent mechanisms may be important in explaining promoter SNP effects. These SNPs may function by altering the conformational properties of the DNA upstream of the TSS, and thus altering the dynamics of TF interactions with each other and the promoter without necessarily being in a binding site (Buckland 2006). The inherent curvature of DNA is often higher at promoters, and this has been shown to be an important factor in the activation of at least some eukaryotic genes (Nishikawa et al. 2003). Manipulations of cloned promoters in reporter vectors have shown that promoters with higher inherent curvature can promote transcription markedly more efficiently than the same promoter carrying mutations that reduced this curvature (Kim, Klooster, and Shapiro 1995). The addition of intercalators that abrogated this curvature greatly reduced this activity difference (Kim, Klooster, and Shapiro 1995). While structural studies show that some TFs, including TBP and p53 (Nagaich, Appella, and Harrington 1997; de Souza and Ornstein 1998), alter the conformation of DNA on binding, it is also the case that DNA which is already in a favourable conformation pre-binding can drastically increase binding affinity (Parvin et al. 1995). Alteration of TF binding efficiency by the introduction of artificial substitutions outside the TFBS that alter conformation has been demonstrated in yeast (Acton, Zhong, and Vershon 1997), although the presence or extent of natural SNPs that function in this way is unknown. A distinct but related property of the DNA itself that can be important in TF binding is the flexibility, or the ability of DNA conformation to be altered by the binding of proteins. This can be important in allowing multiple protein-DNA interactions in close proximity by relieving steric hindrances (either by one factor binding multiple sites or by multiple factors) or by allowing the DNA to loop and bring distant bound factors into contact (Mastrangelo et al. 1991; Suzuki and Yagi 1995; Nagaich, Appella, and Harrington 1997).

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

This project has also explored the qualitative relationship between promoter activity and *in vivo* expression. This has confirmed that promoter sequences contain many of the elements that determine whether a gene is expressed or not, and therefore that the promoter really does integrate the majority of signals in the transcription initiation pathway. Other work has found a more quantitative relationship between promoter activity and gene expression (Cooper et al. 2006), but this was not reproduced here. As suggested before, this may be due to the relative quantitative potential of RT-PCR (as used by Cooper et al) and Affymetrix arrays. Another factor may be the difference in the controls used for the luciferase assays, where a single promoterless plasmid was used in this project versus the average of 102 cloned non-functional DNA elements by Cooper et al. This latter control may form a more consistent baseline as any non-specific activation of transcription due to stochastic biological variation in different cell growths would perturb the baseline by relatively low levels. Indeed, Promega have recently released the pGL4 luciferase plasmid series, where a large number of cryptic TFBS were removed from the vector backbone relative to the pGL3 plasmids. These may have been a source of variation in background levels.

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

The bias towards expansion of upregulatory changes seems at odds with some theoretical models of transcriptome evolution, which propose that downregulatory changes should be more common that upregulatory ones (Khaitovich, Paabo, and Weiss 2005). It also does not agree with recent findings by the Dermitzakis lab at the Sanger Institute that SNPs found by whole genome association to expression phenotypes agree with this model (Stranger et al unpublished). However, it is important to note that while Stranger et al were measuring mRNA levels, this project was measuring *in vitro* promoter activity, with the latter being a component of the former. A possible explanation for the discrepancy is that these association studies may be finding regulatory SNPs in distant enhancer or silencer elements rather than the promoter, and that such functional SNPs may have more powerful effects than those at promoters. This is suggested by the fact that the majority of SNPs identified by Stranger et al are more than 10 kb away from the TSS of the genes they influence (data not shown). The effects of these elements on transcription may be sufficiently powerful that where they contain functional variation, this dominates over promoter sequence variation, and precludes it from identification in association studies. This may also explain discrepancies in the difference between human and chimpanzee promoter activities and the corresponding difference in transcript levels (Heissig et al. 2005). Heissig and colleagues found seven genes that showed significant differences between chimpanzees and humans both in luciferase reporter assays and measures of transcript abundance. However, in 4/7 genes these differences were in the opposite direction to each other (Heissig et al. 2005). It may therefore be proposed that globally, variation in proximal promoters and in distal regulatory elements are influenced differently by selection.

6.2 Future work

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

The most obvious method would be to look for differences in the mRNA transcripts produced by variant promoters. The most well-established method for doing this is probably quantitative RT-PCR from cell lines or mRNA from heterozygotes (Yan et al. 2002b; Bray et al. 2003; Pastinen and Hudson 2004). This would require the identification of individuals who were heterozygous both for the promoter haplotypes of interest and for a transcribed marker SNP that could be used to distinguish the two transcripts. It would also necessitate knowledge of the phase of the promoter haplotypes and the maker SNP in order to be able to say which promoter haplotype is driving the expression of which transcript, enabling the assignment of direction to functional changes. With the HapMap project now having completed phase 1, there is a ready source of cell lines from a range of individuals that can be used for this kind of work (Consortium 2005b). The genotype information would also enable the inference of phase between transcribed markers and the promoter SNPs (Stephens, Smith, and Donnelly 2001).

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

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

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

Finally, current knowledge of functional promoter polymorphisms can be used to build a database of polymorphisms for which function is known *a priori*, and use this for meta-analysis to examine the properties of functional SNPs more thoroughly. This database can then be put through the above battery of methods in order to complete the knowledge required for each of the polymorphisms. Two sets of promoter polymorphisms tested *in vitro* under homogeneous experimental conditions are already available; the data presented in the thesis and that produced by Buckland et al. Together, these consist of 79 isolated and confirmed promoter polymorphisms. There have also been two efforts to curate information from the wider literature, which contains data on many more promoter variants distributed among a large number of papers. Rockman and Wray produced a survey of 140 functional SNPs tested in reporter assays in the literature, and in many cases were able to find supporting published evidence in the form of EMSA experiments or associations with expression or disease phenotypes. In addition, the ORegAnno database of regulatory elements (Montgomery et al. 2006) contains a set of 172 promoter polymorphisms that have been partly manually curated and partly submitted by external contributors. In both these curated datasets, the SNPs are not always clearly-mapped to the genome and the evidence supporting each SNP is very heterogeneous (in some cases, for example, there is differential binding data from EMSA but no luciferase assay). These would then be put through the methods proposed to complete the evidence for them, with the expectation that there would be a high rate of functional confirmation. Indeed, I was able to construct a preliminary database that would hold the integrated results of such a meta-analysis of published functional promoter SNPs, and was able to populate it with data from both the Buckland set and from individual papers. This work was not presented in this thesis, as more work is needed to establish an ontology for populating it with a dataset that can be consistently analysed.

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