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

3.1 Introduction

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

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

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

After selecting the target genes, the next step was to choose the population in which the SNP discovery phase would be carried out. Bearing in mind that the aim of this particular study was not only to discover and genotype SNPs but subsequently clone haplotypes and functionally interrogate individual SNPs, the selection of a panel that would maximise the number of SNPs discovered would not necessarily be ideal. There were two possible strategies to follow; either selecting a panel of individuals of diverse ethnic origin or a larger population from a single ethnic group. The ethnically diverse panel would likely yield more SNPs than the single population, as the individuals will be more diverged. However, as the SNPs would have been arising in parallel lineages prior to being placed in the same pool, the haplotypes would be more different from each other than would be the case if the population were ethnically homogeneous. This is equivalent to the effect of admixture on the linkage disequilibrium patterns in a population (Hartl and Clark 1997; Huttley et al. 1999; Pritchard and Przeworski 2001). When two or more previously isolated populations mix, linkage disequilibrium increases, particularly when the admixture is sudden. With a larger single population, there may be fewer SNPs discovered (depending on the relative sizes of the panels), but these SNPs will have been segregating in the same population, and there will have been more recombination between them. This increase in the number of combinations of alleles means that, when using the naturally occurring haplotypes to investigate the effect of polymorphism on gene expression, the effect of individual SNPs can be interrogated more easily. Another important consideration was that for a given number of individuals, an ethnically diverse panel comprising a smaller set of samples from different populations would lead to a bias against the discovery of rare SNPs compared to a panel with no substructure. For example, in a panel of 10 individuals in a single population, the lowest minor allele frequency that could be determined by diploid resequencing would be 0.05 (a single heterozygote in the panel). If this panel was composed of 2 individuals each from 5 sub-populations, then the lowest determinable allele frequency for a lineage-specific SNP would be 0.25 (a single heterozygote out of the 2 individuals in the sub-population). Of course, SNPs with minor allele frequencies below these thresholds would still be discovered, but they would be disproportionately less likely to be found compared to more common alleles, and their true allele frequency would not be determinable below these values. For these reasons, it was decided that that a moderate-size panel of individuals from a single population would be used in order to obtain more haplotypes for the given number of SNPs.

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

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

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

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

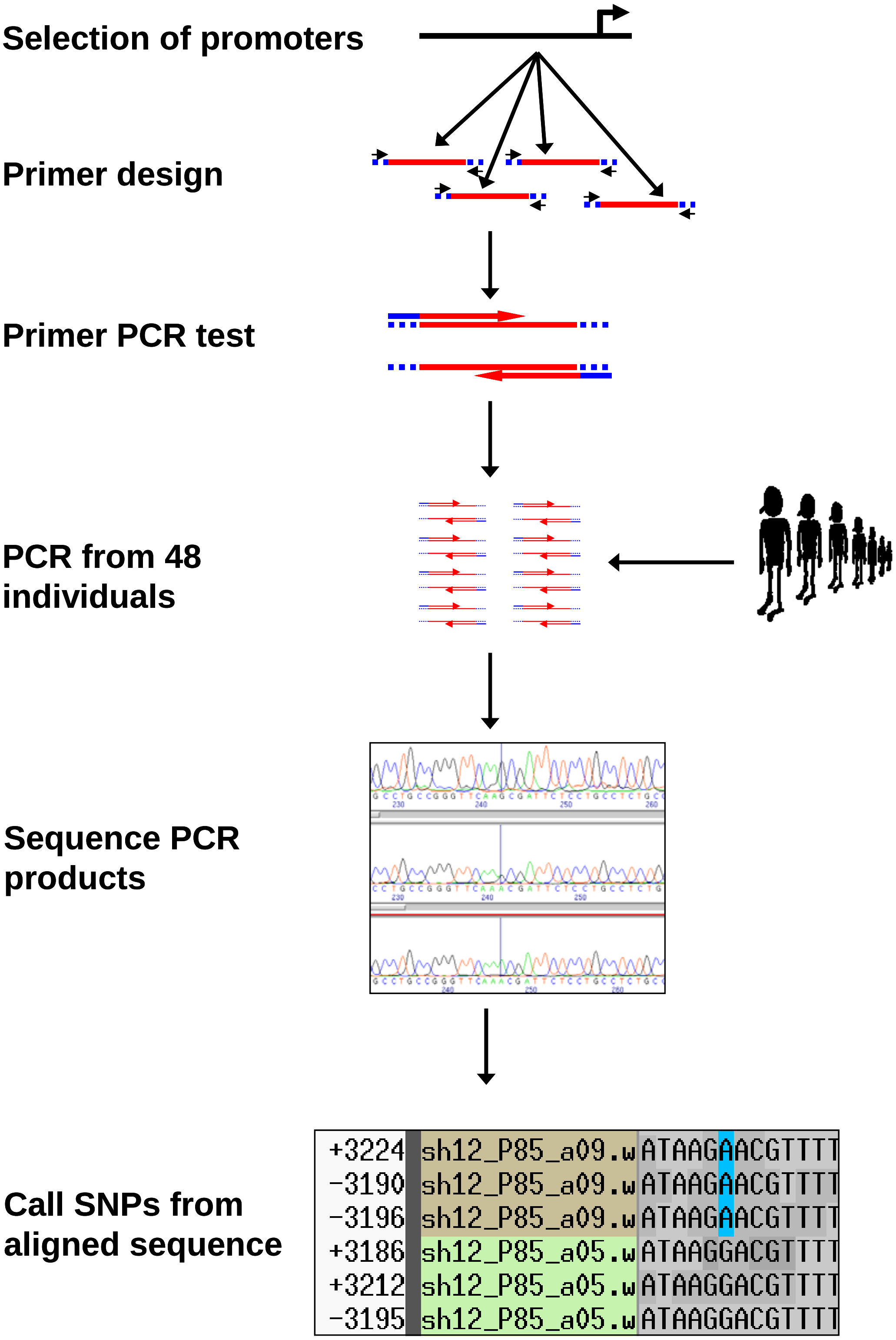

Selection of promoters
Primer design
Primer PCR test
PCR from 48 individuals
Sequence PCR products
Call SNPs from aligned sequence
+3224 sh12_P85_a09.w ATAAGAACGTTTT
-3190 sh12_P85_a09.w ATAAGAACGTTTT
-3196 sh12_P85_a09.w ATAAGAACGTTTT
+3186 sh12_P85_a05.w ATAAGGACGTTTT
+3212 sh12_P85_a05.w ATAAGGACGTTTT
-3195 sh12_P85_a05.w ATAAGGACGTTTT

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

3.2 Results

3.2.1 Selection of promoters for SNP-mining

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

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

3.2.2 Primer design

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

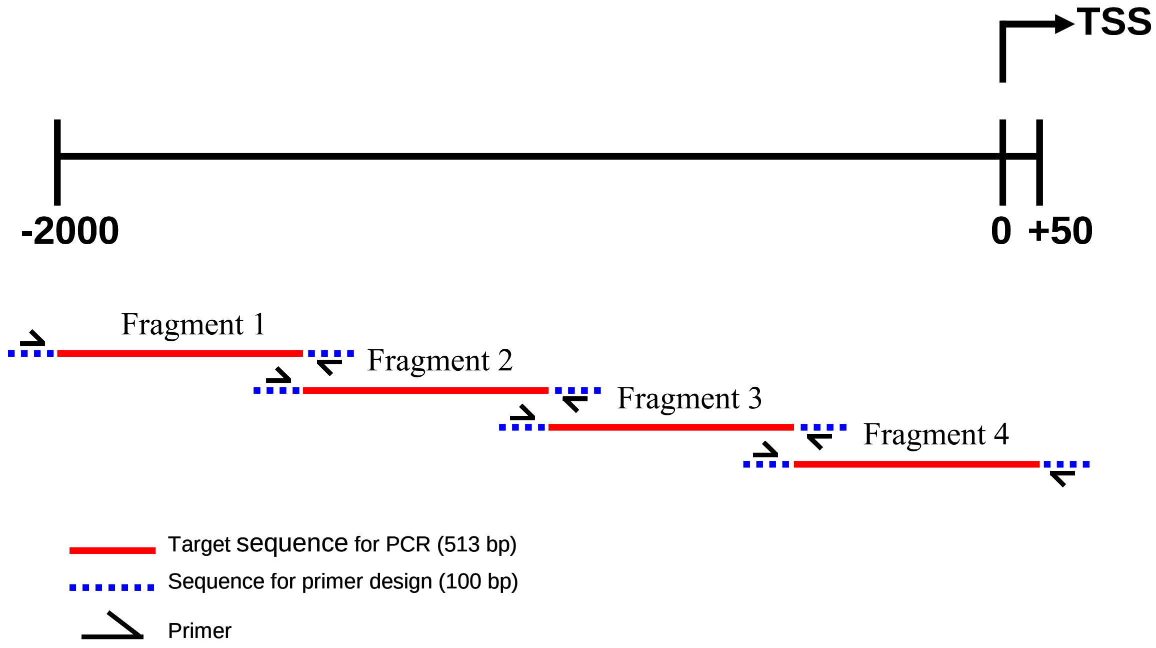

TSS
-2000
0
+50
Fragment 1
Fragment 2
Fragment 3
Fragment 4
Target sequence for PCR (513 bp)
Sequence for primer design (100 bp)
Primer

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

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

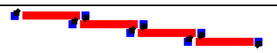
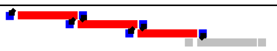
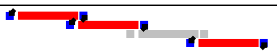
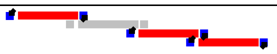
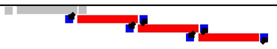
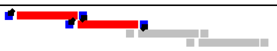
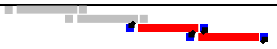
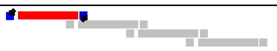
# Fragments	Promoter coverage		# Promoters
4	1,2,3,4		312
3	1,2,3		2
	1,2,4		6
	1,3,4		9
	2,3,4		6
2	1,2		2
	3,4		1
1	1		5

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

3.2.3 Primer tests and PCR optimisation

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

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

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

Initially, all primer pairs were tested using a standard protocol for genomic PCR that had been optimised by Bentley et al (unpublished). 892 (62%) gave clean bands with the standard protocol, 245 showed non-specific amplification and 228 showed weak or no amplification (Table 3). The latter two categories were re-tested in new PCRs with different conditions designed to compensate for amplification problems.

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

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

3.2.4 PCR and sequencing of promoter fragments

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

3.2.5 ExoTrace pipeline for sequence analysis and SNP detection

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

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

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

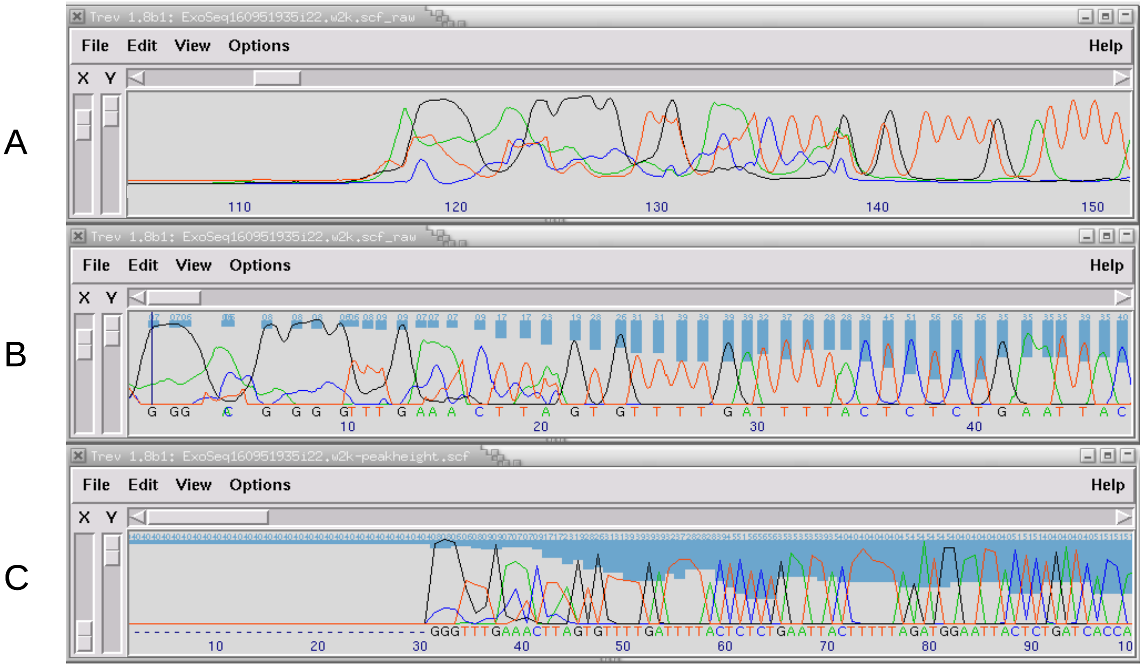
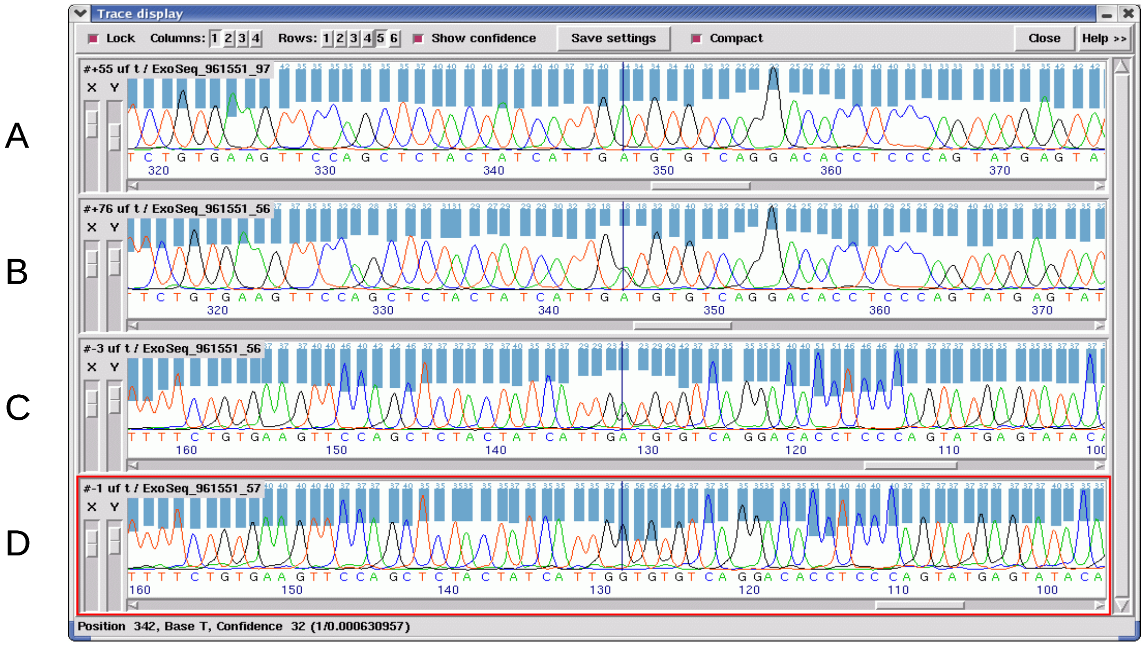

A
B
C
Trev 1.8b1: ExoSeq160951935122.w2k.scf_raw
Trev 1.8b1: ExoSeq160951935122.w2k-peakheight.scf
File Edit View Options Help
X Y
110 120 130 140 150
10 20 30 40
10 20 30 40 50 60 70 80 90 10

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

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


Trace display
Lock
Columns: 1 2 3 4
Rows: 1 2 3 4 5 6
Show confidence
Save settings
Compact
Close
Help >>
A
#+55 uf t / ExoSeq_961551_97
B
#+76 uf t / ExoSeq_961551_56
C
#-3 uf t / ExoSeq_961551_56
D
#-1 uf t / ExoSeq_961551_57
X
Y

Position 342, Base T, Confidence 32 (1/0.000630957)

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

3.2.6 Second round of primer design

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

A second round of primer design and re-sequencing was started, with a primer design strategy to compensate for this problem. Each promoter containing at least one run of 8 or more of the same base were selected for the second round. Primers were designed using Primer3 and the same parameters as the first round. Rather than split the promoters in to equal blocks of target sequences, they were split using the polyN runs as boundaries (Figure 8). 129 promoters were found to contain polyNs, and the 558 new primer pairs were designed for them.

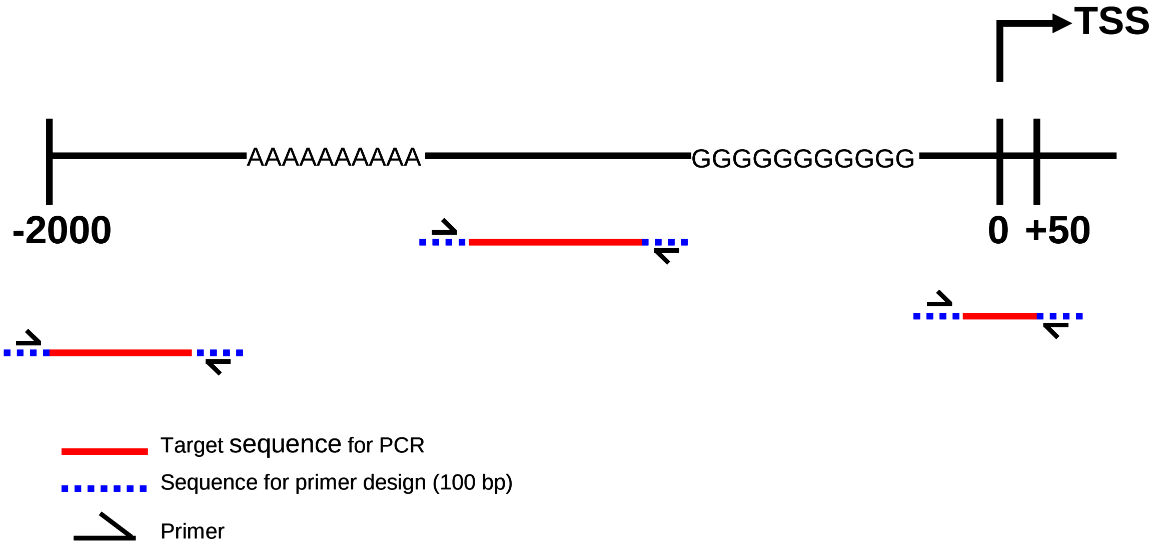


Figure 8. Schematic of the strategy for primer design around polyN motifs.

3.2.7 PCR tests of the second batch of primers

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

# Amplicons	Annealing Temperature / °C	
	60°C	65°C
Tested	558	558
Successful	346	320
Amplified	384	-

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

3.2.8 Promoter sequencing results

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

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

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

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

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

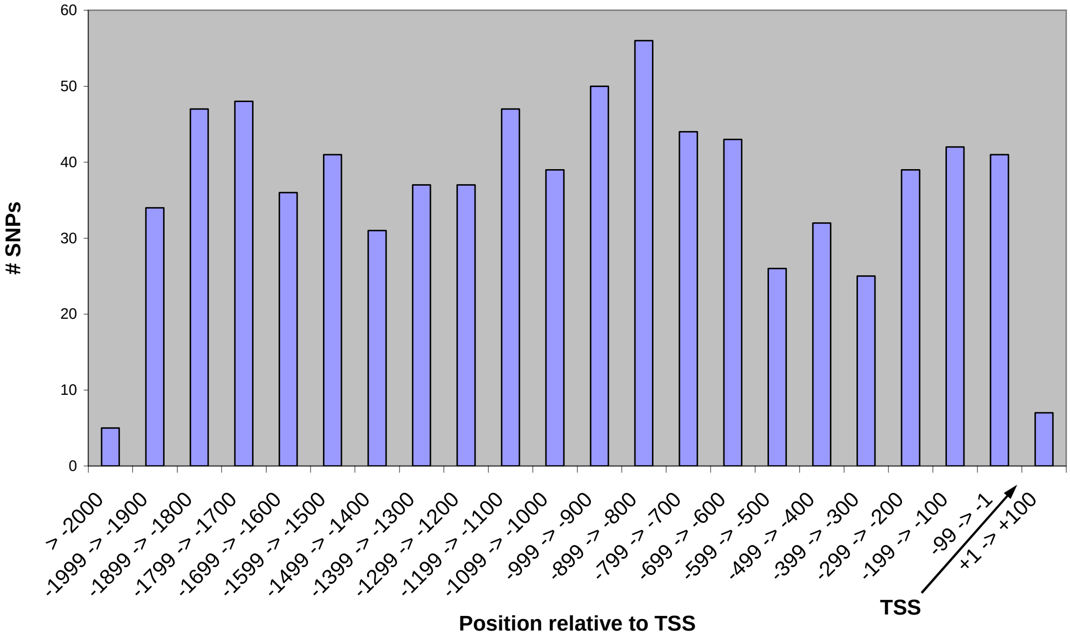


Figure 9. Distribution of SNPs relative to the transcription start site (TSS).

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

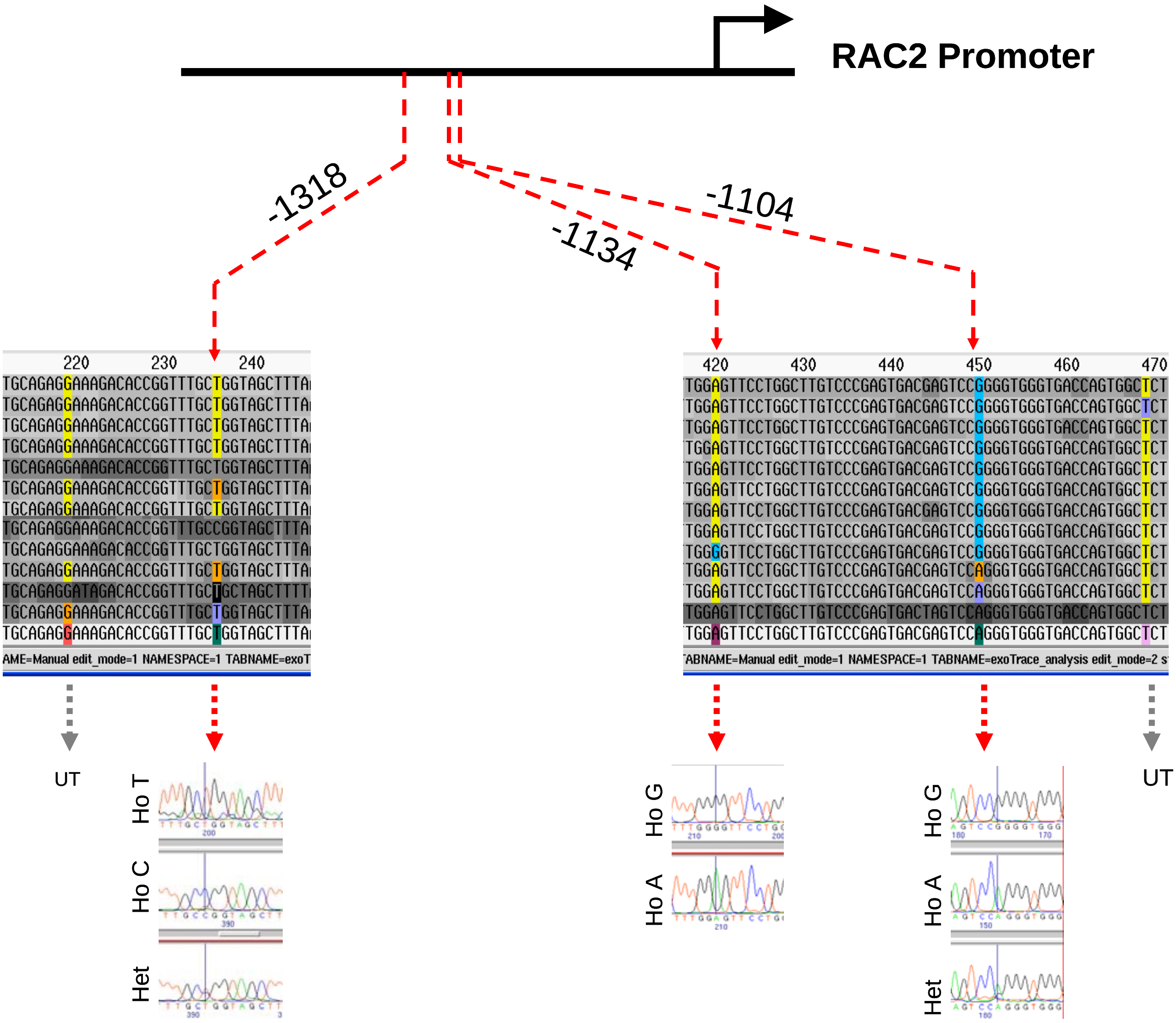


Figure 10. Schematic of the SNP-finding process using the RAC2 promoter as an example. A) Three SNPs were found in this promoter; a C/T SNP at -1318, an A/G SNP at -1134 and a third A/G SNP at -1104 from the TSS. B) All successfully sequenced PCR products were aligned and ExoTrace was used to detect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

3.2.9 Distribution of SNP types and allele frequencies

The minor allele frequency of each SNP was calculated by counting the homozygous and heterozygous calls on each of the 48 samples sequenced. This gives a frequency resolution of 1/96, or 0.0104, and means that alleles as rare as 0.01 minor allele frequency can be detected. This assumes that all 48 samples were amplified successfully and sequenced to good quality. In practice, there is often a loss of a small number of samples due to stochastic failures in sequencing or PCR, meaning that many SNPs are called from fewer than 48 samples (and in some cases substantially fewer). This would be expected to push the minor allele frequency distribution in favour of common SNPs. As would be expected, the majority of SNPs found in the promoter re-sequencing had small minor allele frequencies (Figure 11).

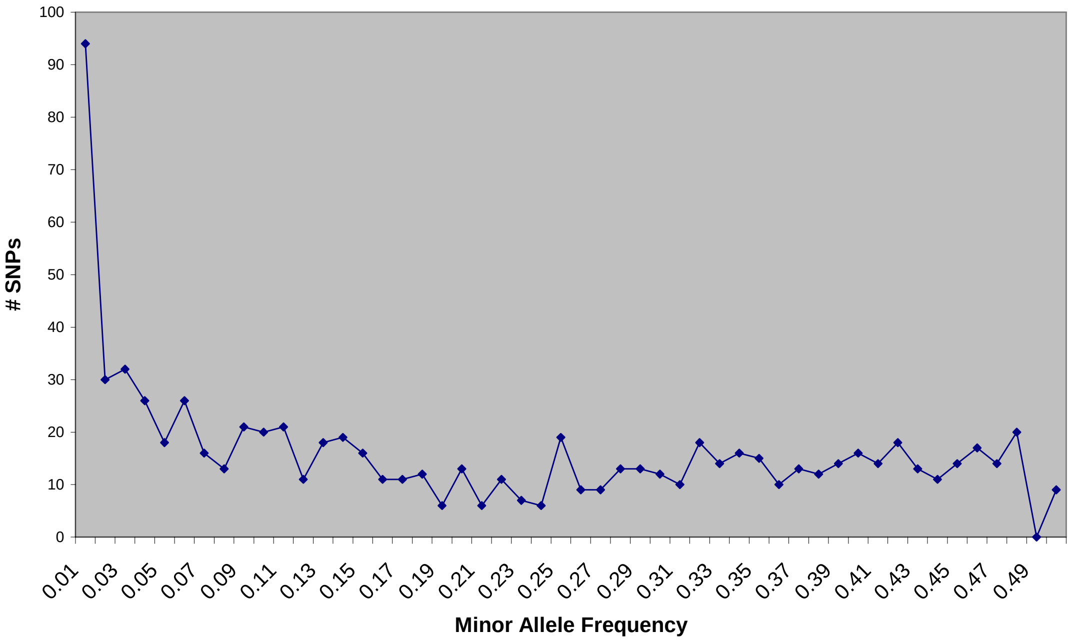


Figure 11. Distribution of the minor allele frequencies of chromosome 22 promoter SNPs.

The distribution of SNP types was compared to a control set from chromosome 22 as a whole in order to see whether there are any differences in the kind of SNPs found in promoters relative to what would be expected. The control set was made up of all SNPs in dbSNP build 125 from chromosome 22 that could be aligned to the chimpanzee genome. This was to enable later use of the chimp sequence to infer direction (see section 3.2.12). The proportions of the different SNP types did not deviate significantly (p-value = 0.19, χ^2) from that expected in the whole genome (Figure 12 A and B). This was somewhat surprising, as an under-representation of C/T SNPs due to the lack of methyl-cytosine deamination at promoters may have been expected. There was a small increase in C/G SNPs at the expense of transitions, consistent with higher GC content, but this was very small and not significant. The SNPs were divided into two sets according to their presence in CpG islands, according to the CpG island annotation on the UCSC genome browser (NCBI build35). This revealed a marked difference in the distributions of SNPs in CpG islands relative to chromosome 22, with far fewer A/T SNPs and transitions. This may be due to a combination of lack of cytosine methylation and elevated GC content. However, the distribution of promoter SNPs outside CpG islands is not significantly different from that of promoter SNPs generally, or from chromosome 22 as a whole (p-value = 0.53, χ^2).

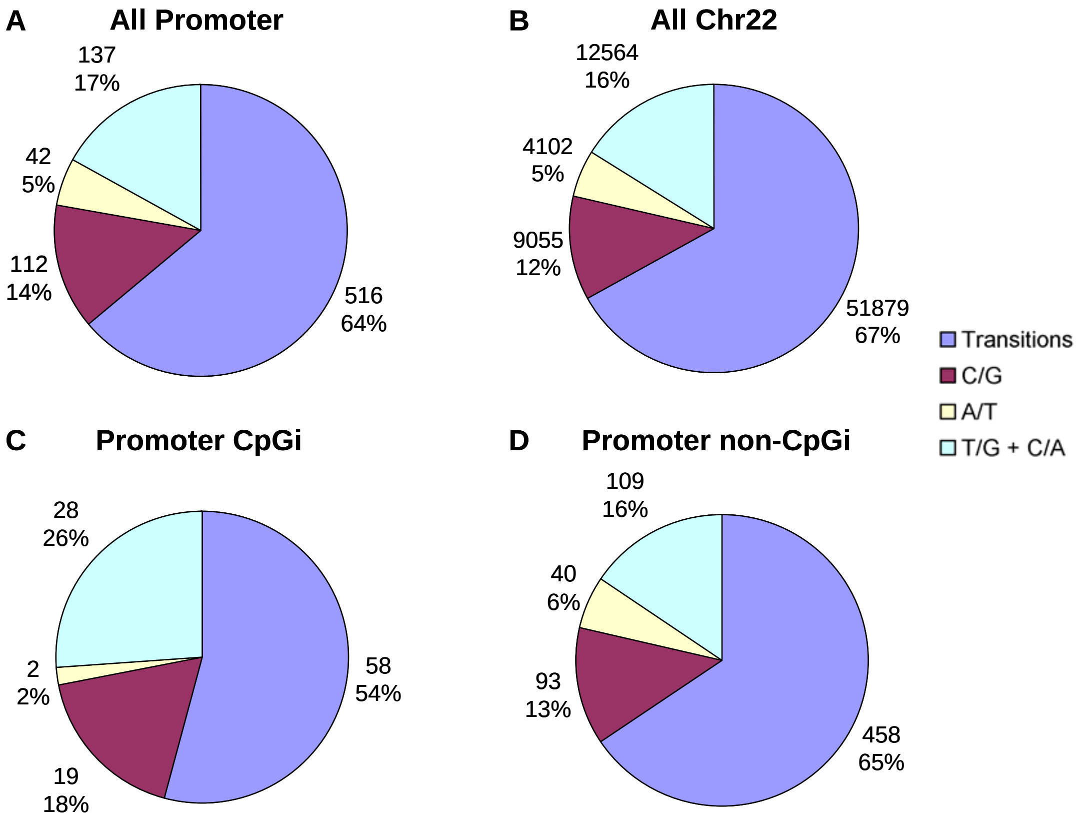


Figure 12. Distributions of the SNP alleles relative to chromosome 22 and to CpG islands in promoters. A) All SNPs from the promoter re-sequencing dataset. B) SNPs from chromosome 22 (Collins et al.). C) Promoters SNPs within CpG islands (according to the UCSC genome browser). D) Promoter SNPs outside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

3.2.10 Comparison of polymorphic promoters with downstream gene function

If promoter sequence polymorphism has an effect on the level of gene expression, then one can hypothesise than some functional classes of genes would be more tolerant of such changes than others. For example, genes involved in crucial processes such as cell cycle control or DNA damage repair might be hypothesised to have lover mutation rates at their promoters compared to other genes such as extracellular receptors due to purifying selection eliminating variation in the former. A recent study has found evidence that genes are preferentially located in mutational hot or cold spots depending on their function (Chuang and Li 2004). In order to test this idea, the Gene Ontology (GO) terms associated with genes having polymorphic promoters was compared to those for the genome as a whole. Five different lists of 1000 randomly selected human genes were generated by Juanma Vaquerizas at the European Bioinformatics Institute to use as comparisons with the list of genes with polymorphic promoters discovered in this project. The FatiGO tool (Al-Shahrour, Diaz-Uriarte, and Dopazo 2004) was used to compare these lists of genes across all GO hierarchies and levels. No significant over- or under-representations of GO terms were found for any level of the GO structure (Figure 13).

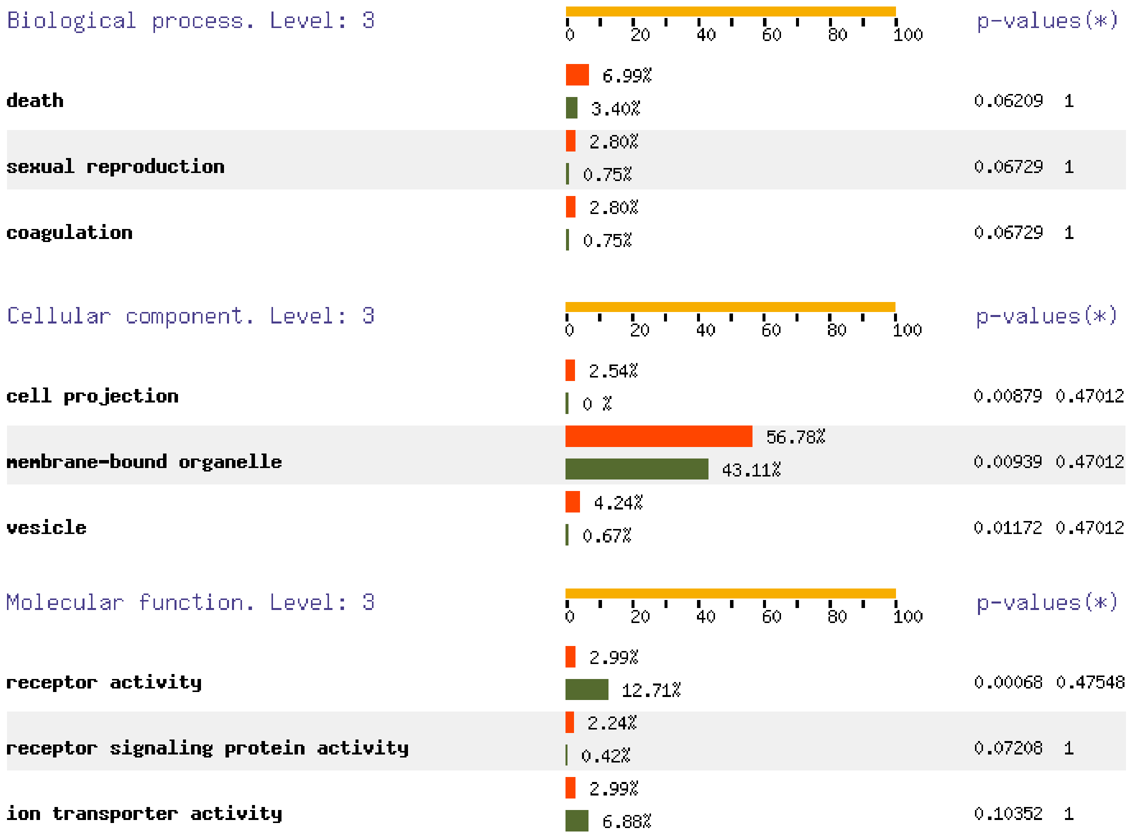


Figure 13. Comparison of the Gene Ontology terms for genes with polymorphic promoters (orange bars) against a list of 1000 randomly selected genes (green bars). This analysis was carried out using the FatiGO web tool, and was repeated for five different control lists of 1000 random genes. In all cases,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

3.2.11 Analysis of the genomic context of promoter SNPs

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

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

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

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

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

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

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

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

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

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

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

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

3.2.12 Evolutionary analysis of the SNPs using the primate genomes

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

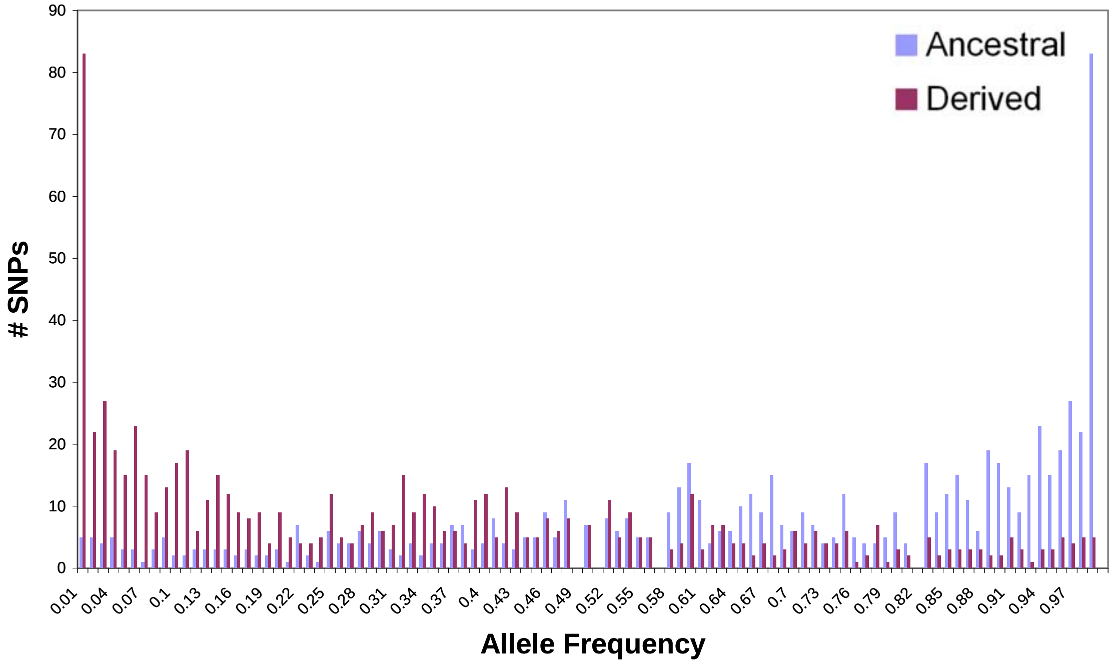


Figure 14. Allele frequency spectrum for ancestral and derived alleles rooted with the chimpanzee and macaque genomes. The two distributions are symmetrical due to the relationship between the two allele frequencies (i.e. one frequency is 1 minus the other frequency). There is a marked bias of derived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

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

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

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

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

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

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

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

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

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

3.2.13 Association of promoter SNPs with gene expression levels

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

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

Only one promoter SNP was significantly associated with an expression phenotype in the downstream gene. This was a C/T SNP 1747 upstream of the TSS of the *SNAP29* gene. The *SNAP29* protein is involved in intracellular vesicle trafficking in neurons, and truncation of the protein has been linked to severe neurocutaneous abnormalities (Sprecher et al. 2005). Interestingly, previous studies have reported a significant association between another SNP in the *SNAP29* promoter and schizophrenia (Saito et al. 2001; Wonodi et al. 2005). This was an A/G SNP at -849 bases from the TSS, and is present in dbSNP as rs165596. The G allele in this SNP was found to be significantly overrepresented in schizophrenia patients relative to control groups. This SNP was not detected in the promoter SNP mining as the amplified fragment in which it would be located failed to return usable sequence. The -1747 C/T SNP is novel and has never been reported before. However, the relationship between the two SNPs could still be determined because rs165596 was genotyped in the HapMap project.

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

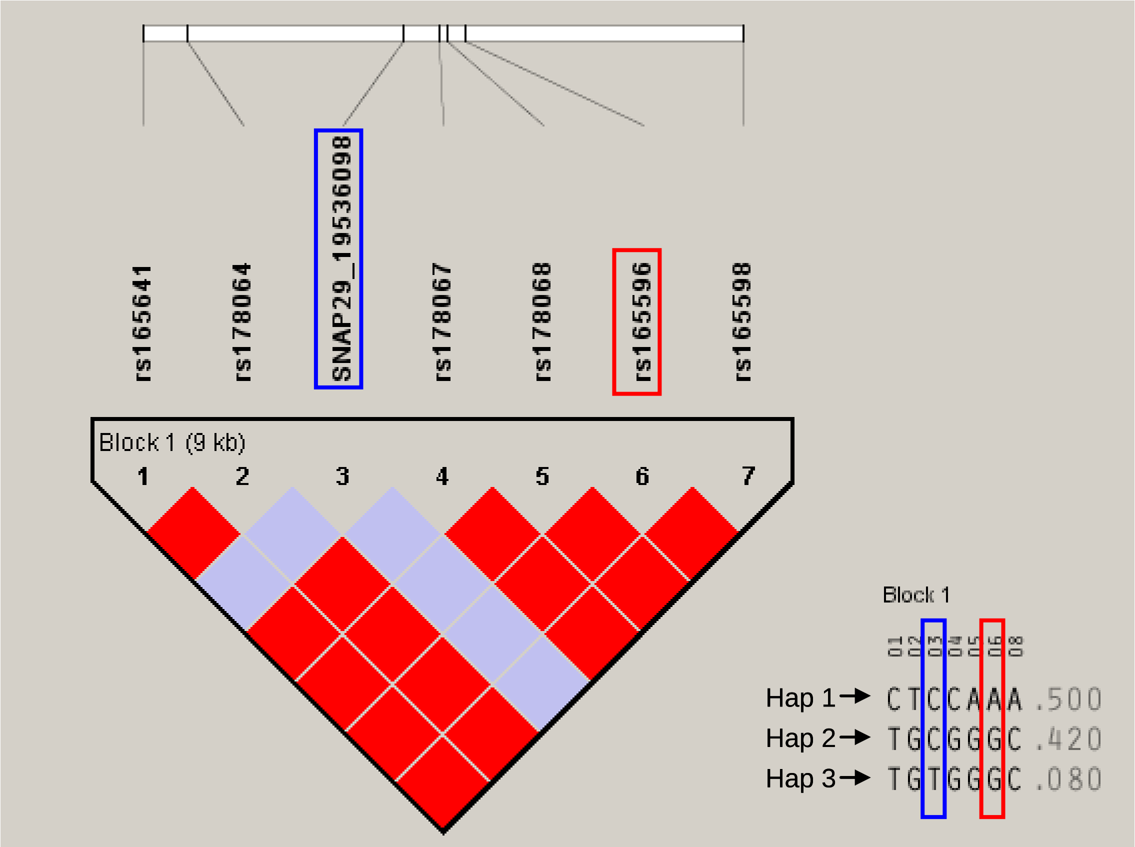


Figure 15. Linkage of the T allele of the novel -1747 *SNAP29* SNP with the G allele of rs165596.

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

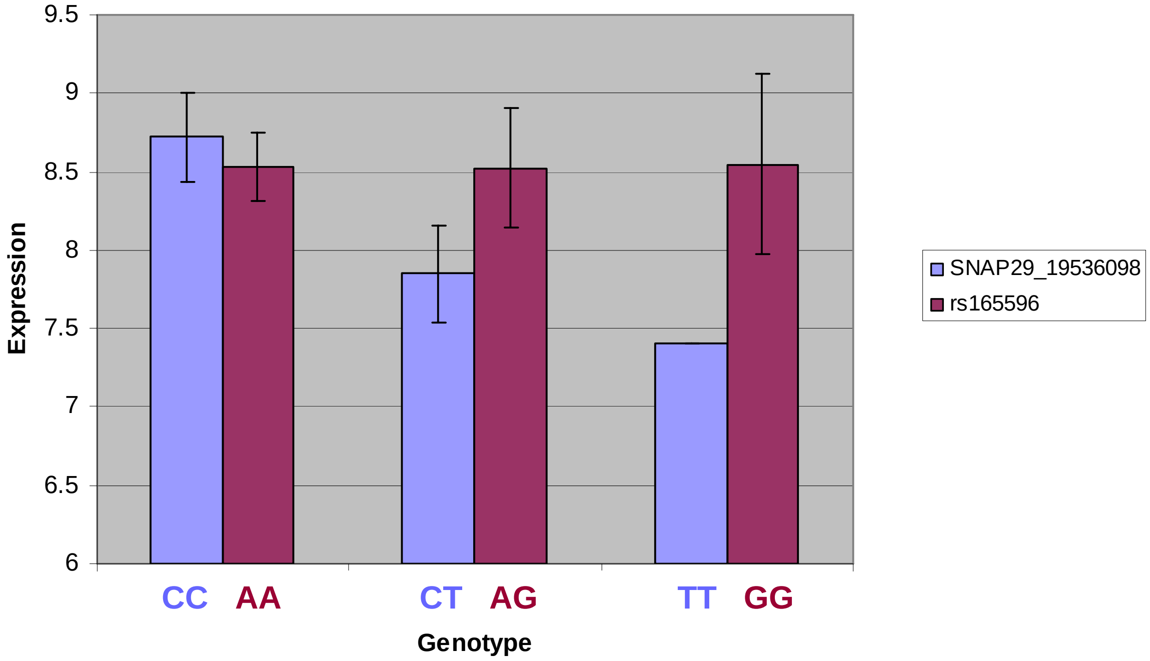


Figure 16. Association of the genotypes at the -1747 *SNAP29* SNP and rs165596 with *SNAP29* expression

3.3 Conclusions

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

The most valid way to assess the rate of promoter SNP ascertainment is to compare it to other re-sequencing projects using the same number of individuals from the same population. The only major project currently using the same 48-person panel is the Sanger Institute ExoSeq project, which aims to mine exons across the human genome for SNPs by re-sequence. While data from this project has yet to be published, they report a rate of 9.27 SNPs per kb. This is slightly under six times as high as the rate from the promoter re-sequencing. As the ExoSeq project is a long term project with a team dedicated to its completion, they were able to repeat failed PCRs or sequencing reactions, and this would increase the ascertainment rate. Although the aim of the ExoSeq project is to re-sequence exons, their primer design pipeline allows 125 bases of flanking sequence around each exon, thus including a significant amount of intron sequence. This in fact accounts for a large proportion of the SNPs discovered, and because introns are thought to be under less selective constraint than promoters, this would have driven up the number of SNPs found per kilobase relative to the promoter project. Also, exons are likely to contain far less low complexity sequence than promoters, making them easier to sequence and thus easier for ExoTrace to detect SNPs. A smaller study by T. Eades at the Sanger Institute is using this panel to re-sequence non-coding regions that are highly conserved between humans and mice. This has yielded 54 SNPs from 40 kilobases of sequence, a rate of 1 SNP per 740 bases or 1.35 SNPs per kb, somewhat lower than the rate for promoters. This is more likely due to the pre-selection of conserved sequences that will naturally contain fewer polymorphisms rather than a reflection of the relative SNP ascertainments of the two studies.

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

Comparison of the distributions of different SNP types revealed no significant difference between promoters and chromosome 22, despite the known lack of methylation at promoters which would have been proposed to influence the SNP distribution. Analysis of the rooted polymorphisms confirmed that most C/T SNPs are caused by a cytosine mutating to a thymine rather than the reverse, but still failed to show that this process happened any less frequently at promoters than in the rest of the chromosome. Restriction of the analysis to SNPs within 500 bases of the TSS, where lack of methylation is the most marked (Eckhardt et al, unpublished) did reveal a significant excess of C/G mutations, but this is more consistent with elevated GC content than with a methylation-related phenomenon. Indeed, even when only the rooted SNPs in CpG islands were analysed, the expected bias away from C to T mutations does not arise. A significant over-representation of C to A and G to T changes at the expense of A to G and T to C was observed, again consistent with elevated GC content leading to more G and C from which mutations can arise. While there was also an excess of GC SNPs in CpG islands, this fell just short of statistical significance. A possible explanation for these findings is that methyl-cytosine deamination is a relatively ancient process, dating as far back as the onset of DNA methylation in the mammalian lineage. As such, many of the methyl-cytosines in the human genome may have long since mutated to thymine and become the dominant alleles if not becoming completely fixed. As the number of CpG dinucleotides remaining in the human genome is relatively low (only 20% of the level expected), the rate of C/T SNP generation by methyl-cytosine deamination may have dropped significantly over evolutionary time. The lack of a bias away from these mutations in promoters may therefore reflect a corresponding drop in the rate of methyl-cytosine deamination in the wider genome, rather than signifying that promoters are methylated.

16.1% of the promoter SNPs in this study were found within putative regulatory elements. The precise figure is probably not meaningful, as the overall total was greatly influenced by the two TFBS databases, and the number of these elements found varies greatly with the parameters used. More importantly, there was no significant under-representation of SNPs in these elements overall. Such a bias might have been expected if the majority of these elements represented real functional sites that might be susceptible to purifying selection. Examination of individual categories showed only one with fewer SNPs than would be expected given the base coverage of the elements. However, this was the phastcons category, which is highly conserved by definition and therefore almost certain to contain fewer SNPs regardless of any functional implications. Given the equal distributions of SNPs between these elements and promoters overall, there is no sign from the SNP data alone that these elements are predictive of functional SNPs *a priori*.

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