Chapter 6

The identification and analysis of single nucleotide polymorphisms

6.1 Introduction

Genetic variation, locus specific differences in sequence between individuals, has arisen within the human population primarily through unique mutational events some time in the past. The genetic variants (polymorphisms) in present-day chromosomes reflect these historical mutational events, and analysis of them can therefore be used as a means of determining evolutionary relationships between populations and individuals.

Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation within the human genome. On this basis, comparison of any two human genomic sequences would identify approximately 2.5 million single base pair variations, at a frequency of 1 SNP every 1.25 kb (Li *et al.*, 1991, ISNPMG, 2001). A subset of these variants will cause an alteration in an encoded function, for example with respect to transcription, translation, or the structure or catalysis of a protein. More over, any of these functional changes which are deleterious to health represent a genetic contribution to a human disease state. These functional variants, and other polymorphisms which are linked to them, can be investigated to characterise genetic contributions to human disease. From a disease perspective, the sequence polymorphisms of an individual can contribute to the severity, age of onset and susceptibility to disease, or even the physiological response to the drug treatment.

The association between a polymorphism and a disease phenotype can be studied by direct or indirect approaches (reviewed by Collins, 1998). In the direct approach, a functional variant allele is tested directly for association with the phenotype under study, for example a case-control study. The frequency of a variant allele is compared in the case population and in an equivalent group of matched controls. A skew in the frequency of an allele is indicative of an association, which can then be confirmed by additional population genetic studies and functional studies. This approach depends on prior identification of putative functional variants; given the range and possible low frequency of many of these, a systematic study is required to provide the candidate functional variants for such studies. The alternative, indirect approach is carried out by testing variants (usually SNPs) selected by genomic position, for association with the phenotype in the association study. This approach depends on adequate association (i.e. linkage disequilibrium) between the SNP(s) used in the test, and the unknown functional variant, which is likely to be nearby in the genome. An association detected by the indirect approach therefore results in identification of candidate regions within the genome to target the search for the functional variant(s) which then require confirmation in the same way described above.

While large-scale studies will lead to a well-characterised panel of SNPs for genome-wide indirect association studies to search for new genes and variants involved in human disease, a more targeted small scale approach can be carried out by selecting candidate genes for SNP discovery and association studies. The identification of SNPs within coding sequence (cSNPs) in genes of medical importance, and associating coding and structural changes with gene function, will be essential to our understanding and treatment of human disease. Additional studies of gene regulatory regions may also be required to find other functional variants which alter transcriptional processes, but search for cSNPs represents a valuable first step in targeted SNP searches in candidate genes. This chapter outlines an approach to the identification of cSNPs within a family of genes localised to 1pcen – 1p13 (GSTM1-5) which has inferred association with increased cancer susceptibility. Seven other genes of medical interest, the majority of which are involved in drug metabolism, are also included in this investigation. Diseases associated with genetic variation within these additional loci include gastric (Tsukino *et al*., 2002) and lung (Nakachi *et al*., 1991) cancer as well as susceptibility to treatment by chemotherapy (Allan *et al*., 2001).

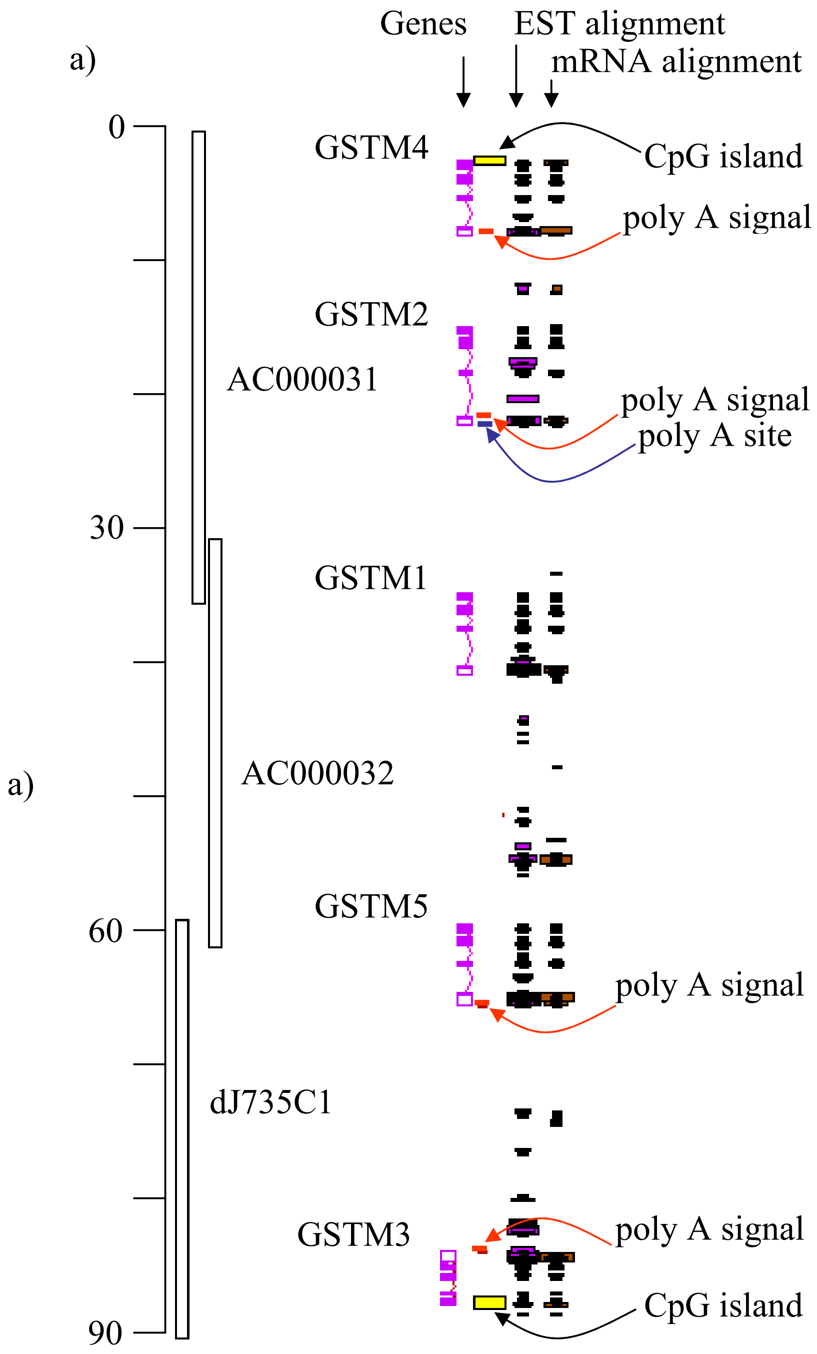
6.2 Gene Annotation

Correct annotation of coding features in their genomic context is central to the accurate design of primers for the identification of single nucleotide polymorphisms that are located within genes. The genes, GSTM 1 - 5, localised to chromosome 1p13 by mRNA / EST BLAST alignment and were annotated during the course of the full genic characterisation of 1pc – 1p13, as described in the previous chapter. The remaining genes considered in this study were either previously annotated as part of a chromosome specific sequencing project, GSTT1 on chromosome 22 (by others), or by *de novo* annotation of clone based sequence within a project specific ACeDB database. These genes, GSTP1, CYP1A1, CYP1A2, CYP2A6, CYP3A4 and NFE2L2, were localised to chromosomes 11, 15, 15, 19, 7 and 2, respectively, by BLAST alignment and the full length structure is manually annotated within clone based sequence.

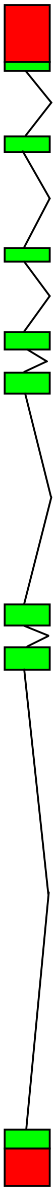
6.3 Identifying SNPs within Gene Families

To date, 69% (19626/28574) of known and predicted proteins show homology to existing proteins within Interpro (http://www.ebi.ac.uk/proteome/), suggesting a large proportion of genes in the human genome are members of a gene family. It is therefore an important consideration when designing primers to identify cSNPs that the assay will identify SNPs from the correct gene and not from a closely related family member. GSTM 1 – 5 were used as a model to assess the difficulties of designing and generating sequence from exon specific primers from closely related genes.

Glutathione S-transferases are a functionally diverse multi-gene family of soluble enzymes which are involved in the detoxification of a wide variety of chemicals via conjugation and reduction of glutathione (Booth *et al*., 1961, Mannervik and Danielson, 1988). Mammalian GSTs are divided into five main classes, alpha (A), mu (M), pi (P), theta (T) and Zeta (Z) and exist as dimeric proteins, with only subunits within the same class forming homodimers. In general, members of the same class share more that 40 – 50% sequence identity but less than 25-30% sequence identity with GSTs in other classes (Hayes and Pulford, 1995). Within the GSTM family, four of the five members (GSTM1, 2, 4 and 5) share a > 70% genomic sequence homology whilst the fifth gene, GSTM3, which is transcribed in the opposite direction, shares a low genomic sequence homology but a high percentage identity with the other four genes at the coding nucleotide (>77%) and amino acid level (>78%). All five genes are evenly distributed within a 90kb interval spanned by one PAC clone (dJ735C1) and two previously sequenced cosmids ctgm1, (AC000031) and cgtm12 (AC000032) (Xu *et al*., 1998), as displayed in figure 6.1a. The structure of all five genes is assumed to be complete as polyA signals were identified within all genes, whilst a CpG islands have been localised to the 5’ end of GSTM4 and GSTM3. The consensus GSTM gene structure contains 8 exons, the majority of which are identical in size between family members. The intron size is also conserved with introns 3, 5 and 7, showing most variation (figure 6.1b).


a)
Genes
EST alignment
mRNA alignment
0
30
60
90
GSTM4
CpG island
poly A signal
GSTM2
AC000031
poly A signal
poly A site
GSTM1
AC000032
a)
GSTM5
poly A signal
dJ735C1
GSTM3
poly A signal
CpG island

b)



	M4	M2	M1	M5	-M3
3'UTR	263	14	51	94	171
Exon1	**36**	**36**	**36**	**36**	**48**
Intron1	287	273	260	269	331
Exon2	**76**	**76**	**76**	**76**	**76**
Intron2	427	424	427	425	339
Exon3	**65**	**65**	**65**	**65**	**65**
Intron3	310	300	310	300	1062
Exon4	**82**	**82**	**82**	**82**	**82**
Intron4	100	99	95	95	93
Exon5	**101**	**101**	**101**	**101**	**101**
Intron5	941	1742	945	1185	339
Exon6	**96**	**96**	**96**	**96**	**96**
Intron6	90	90	87	87	88
Exon7	**111**	**111**	**111**	**111**	**111**
Intron7	2054	3163	2641	2096	287
Exon8	**90**	**90**	**90**	**90**	**99**
3'UTR	536	540	539	930	534

Figure 6.1: An ACeDB display of GSTM 1 – 5 and a generic GSTM gene structure. a) The annotated structures of GSTM 1 – 5 contained within two overlapping cosmids (AC000031 and AC000032) and one PAC (dJ735C1) clone. Supporting EST and mRNA sequence, in addition to poly A signals, sites and CpG islands are also shown. b) A generic GSTM gene structure with base pair sizes of exons and introns. Red boxes and green boxes represent untranslated and translated sequence, respectively.

To facilitate the identification of primer pairs that would uniquely flank GSTM exons the genomic sequence of each of the GSTM loci, including approximately 1000 bp preceding the 5‘ UTR and approximately 500 bp beyond the 3’ UTR, was aligned. Sequence was exported from 1ace in Fasta format and analysed within ClustalW (Higgins *et al*., 1994) prior to editing in GeneDoc (Nicholas *et al*., 1997) (figure 6.2).

Figure 6.2 (inserted at end of chapter 6)**:** A genomic sequence alignment of GSTM 1 – 5**:** Bases that were in common between all 5 nucleotide sequences were coloured red, 4 sequences green, 3 sequences blue and 2 or less black. 5’ and 3’ untranslated regions have been shaded yellow whilst coding sequence has been shaded grey and the corresponding exon denoted. PolyA signals at the 3’and sites, where detected, were boxed in black and red, respectively. Coloured shading (yellow, orange, dark green, pink, blue and light green) indicate the positions where primers could be designed within the genomic sequence relative to coding features.

6.4 Primer Design

Attempts were made to design exon flanking primers for each of the 91 exons contained within the 12 target genes. Primers were designed as outlined in materials and methods (section 2.16.1) but with an optimal PCR product size of 600bp. In some instances, because of sequence and product size constraints, primer pairs were designed across multiple exons (e.g. GSTM4 exons 3/4 and 6/7) or overlapped to ensure complete exon coverage (e.g. NFE2L2 exon 5a, b), as listed in table 6.1. All 5’ and 3’ UTR primer pairs were designed to incorporate as much of the untranslated region as possible. Three exons did not have working assays associated with them; primers from exon 8 of GSTM1 and GSTM2 failed primer design due to the high percentage of sequence conservation within the introns of the two genes, whilst the sense primer of exon 2 from GSTM2 failed during primer synthesis due to the enforced high GC content (80%) of the primer.

Table 6.1: A summary of the primers designed to the exons of 12 genes for the detection of coding polymorphisms. EMBL accession number associated with the primer pairs and estimated PCR product sizes are also listed.

Gene	Chr	Exon	stSG #	Size (bp)
GSTM4	1	e1	452701	528
		e2	452702	724
		e3/4	452703	566
		e5	452704	527
		e6/7	452705	701
		e8	452706	689
GSTM2	1	e1	452707	408
		e2	452708	534
		e3/4	452709	719
		e5	452710	614
		e6/7	452711	508
		e8		
GSTM1	1	e1	452712	448
		e2	452713	535
		e3	452714	484
		e4/5	452715	468
		e6/7	452716	709
		e8		
GSTM5	1	e1	452717	448
		e2	452718	548
		e3/4	452719	878
		e5	452720	409
		e6/7	452721	663
		e8	452722	621
GSTM3	1	e1	452723	520
		e2/3	452724	622
		e4/5	452725	463
		e6/7	452726	567
		e8	452727	748
GSTP1	11	e1	158595	465
		e2	158596	384
		e3	158597	385
		e3/4	158591	399
		e5	158592	399
		e6	158593	399
		e7	158594	498
GSTT1	22	e1	140015	348
		e2	140017	351
		e3	452728	276
		e4	452729	406
		e5	140020	659
CYP1A1	15	e1a	452730	556
		e1b	452731	533
		e2/3/4	452732	564
		e5/6	452733	637
		3' UTRa	452734	620
		3' UTRb	452735	697
CYP1A2	15	e1a	452736	700
		e1b	452737	536
		e2/3	452738	753
		e4	452739	511
		e5	452740	484
		e6	452741	464
CYP2A6	19	e1/2	452742	722
		e3/4	452743	626
		e5	452744	588
		e6	452745	404
		e7	452746	467
		e8	452747	443
		e9	452748	538
CYP3A4	7	e1	452755	613
		e2	452756	519
		e3	452757	510
		e4	452758	549
		e5/6	452759	589
		e7	452760	592
		e8	452761	476
		e9	452762	512
		e10	452763	523
		e11	452764	654
		e12	452765	419
		e13a	452766	669
		e13b	452767	791
NFE2L2	2	e1	452768	589
		e2	452769	419
		e3	452770	577
		e4	452771	527
		e5a	452772	649
		e5b	452773	695
		3’ UTR	452774	756

6.5 DNA screening

DNA from the Coriell CEPH/Utah reference family collection (http://locus.umdnj.edu/ccr) was used as template for the generation of exon specific PCR products. A summary of the 47 DNAs used, which represent the founders of the collection, are listed in table 6.2. The pedigree from which one of the DNA originated, CEPH/Utah pedigree 1333 (red underline), is shown in figure 6.3. All individuals used in this study are Caucasian and of Northern European extraction.

Table 6.2: A summary of the CEPH/Utah DNA used for the generation of exon sequence. The maternal grandmother DNA listed in red was not used in this study.

Family ID	Paternal Grandfather	Paternal Grandmother	Maternal Grandfather	Maternal Grandmother
1331	NA07007	NA07340	NA07016	NA07050
1333	NA07049	NA07002	NA07017	NA07341
1341	NA07034	NA07055	NA06993	NA06985
1346	NA12043	NA12044	NA12045	NA12046
1347	NA11879	NA11880	NA11881	NA11882
1362	NA11992	NA11993	NA11994	NA11995
1408	NA12154	NA12236	NA12155	NA12156
1416	NA12248	NA12249	NA12250	NA12251
1423	NA11917	NA11918	NA11919	NA11920
1334	NA12144	NA12145	NA12146	NA12239
1340	NA06994	NA07000	NA07022	NA07056
1420	NA12003	NA12004	NA12005	NA12006

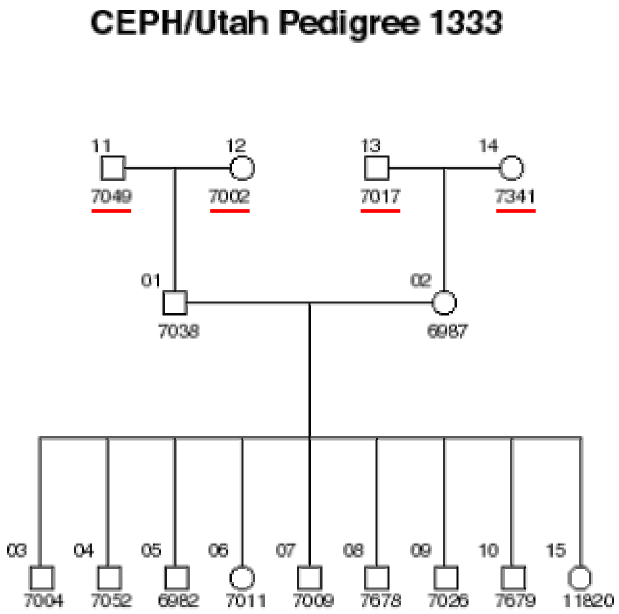


Figure 6.3: A CEPH pedigree. Pedigree represents four of the DNAs used (red underlines) in the *de novo* generation of cSNPs from the CEPH/Utah family collection.

The plate format of the exon specific – CEPH DNA PCR reactions was designed so that all primer pairs were screened across four CEPH DNAs within one 384 well microtitre plate. An aliquot of PCR products from each well of the first, DNAs 1-4, and second plates, DNAs 5-8, were separated by gel electrophoresis to determine the success rate of PCR reactions prior to sequencing (by others). Figure 6.4 shows an example of the resultant PCR products separated by gel electrophoresis, generated at a Tm of 60^{o}C from CEPH DNAs 5-8, of primer pairs designed to exons 1, 2, 3 and 4 of GSTM4. Product sizes of the three exon specific primers were shown to correspond with their estimated size from genomic sequence, 528 bp, 744 bp and 526 bp respectively, but variation in intensity and multiplicity of bands was observed.

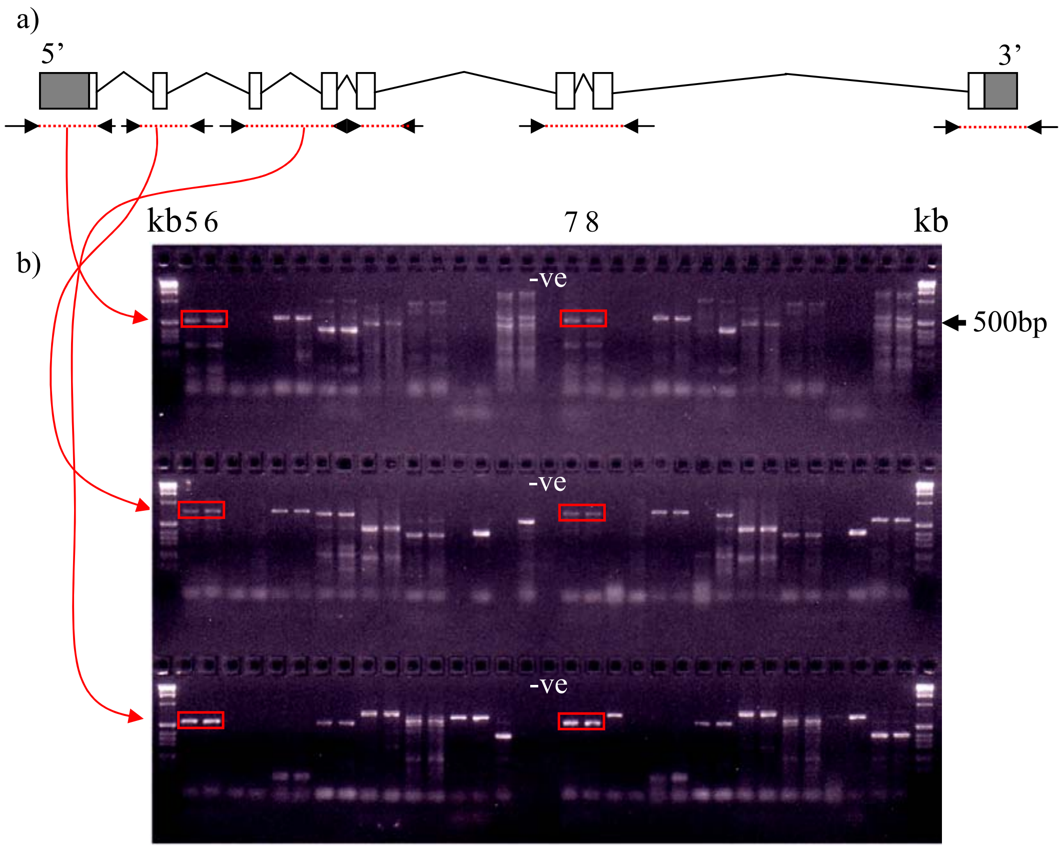


Figure 6.4: Screening of CEPH DNAs with exon specific primer pairs designed to GSTM4. a) Exon specific primer pairs were designed to an annotated gene structure of GSTM4. b) Red arrows indicate the amplification of a 500 bp product from CEPH DNAs 5 – 8 from exons1, 2, 3 and 4 at 60°C. A size marker was included to size PCR products (kb).

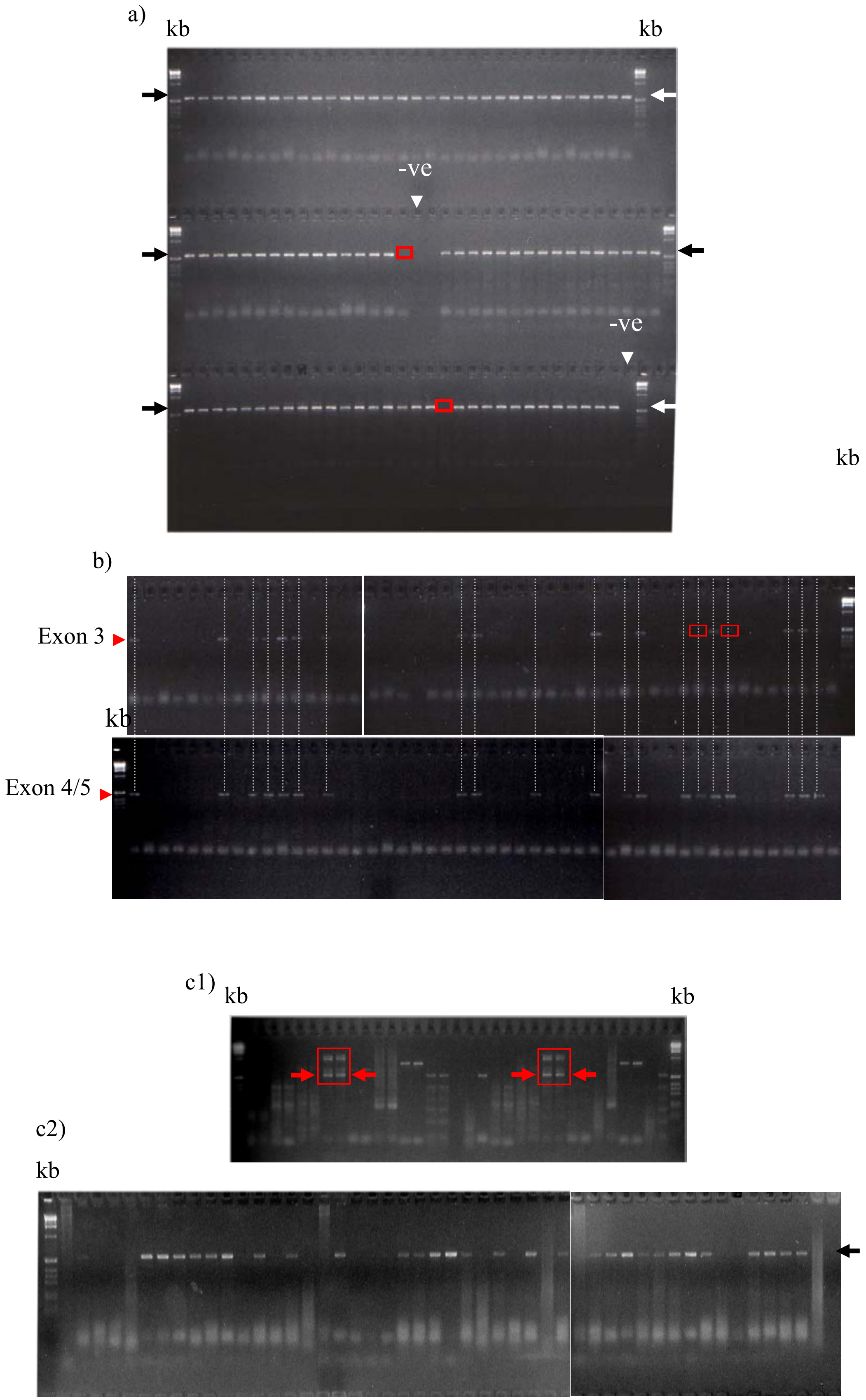
Initial collation of PCR results from electrophoresis of the first two plates (CEPH DNAs 1-8) indicated that well locations C7 – G7 from the primer microtitre plate (exon 5 – GSTP1, 3' UTRa – CYP1A1, exon 7 – CYP2A6, exon 4 – CYP3A4 and exon 3 – NFE2L2, respectively) consistently failed to generate a PCR product. This error was systematic (caused by blocked pins on a liquid handling robot and not through any primer design or DNA template problem) and the missing PCR reactions were repeated. In this instance, each of the 47 CEPH DNAs was tested with the missing primer pairs, in addition to two working GSTM1 primers. Repetition of 158592, 452734, 452746, 452758 and 452770 resulted in three of the five PCR reactions yielding strong single bands following electrophoresis, whilst 158592 generated multiple bands and 452734 failed to generate any product. Products from the repeated round of PCR reactions were sequenced (by others) with those generated from the initial experiment (figure 6.5a).

GSTM1 is known to be a null allele within 50% of the population (Board *et al*., 1990); therefore parallel generation of PCR products from all DNAs would permit an estimation of the percentage of CEPH individuals that miss both copies of the gene. Whilst the intensity of the PCR products when assessed by agarose gel electrophoresis was faint (figure 6.5b), a reproducible relationship could be observed between the DNAs that produced a band from exon 3 and exon 4 / 5 specific primers, thereby confirming the presence of the gene in these individuals. Eighteen of the forty seven DNAs (38%) generated a PCR product for the aforementioned primers whilst a further two DNAs yielded products for one of the primer pairs. The lower observed occurrence of null alleles within the sample set (38% observed compared to 50% expected) may be due to a population bias for the occurrence of the null status of the gene within the

Northern European Caucasian population or may be due to the failure of the faint PCR products to be detected from the electrophoresis gels.

To investigate whether increasing the annealing temperature of primers during the PCR reaction would obviate the production of multiple products, six pairs of primers which generated multiple bands in the first experiment (158592, 452712, 452717, 452731, 452764 and 452767) were tested across each of the 47 CEPH DNAs at 65^{o}C. Two of the six re-tested primer pairs yielded a strong single band following electrophoresis, whilst the remainder failed to generate any product. One of the two successful primer pairs that worked at the more stringent annealing temperature, 453731 covering the 3’ end of CYP1A1 exon one, is shown in figure 6.5c. The use of a higher annealing temperatures in PCR reactions as a ‘clean-up step’ has subsequently been adopted by the Sanger Institute project involved with the large scale production of exon specific sequences.

Figure 6.5 (see over)**:** PCR products from exon primers using CEPH DNA as template. a) PCR primers from exon 4 of CYP3A4 (452758) and exon 3 of NFE2L2 (452770) screened across 47 CEPH DNAs at an annealing temperature of 60^{o}C. Included are size markers (kb), expected PCR product sizes (black arrows) and negative controls ($T_{0.1}E$). DNAs for which no PCR product was generated at boxed in red. b) PCR products generated by primers designed to exons 3 and 4 / 5 and screened across 47 CEPH DNAs generated at an anneling temperature of 60^{o}C. Dotted grey lines join CEPH DNAs that are thought to contain the GSTM1 gene. DNAs for which no PCR product was generated at boxed in red. c1) Initial screening of CEPH DNAs produces multiple bands at 60^{o}C (red box). Increasing the stringency of primer annealing temperature to 65^{o}C c2) results in the generation of a single PCR product (black arrows).


a)
kb
kb
-ve
-ve
kb
b)
Exon 3
kb
Exon 4/5
c1)
kb
kb
c2)
kb

Analysis of the PCR reactions from the 77 primer pairs, covering 91 coding exons of the 12 target genes listed in table 6.3, revealed that 71 (92%) of primers generated bands from at least one of the eight CEPH DNAs when products were separated by gel electrophoresis. The number of working PCR assays was further reduced when primers that produced multiple bands, without the presence of a single band in one of the eight DNAs, were subtracted. These 62 assays (81% of the total) contained two primer pairs (CYP1A1 – exon 1b and GSTM5 – exon 1) for which the products were generated at a primer annealing temperature of 65°C.

Table 6.3: A summary of exon specific PCR reaction results using CEPH DNAs 1 – 8 as a template. GSTM1 – 5 coloured shading corresponding to primers in figure 6.2, the presence ‘Y’ (yellow shaded if there were multiple bands) or absence ‘N’ of PCR products upon gel electrophoresis is listed under the CEPH DNA columns. Pink shading indicates products generated at an annealing temperature of 65°C. Grey box, CEPH DNA 3, indicates possible degradation due to high failure rate.

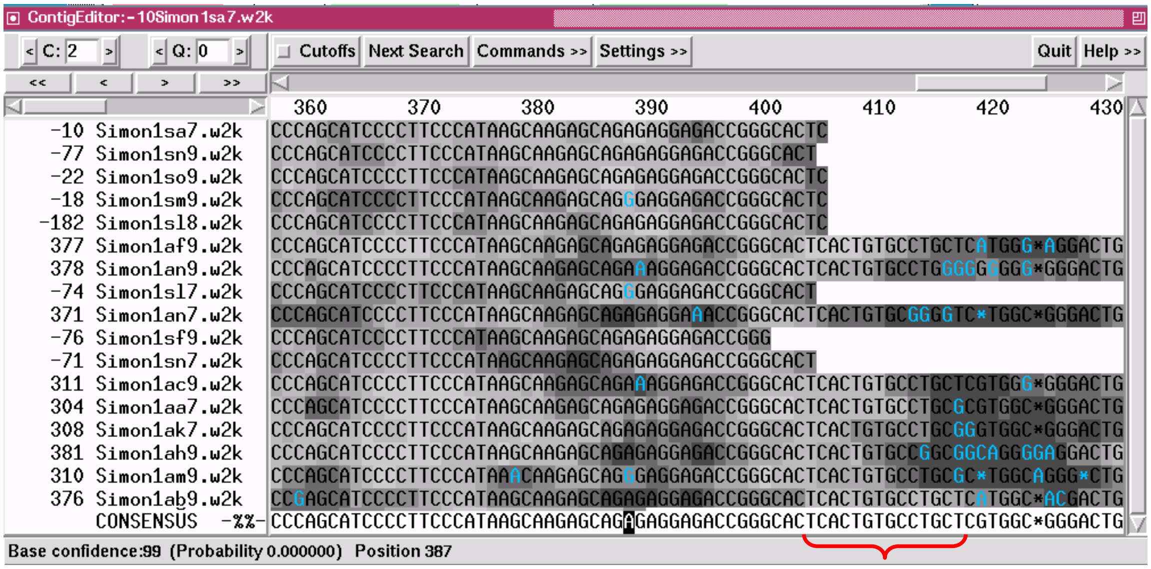
		CEPH DNAs										CEPH DNAs							
Target	**Exon**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Target**	**Exon**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
GSTM4	e1	Y	Y	N	N	Y	Y	Y	Y	**CYP1A1**	e1a	Y	Y	Y	Y	Y	Y	Y	Y
	e2	N	N	N	Y	Y	Y	Y	Y		e1b	Y	Y	Y	Y	Y	Y	Y	Y
	e3/4	N	N	Y	Y	Y	Y	Y	Y		e2/3/4	Y	Y	Y	Y	Y	Y	Y	Y
	e5	N	N	N	Y	Y	Y	N	Y		e5/6	N	N	N	N	N	N	N	N
	e6/7	N	N	N	Y	Y	Y	Y	Y		3' UTRa	N	N	N	N	N	N	N	N
	e8	N	N	N	Y	Y	Y	Y	Y		3' UTRb	Y	Y	Y	Y	Y	Y	Y	Y
GSTM2	e1	N	N	N	Y	Y	Y	Y	Y	**CYP1A2**	e1a	Y	Y	N	Y	Y	Y	Y	Y
	e2										e1b	Y	Y	N	Y	Y	Y	Y	Y
	e3/4	N	N	N	Y	Y	Y	Y	Y		e2/3	Y	Y	N	Y	Y	Y	Y	Y
	e5	N	N	N	Y	Y	Y	Y	N		e4	Y	N	Y	N	Y	N	Y	N
	e6/7	N	N	Y	Y	Y	Y	Y	Y		e5	Y	Y	N	Y	Y	Y	Y	Y
	e8										e6	Y	Y	Y	Y	Y	Y	Y	Y
GSTM1	e1	Y	N	N	N	N	N	Y	Y	**CYP2A6**	e1/2	Y	Y	Y	Y	Y	Y	Y	Y

	e2	Y	N	N	N	Y	N	Y	N		e3/4	Y	Y	N	Y	Y	Y	Y	Y
	e3	Y	N	N	N	N	N	Y	N		e5	Y	Y	N	Y	Y	Y	Y	Y
	e4/5	Y	N	N	N	N	N	Y	N		e6	Y	Y	Y	Y	Y	Y	Y	Y
	e6/7	Y	N	N	N	N	N	Y	N		e7	Y	Y	Y	Y	Y	Y	Y	Y
	e8										e8	Y	Y	N	Y	Y	Y	Y	Y
											e9	Y	Y	Y	Y	Y	Y	Y	Y
GSTM5	e1	N	N	Y	Y	N	N	Y	Y										
	e2	Y	Y	Y	Y	Y	Y	Y	Y	**CYP3A4**	e1	N	N	N	N	N	N	N	N
	e3/4	Y	Y	N	Y	Y	Y	Y	Y		e2	Y	Y	Y	Y	Y	Y	Y	Y
	e5	Y	Y	N	N	Y	Y	N	N		e3	Y	Y	Y	Y	Y	Y	Y	Y
	e6/7	Y	Y	N	Y	Y	Y	Y	Y		e4	Y	Y	Y	Y	Y	Y	Y	Y
	e8	Y	Y	N	Y	Y	Y	Y	Y		e5/6	Y	Y	N	Y	Y	Y	Y	Y
											e7	Y	Y	N	Y	Y	Y	Y	Y
GSTM3	e1	Y	Y	Y	Y	Y	Y	Y	Y		e8	Y	Y	Y	Y	Y	Y	Y	Y
	e2/3	Y	Y	N	Y	Y	Y	Y	Y		e9	Y	Y	Y	Y	Y	Y	Y	Y
	e4/5	Y	N	N	N	Y	N	N	N		e10	Y	N	Y	N	Y	N	Y	N
	e6/7	Y	Y	Y	Y	Y	Y	Y	Y		e11	N	N	N	N	N	N	N	N
	e8	Y	Y	N	Y	Y	Y	Y	Y		e12	N	Y	Y	Y	Y	Y	N	Y
											e13a	Y	Y	N	Y	Y	Y	Y	Y
GSTP1	e1	N	N	N	N	N	N	N	N		e13b	Y	Y	Y	Y	Y	Y	N	Y
	e2	N	N	N	N	N	N	Y	Y										
	e3	Y	Y	Y	Y	Y	Y	Y	Y	**NFE2L2**	e1	N	Y	N	N	N	Y	N	Y
	e3/4	N	N	N	N	N	N	N	N		e2	Y	Y	N	Y	Y	Y	Y	Y
	e5	Y	Y	N	N	Y	Y	N	N		e3	Y	Y	Y	Y	Y	Y	Y	Y
	e6	Y	Y	N	N	Y	Y	Y	Y		e4	Y	Y	N	Y	Y	Y	Y	Y
	e7	Y	Y	N	Y	Y	Y	Y	Y		e5a	Y	Y	Y	Y	Y	Y	Y	Y
											e5b	Y	Y	N	Y	Y	Y	Y	Y
GSTT1	e1	Y	Y	N	Y	Y	Y	N	Y		3' UTR	Y	Y	Y	Y	Y	Y	Y	Y
	e2	Y	Y	N	Y	Y	Y	N	Y										
	e3	Y	N	Y	N	Y	N	N	Y										
	e4	Y	Y	Y	Y	Y	Y	N	Y										
	e5	Y	Y	Y	Y	Y	Y	N	Y										

6.6 Sequence generation and assembly

Dye-terminator sequencing, using exon specific primers and PCR products is described in section 6.5, was carried out by Sarah Lindsay and the high throughput sequencing facility at the Sanger Institute. PCR products were sequenced and then run on ABI 3700 and 3730 sequencing machines. Bases within sequence reads were ‘called’ using Phred (Ewing *et al*., 1998) and assembled within Phrap (Ewing *et al*., 1998) utilising, amongst other criteria, Phred assigned base-quality values (Q-values). Sequence reads were assembled and imported into gene specific directories by Sarah Hunt. Vector masked and quality clipped sequences reads, together with the consensus sequence, were viewed within gene specific Gap4 (Bonfield *et al*., 1995) databases.

Genomic coverage of the assembled exon specific sequence contigs was determined within the respective ACeDB databases by exact alignment of 10 – 15 good quality bases (light sequences within the assembled contig, figure 6.6) from either end of a sequence assembly. The number of reads contained within an assembly was determined to be the total number of reads contained within a sequence contig.

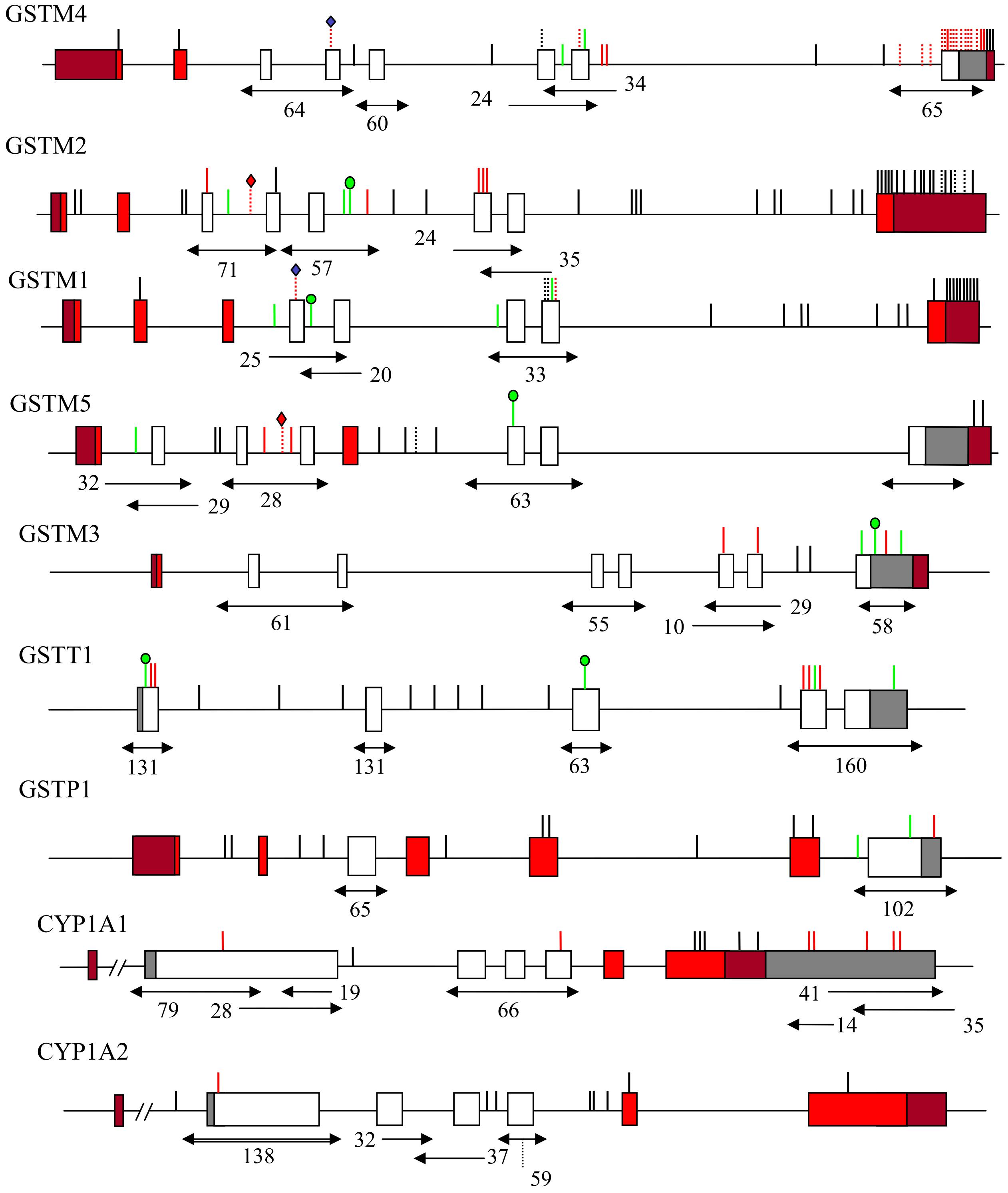


```
 -10 Simon1sa7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTC
 -77 Simon1sn9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACT
 -22 Simon1so9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTC
 -18 Simon1sm9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGGGAGGAGACCGGGCACTC
-182 Simon1sl8.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTC
 377 Simon1af9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCTGCTCATGGG*AGGACTG
 378 Simon1an9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAAAGGAGACCGGGCACTCACTGTGCCTGGGGGGGG*GGGACTG
 -74 Simon1sl7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGGGAGGAGACCGGGCACT
 371 Simon1an7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAAACCGGGCACTCACTGTGCGGGTC*TGGC*GGGACTG
 -76 Simon1sf9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGG
 -71 Simon1sn7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACT
 311 Simon1ac9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAAAGGAGACCGGGCACTCACTGTGCCTGCTCGTGGG*GGGACTG
 304 Simon1aa7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCTGCGCGTGGC*GGGACTG
 308 Simon1ak7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCTGCGGTGGC*GGGACTG
 381 Simon1ah9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCGGCGGCAGGGAGGACTG
 310 Simon1am9.w2k  CCCAGCATCCCCTTCCCATAACCAAGAGCAGGGAGGAGACCGGGCACTCACTGTGCCTGCGC*TGGCAGGG*CTG
 376 Simon1ab9.w2k  CCGAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCTGCTCATGGC*ACGACTG
     CONSENSUS  -%%- CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCTGCTCGTGGC*GGGACTG
Base confidence:99 (Probability 0.000000) Position 387
```

Figure 6.6: Assembly of *de novo* exon specific sequences shown in Gap4. Red brackets indicate the consensus sequence that was used to determine the extent of sequence coverage by alignment within 1ace.

6.7 Exon coverage of sequence contigs

A minimum of ten good quality sequence reads, including both sense and anti-sense reads, was used as the criteria for determining whether *de novo* sequence had successfully been generated to identify known and novel SNPs within an exon. Fifty six of the 77 initial primer pairs (73%) provided data which could be assembled into sequence contigs using the above criteria. This represents 79% of primer pairs excluding those for which PCR assays failed to produce a product. These 56 sequence contigs account for 67 of the 91 exons (74%) from the 12 target genes. Sequences derived from 17 (30%) of the PCRs were assembled separately as groups of sense or anti-sense reads. These were paired up ad counted as a single contig when calculating coverage figures.


GSTM4
64
60
24
34
65
GSTM2
71
57
24
35
GSTM1
25
20
33
GSTM5
32
29
28
63
GSTM3
61
55
10
29
58
GSTT1
131
131
63
160
GSTP1
65
102
CYP1A1
79
28
19
66
41
14
35
CYP1A2
138
32
37
59

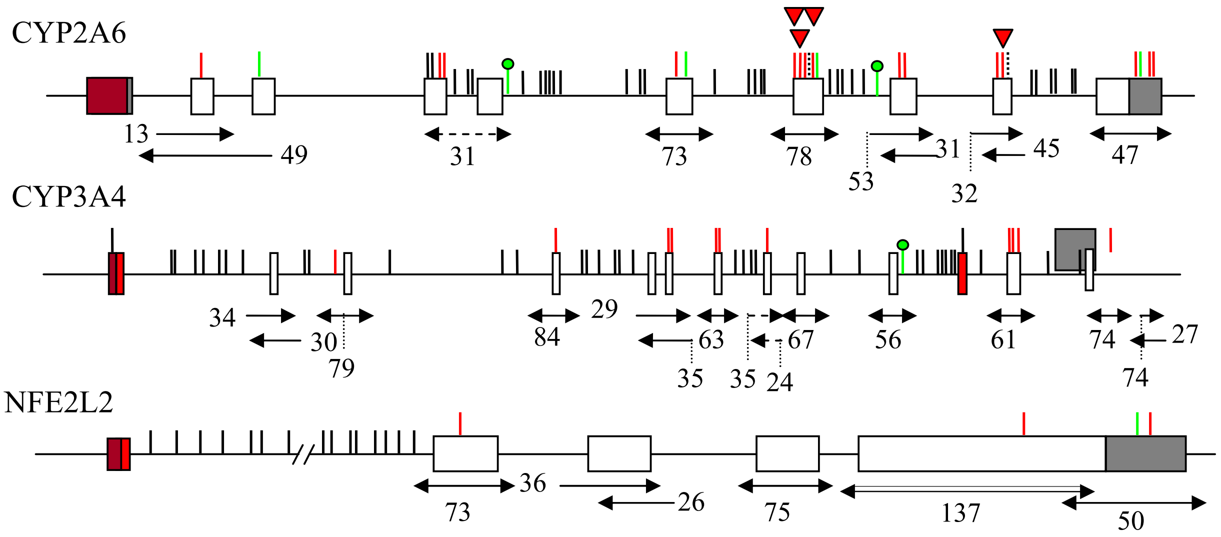
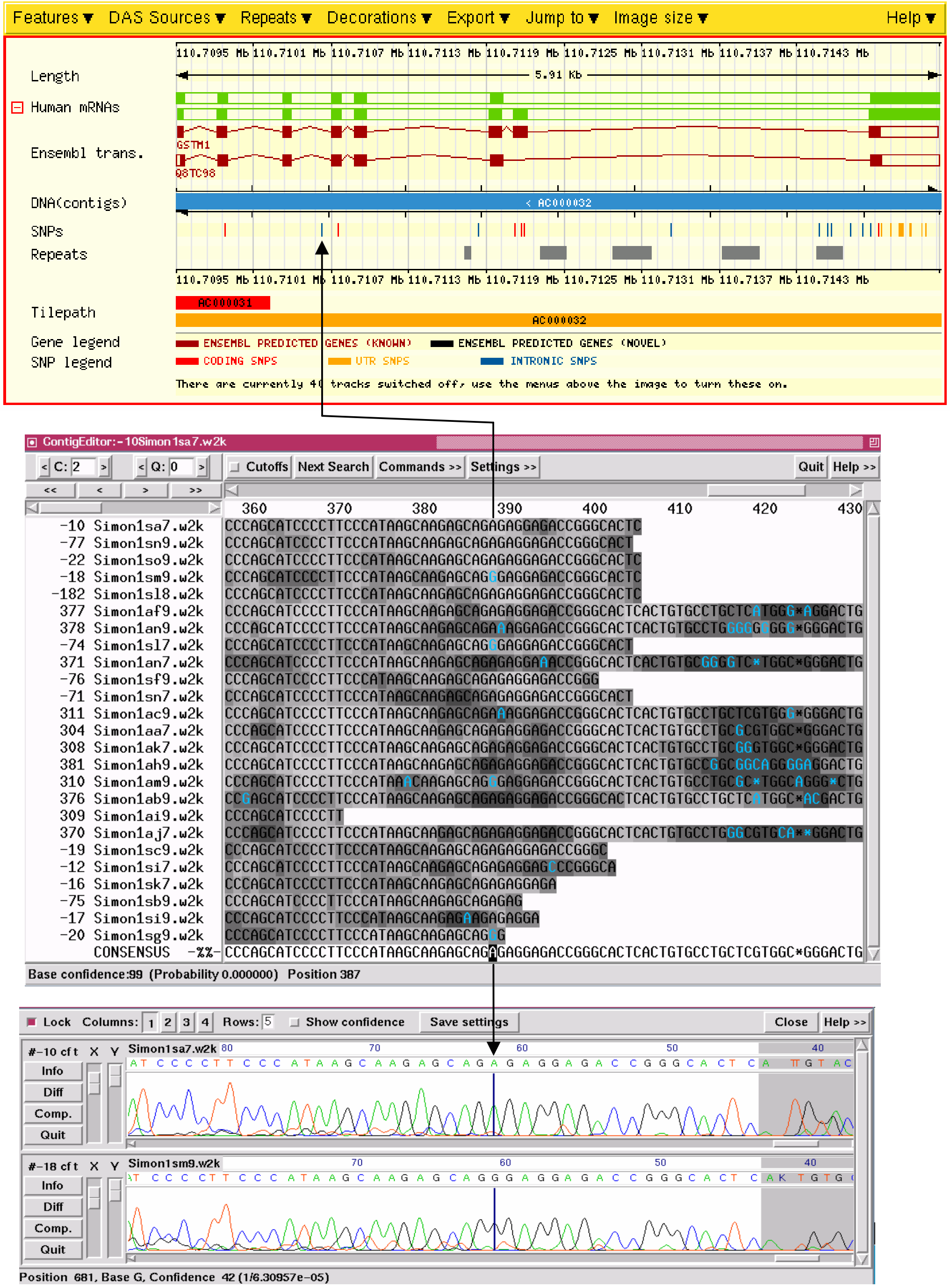


Figure 6.7: A summary representation of the *de novo* sequence coverage of exons from the target genes. Clear and grey boxes represent *de novo* sequence coverage of exons and untranslated region, respectively (not drawn to scale); exons and UTR without sequence coverage are red and dark red, respectively. Double headed arrows indicate sequence from both primer pairs assembled into one contig (number indicates the total number of reads assembled); single headed arrows specify assembly of sequence from one primer; double lined arrows are where four primers have been assembled. Black vertical lines represent known SNPs not covered by sequence coverage; black dotted vertical lines indicate where the flanking sequence from the known SNP did not align to genomic sequence; red vertical lines are known SNPs not identified by this study; vertical red dotted lines are known SNPs with multiple loci within genomic sequence (red and blue diamonds denote where extended genomic flanking sequence alignment was required to localise the SNP); green vertical lines are known SNPs found within *de novo* sequence; green vertical lines with round heads are novel SNPs identified here.

6.7.1 Validation and localisation of known SNPs

One hundred and three (35%) of the two hundred and ninety one known SNPs, which are derived from dbSNP (http://www.ncbi.nlm.nih.gov/SNP/index.html URL), and localised to the 12 genes of interest within Ensembl (http://www.ensembl.org/), are covered by the *de novo* sequence contigs (figure 6.8a). To assess whether these SNPs were present within annotated gene structures the known SNPs, and their flanking sequence, were aligned to the genomic sequence within respective ACeDB databases. This alignment identified 17 SNPs which are primarily found within GSTM4, but also localised to multiple genes and 8 SNPs for which the submitted sequence could not be found when carrying out the alignment using 15-20 bp either side of the SNP, including each allele in the search.

The remaining 76 known SNPs were localised along with their flanking sequence within the *de novo* sequence contigs by sequence matching within respective Gap4 databases. Each read of the assembled sequence contigs were then checked manually for the presence or absence of the relevant SNP allele, in either homozygous or heterozygous form (figure 6.8c). An important consideration when identifying known SNPs within the sequence assembly was the orientation of the sequence contig relative to the forward or reverse read data associated with the submitted SNP. The majority, 59 (78%), of the 76 uniquely localised known SNPs were not found within the CEPH DNAs tested here. Of the remaining 17 SNPs (22%), 8 were determined to be located within coding sequence, and 9 were located within non-coding sequence.



```
                 360       370       380       390       400       410       420       430
   -10 Simon1sa7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTC
   -77 Simon1sn9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACT
   -22 Simon1so9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTC
   -18 Simon1sm9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGGGAGGAGACCGGGCACTC
  -182 Simon1sl8.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTC
   377 Simon1af9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCTGCTCATGGG*AGGACTG
   378 Simon1an9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAAAGGAGACCGGGCACTCACTGTGCCTGGGGGGGG*GGGACTG
   -74 Simon1sl7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGGGAGGAGACCGGGCACT
   371 Simon1an7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAAACCGGGCACTCACTGTGCGGGTC*TGGC*GGGACTG
   -76 Simon1sf9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGG
   -71 Simon1sn7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACT
   311 Simon1ac9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAAAGGAGACCGGGCACTCACTGTGCCTGCTCGTGGG*GGGACTG
   304 Simon1aa7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCTGCGCGTGGC*GGGACTG
   308 Simon1ak7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCTGCGGGTGGC*GGGACTG
   381 Simon1ah9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCGGCGGCAGGGAGGACTG
   310 Simon1am9.w2k  CCCAGCATCCCCTTCCCATAAACAAGAGCAGGGAGGAGACCGGGCACTCACTGTGCCTGCGC*TGGCAGGG*CTG
   376 Simon1ab9.w2k  CCGAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCTGCTCATGGC*ACGACTG
   309 Simon1ai9.w2k  CCCAGCATCCCCTT
   370 Simon1aj7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCTGGGCGTGCA**GGACTG
   -19 Simon1sc9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGC
   -12 Simon1si7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGCCCGGGCA
   -16 Simon1sk7.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGA
   -75 Simon1sb9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAG
   -17 Simon1si9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGAAGAGAGGA
   -20 Simon1sg9.w2k  CCCAGCATCCCCTTCCCATAAGCAAGAGCAGGG
       CONSENSUS  -%%-CCCAGCATCCCCTTCCCATAAGCAAGAGCAGAGAGGAGACCGGGCACTCACTGTGCCTGCTCGTGGC*GGGACTG
Base confidence:99  (Probability 0.000000)  Position 387
```

```
#-10 cf t  X  Y  Simon1sa7.w2k 80        70        60        50        40
AT C C C C T T C C C A T A A G C A A G A G C A G A G A G G A G A C C G G G C A C T C A TT G T A C

#-18 cf t  X  Y  Simon1sm9.w2k           70        60        50        40
AT C C C C T T C C C A T A A G C A A G A G C A G G G A G G A G A C C G G G C A C T C A K T G T G
Position 681, Base G, Confidence 42 (1/6.30957e-05)
```

Figure 6.8: Identification of a known SNP which is within intron 3 of GSTM1. Represented is the known SNP within Ensembl, a), the specific nucleotide position within a sequence assembly displayed in a Gap4 database, b), and the homozygous sequence reads for each allele.

6.7.2 Identification of novel SNPs

To more readily identify novel SNPs within assembled Gap4 sequence contigs the quality value of the reads was converted into grey-scale and the ambiguous base calls, and potential novel SNPs, were coloured blue (figure 6.8b). Ambiguous bases were then checked manually within all traces. A putative novel SNP was identified when the base containing the homozygous minor allele had a quality value of 30 (Q-30) or the heterozygote, containing both alleles, was surrounded by bases with > Q-30 quality values. Nine novel SNPs, three of which were located within coding sequence, were identified within the 12 target genes. Additional support for a novel SNP was obtained by checking the reverse read for each read pair. A breakdown of known and novel SNP numbers are listed in table 4 whilst a summary of the placement of all SNPs within annotated gene structures can be found in figure 6.7.

Table 6.4: A summary of the known and novel SNPs associated with 12 target genes.

Gene	Known SNPs	With Seq Coverage	Uniquely Assigned	Align to Genomic	Not Observed	Observed	Coding	Novel	Coding
GSTM4	32	23	8	7	5	2	1	0	-
GSTM2	43	8	7	7	5	2	-	1	-
GSTM1	26	7	5	3	-	3	1	1	-
GSTM5	12	4	4	3	2	1	-	1	1
GSTM3	7	5	4	4	3	1	1	1	-
GSTP1	13	3	3	3	1	2	1	0	
GSTT1	16	7	7	7	5	2	1	2	2
CYP1A1	13	7	7	7	7	0	-	0	-
CYP1A2	9	1	1	1	1	0	-	0	-
CYP2A6	54	23	23	19	16	3	3	2	-
CYP3A4	47	11	11	11	11	0	-	1	-
NFE2L2	19	4	4	4	3	1	-	0	-
TOTAL	**291**	**103**	**84**	**76**	**59**	**17**	**8**	**9**	**3**

6.8 SNP Analysis

Under conditions of random mutation it is expected that half as many nucleotide transitions occur (purine→ purine or pyrimidine→pyrimidine) as transversions (pyrimidine↔ purine), table 6.5 (Dawson *et al*., 2001). In addition, as the strand on which the change has occurred is unknown, no distinction can be made between A↔G and C↔T and between C↔A and G↔T. Analysis of 28 SNPs contained within this data set revealed that 19 transitions (68%) occur approximately twice as often as 9 transversions (32%) (table 6.6). The observed contradiction to the expected frequency has previously been reported (Horton *et al*., 1998; Dawson *et al*., 2001; Halushka *et al*., 1999, Deutsch *et al*., 2001). A possible reason for the occurrence of an increased number of C↔T (G↔A) transversions may be due to the deamination of 5-methylcytosine that occurs frequently at CpG dinucleotides.

Table 6.5: Expected occurrence of transitions and transversions in genomic sequence.

	A	**T**	**C**	**G**
A	-			
T	Transversion	-		
C	Transversion	Transition	-	
G	Transition	Transversion	Transversion	-

Table 6.6: The observed number of transitions and transversions of known and novel SNPs within the 12 target genes.

Variation	**Number**	**Percentage**
C↔T (G↔A)	19	68.00%
Transitions	**19**	**68.00%**
C↔A (T↔G)	6	21.00%
C↔G	3	11.00%
A↔T	-	-
Transversions	**9**	**32.00%**

Of the 28 detected novel and known SNPs, 12 were localised to introns, 5 to 3' untranslated regions, whilst 11 occurred within exons (table 6.7). Seven of the 11 coding SNPs resulted in synonymous changes, i.e. maintaining amino acid type by codon usage, whilst the remaining four SNPs resulted in an amino acid change. Each of the four amino acid changes were caused by substitutions in the first base of the codon, two of which resulted in a conservative change (maintaining a similar amino acid type, for example valine to an isoleucine change maintains an aliphatic, hydrophobic amino acid) and the remaining two changes were non-conservative changes.

Table 6.7: Exonic SNP analysis.

Gene	Exon	Position	Allele	change	Codon	change	Encoded	a. a.	Synon	Cons
GSTM4	7	78	C	↔ T	TTT	↔ TTC	phe	↔ phe	yes	-
GSTM1	7	72	C	↔ T	GAC	↔ GAT	asp	↔ asp	yes	-
GSTM5	6	22	T	↔ C	TTG	↔ CTG	leu	↔ leu	yes	-
GSTM3	8	91	G	↔ A	GTA	↔ ATA	val	↔ iso	no	yes
GSTT1	1	13	C	↔ T	CTG	↔ TTG	leu	↔ leu	yes	-
GSTT1	3	110	A	↔ C	ACG	↔ CCG	thr	↔ pro	no	no
GSTT1	4	155	G	↔ A	GTA	↔ ATA	val	↔ iso	no	yes
GSTP1	7	111	A	↔ C	AGT	↔ ACT	ser	↔ ser	yes	-
CYP2A6	2	37	T	↔ C	TTG	↔ CTG	leu	↔ leu	yes	-
CYP2A6	5	117	C	↔ T	CGC	↔ CGT	arg	↔ arg	yes	-
CYP2A6	6	100	C	↔ T	CGC	↔ TGC	arg	↔ cys	no	no

6.8.1 Validating SNPs within highly homologous genes

One of the difficulties encountered whilst identifying valid SNPs associated with GSTM 1- 5 paralogues was to determine whether a SNP was valid (i.e. two allelic variants at the same locus) or merely the alignment of two locus specific base pair variants, one from each of two closely related gene family members. A comparison between the known SNPs localised to GSTM 1 – 5 and the number which were uniquely placed (columns 2 and 4 in table 6.4) indicates the scale of the problem that was encountered, i.e. 65% of GSTM1 SNPs that had *de novo* sequence coverage localised to other GSTM loci. Three criteria were used to assist with the elucidation of a valid candidate SNP from locus specific variants of paralogous genes; 1) the SNP must be found within *de novo* sequence traces, 2) it must be in Hardy-Weinberg (H-W) equilibrium (i.e. allele 1 = p and allele 2 = q, $p^2 + 2pq + q^2 = 1$) and 3) the alternative allele should not be present as a base variant in a paralogous gene. Valid SNPs were divided into 'known', i.e. those that have a dbSNP entry, and 'novel', i.e. those identified in this study. 'Suspect candidate' SNPs fulfilled criteria 1) and 2) but the alternate SNP allele was be present within a homologous gene. 'Rejected candidates' are not observed in sequence traces (due to BLAST alignment of the flanking sequence of a SNP from a closely related gene family member), the alternative allele was present within paralogues and it was not in H-W equilibrium. The observed occurrence and genic location of known, novel, suspect and rejected GSTM 1 – 5 SNPs is shown in table 8.

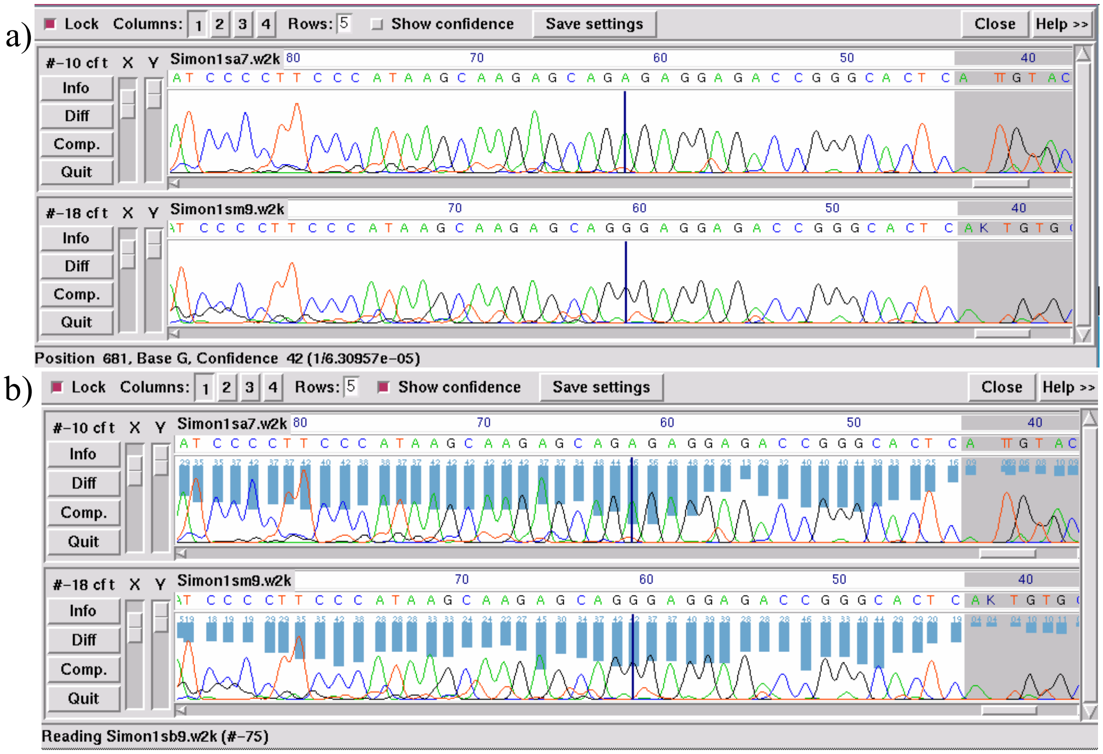
Table 6.8: Summary of categorised GSTM 1 – 5 SNPs.

Status	Exon	Intron	3' UTR	Total
Known	1	4	2	**7**
Novel	-	3	-	**3**
Suspect	3	1	-	**4**
Rejected	6	3	9	**18**

6.8.2 Validating SNPs

6.8.2.1 Known SNPs

The presence of a valid known SNP was easily detected once its location within the *de novo* sequence contig had been identified using flanking sequence matches within the respective Gap4 database. The alternative minor allele, even when present as a heterozygote, was clearly discernable within good quality sequence. An addition example of a known SNP, dbSNP: 737497, is shown in figure 6.9. The SNP was found within the *de novo* sequence as a homozygote in major and minor alleles and localises to intron 3 within GSTM1. The T233C SNP was submitted to dbSNP by The SNP Consortium (TSC). The orientation of the flanking sequence had been designated as a forward read within dbSNP. However, integration of the cosmids (from which the SNP was derived) within the minimum tile path (chapter 4) indicated the submission, and therefore the SNP alleles, was in the wrong orientation. The allele frequency of the major and minor alleles, with Q-values >30, was 0.75 and 0.25, respectively, from 24 chromosomes sampled. The SNP was also in H-W equilibrium.

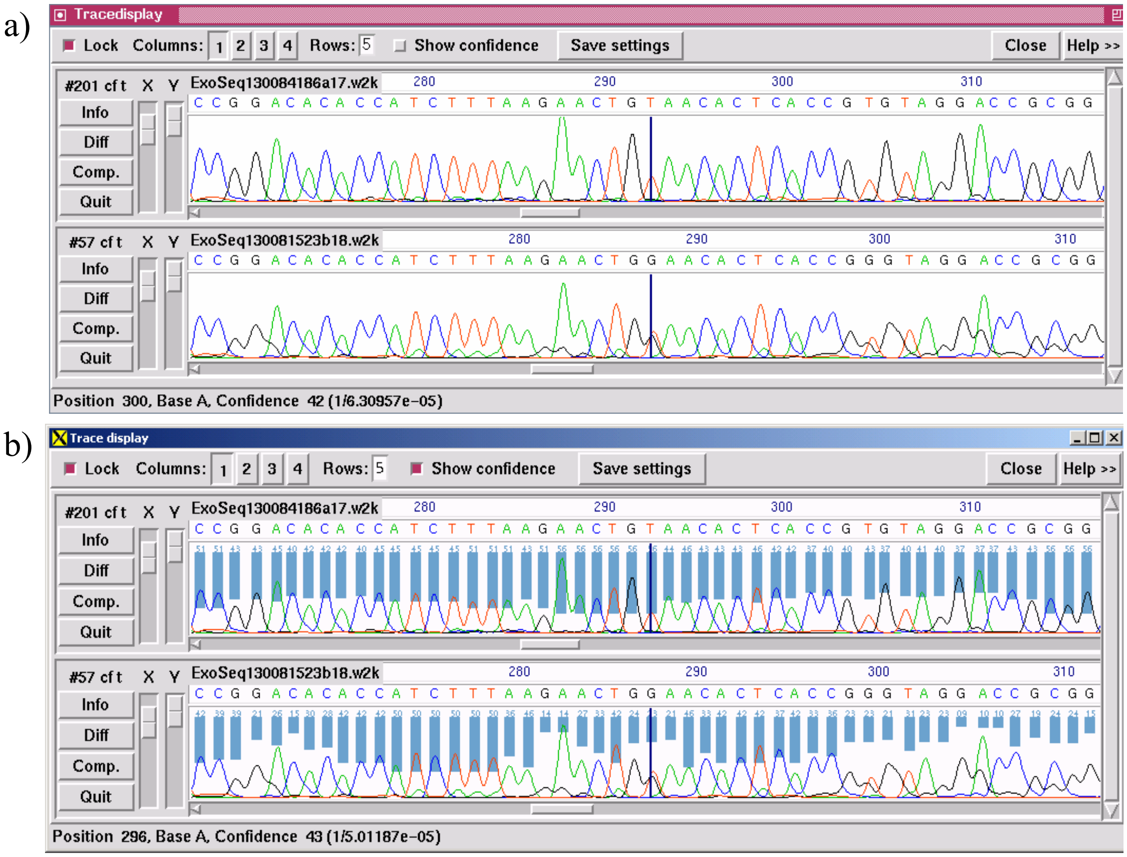

a)
Lock Columns: 1 2 3 4 Rows: 5 Show confidence Save settings Close Help >>
#-10 cf t X Y Simon1sa7.w2k 80 70 60 50 40
Info
Diff
Comp.
Quit
#-18 cf t X Y Simon1sm9.w2k 70 60 50 40
b)
Reading Simon1sb9.w2k (#-75)

Position 681, Base G, Confidence 42 (1/6.30957e-05)

Figure 6.9: The identification of a known T/C SNP, dbSNP: 737497. a) Assembled GAP4 traces used to identify this SNP are displayed in the reverse orientation. The top trace is the homozygous major allele and the bottom trace is the homozygous minor allele. b) Q-values associated with the SNP nucleotides.

6.8.2.2 Novel SNPs

Novel SNPs were validated using support from the base quality values, either of the homozygote minor allele or of the surrounding bases for a heterozygote allele. Where possible, SNPs were verified within reverse reads. All novel SNPs that were identified as part of this study were in HW equilibrium. Figure 6.10 is an example of a novel A152C SNP localised to intron 5 of GSTM2 and identified in the opposite orientation. Allele frequencies of the major (0.96) and minor (0.4) alleles were determined from a sample of 48 chromosomes.



```
#201 cf t   ExoSeq130084186a17.w2k
CCGGACACACCATCTTTAAGAACTGTAACACTCACCGTGTAGGACCGCGG
#57 cf t    ExoSeq130081523b18.w2k
CCGGACACACCATCTTTAAGAACTGGAACACTCACCGGGTAGGACCGCGG
Position 300, Base A, Confidence 42 (1/6.30957e-05)
```

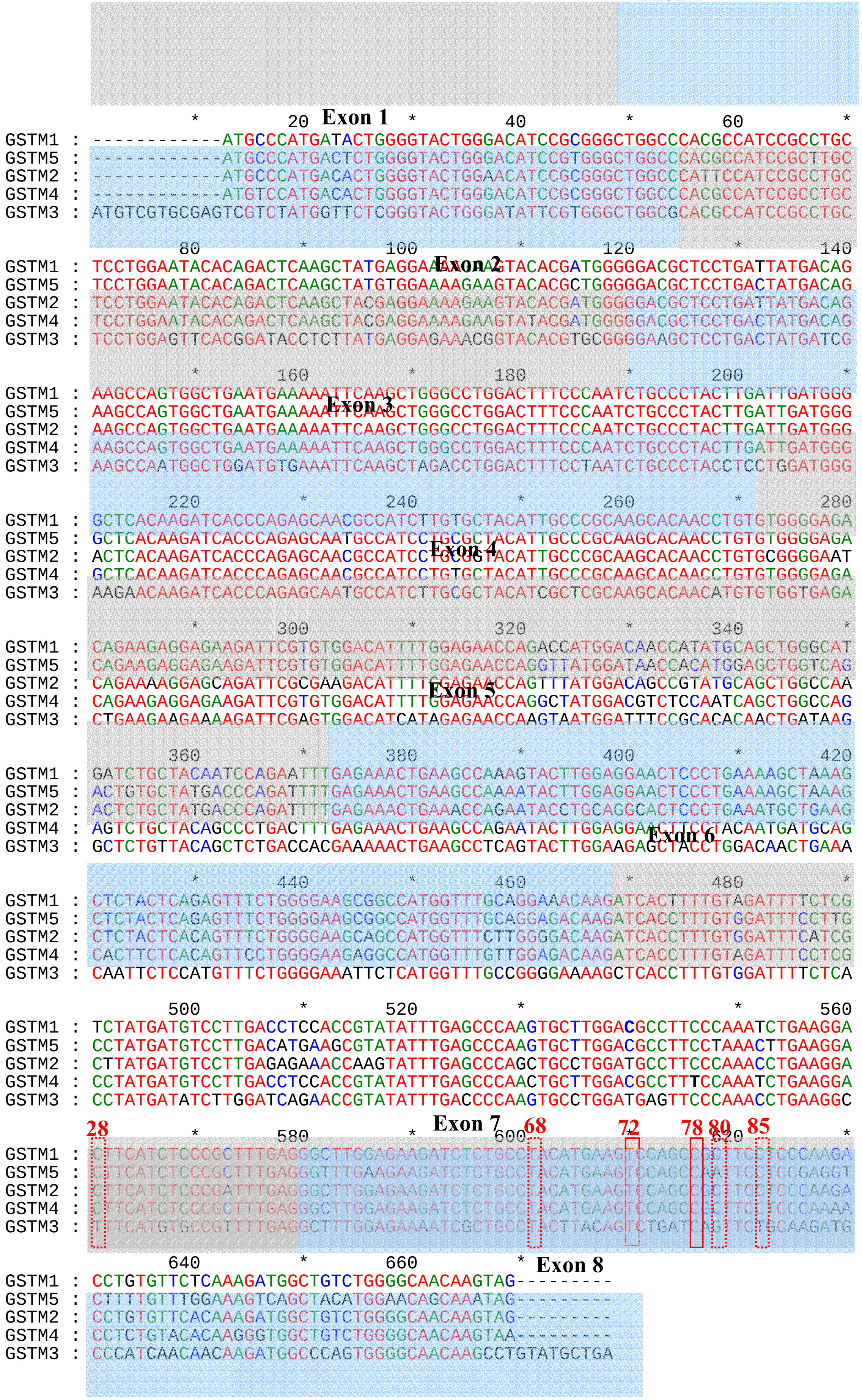
```
#201 cf t   ExoSeq130084186a17.w2k
CCGGACACACCATCTTTAAGAACTGTAACACTCACCGTGTAGGACCGCGG
#57 cf t    ExoSeq130081523b18.w2k
CCGGACACACCATCTTTAAGAACTGGAACACTCACCGGGTAGGACCGCGG
Position 296, Base A, Confidence 43 (1/5.01187e-05)
```

Figure 6.10: A novel A/C SNP identified within intron 3 of GSTM2. a) Assembled Gap4 traces for the homozygous major allele, top trace, and the heterozygous minor allele, bottom trace (in the reverse orientation). b) Assembled traces (from 6.10a) with associated Q.

6.8.2.3 Suspect candidate SNPs

Within Ensembl, dbSNP entries 1056806 and 506008 were localised to exon 7 of GSTM1 and 4 in positions 72 and 78, respectively. These two SNPs were also localised within GSTM1 and GSTM4 in dbSNP. Alignment of the SNPs within the coding sequence, figure 6.11, revealed that the minor allele, T at position 72 and C at position 78, was present within other gene family members. Therefore, it was uncertain if these suspect SNPs were valid, or were incorrectly aligned to the genomic sequence. Errors associated with SNPs that have been identified by sequence based detection may arise from; re-sequencing without careful selection of primers; lesser quality shotgun sequencing and subsequent alignment of the genomic read; the alignment of the SNP derived reference sequence to more than one genomic position. The positions of dbSNP:1056806 and dbSNP:506008 were localised within the assembled sequence contigs of exon 7 in GSTM1 and GSTM4. Analysis of the sequence traces at positions 72 and 78 revealed that major and minor alleles could be identified therefore satisfying criteria 1), however detailed analysis of the flanking sequence for each SNP uniquely assigned 1056806 to GSTM1 and 506008 to GSTM4. Having determined that, within the data set described here, the two SNPs belonged to different genes, the allele frequencies and H-W equilibrium were calculated (therefore satisfying criteria 2). The allele frequencies of dbSNP:1056806 were 0.58 and 0.42, whilst the frequencies of dbSNP:506008 were 0.85 and 0.15. Both loci were determined to be in H-W equilibrium.

Having proven, within this data set, the SNPs were locus specific, there remains the possibility the minor allele may have arisen from the production of sequence from a mispriming event of a gene family member. Whilst there is a high degree of sequence conservation within exon 7 there are single base variants between loci which should be present in the sequence traces if the primers had misprimed. The bases of the alternative alleles, located at positions 28, 68, 80 and 85 of figure 6.11, should be present as heterozygous alleles in all traces. Analysis of the sequence failed to show any bi-allelic reads at these positions therefore supporting the unique assignment of the two submitted SNPs



```
                 *        20         *        40         *        60         *
GSTM1 : ------------ATGCCCATGATACTGGGGTACTGGGACATCCGCGGGCTGGCCCACGCCATCCGCCTGC
GSTM5 : ------------ATGCCCATGACTCTGGGGTACTGGGACATCCGTGGGCTGGCCCACGCCATCCGCTTGC
GSTM2 : ------------ATGCCCATGACACTGGGGTACTGGAACATCCGCGGGCTGGCCCATTCCATCCGCCTGC
GSTM4 : ------------ATGTCCATGACACTGGGGTACTGGGACATCCGCGGGCTGGCCCACGCCATCCGCCTGC
GSTM3 : ATGTCGTGCGAGTCGTCTATGGTTCTCGGGTACTGGGATATTCGTGGGCTGGCGCACGCCATCCGCCTGC

             80         *       100         *       120         *       140
GSTM1 : TCCTGGAATACACAGACTCAAGCTATGAGGAAAAGAAGTACACGATGGGGGACGCTCCTGATTATGACAG
GSTM5 : TCCTGGAATACACAGACTCAAGCTATGTGGAAAAGAAGTACACGCTGGGGGACGCTCCTGACTATGACAG
GSTM2 : TCCTGGAATACACAGACTCAAGCTACGAGGAAAAGAAGTACACGATGGGGGACGCTCCTGATTATGACAG
GSTM4 : TCCTGGAATACACAGACTCAAGCTACGAGGAAAAGAAGTATACGATGGGGGACGCTCCTGACTATGACAG
GSTM3 : TCCTGGAGTTCACGGATACCTCTTATGAGGAGAAACGGTACACGTGCGGGGAAGCTCCTGACTATGATCG

                 *       160         *       180         *       200         *
GSTM1 : AAGCCAGTGGCTGAATGAAAAATTCAAGCTGGGCCTGGACTTTCCCAATCTGCCCTACTTGATTGATGGG
GSTM5 : AAGCCAGTGGCTGAATGAAAAATTCAAGCTGGGCCTGGACTTTCCCAATCTGCCCTACTTGATTGATGGG
GSTM2 : AAGCCAGTGGCTGAATGAAAAATTCAAGCTGGGCCTGGACTTTCCCAATCTGCCCTACTTGATTGATGGG
GSTM4 : AAGCCAGTGGCTGAATGAAAAATTCAAGCTGGGCCTGGACTTTCCCAATCTGCCCTACTTGATTGATGGG
GSTM3 : AAGCCAATGGCTGGATGTGAAATTCAAGCTAGACCTGGACTTTCCTAATCTGCCCTACCTCCTGGATGGG

            220         *       240         *       260         *       280
GSTM1 : GCTCACAAGATCACCCAGAGCAACGCCATCTTGTGCTACATTGCCCGCAAGCACAACCTGTGTGGGGAGA
GSTM5 : GCTCACAAGATCACCCAGAGCAATGCCATCCTGCGCTACATTGCCCGCAAGCACAACCTGTGTGGGGAGA
GSTM2 : ACTCACAAGATCACCCAGAGCAACGCCATCCTGCGGTACATTGCCCGCAAGCACAACCTGTGCGGGGAAT
GSTM4 : GCTCACAAGATCACCCAGAGCAACGCCATCCTGTGCTACATTGCCCGCAAGCACAACCTGTGTGGGGAGA
GSTM3 : AAGAACAAGATCACCCAGAGCAATGCCATCTTGCGCTACATCGCTCGCAAGCACAACATGTGTGGTGAGA

                 *       300         *       320         *       340         *
GSTM1 : CAGAAGAGGAGAAGATTCGTGTGGACATTTTGGAGAACCAGACCATGGACAACCATATGCAGCTGGGCAT
GSTM5 : CAGAAGAGGAGAAGATTCGTGTGGACATTTTGGAGAACCAGGTTATGGATAACCACATGGAGCTGGTCAG
GSTM2 : CAGAAAAGGAGCAGATTCGCGAAGACATTTTGGAGAACCAGTTTATGGACAGCCGTATGCAGCTGGCCAA
GSTM4 : CAGAAGAGGAGAAGATTCGTGTGGACATTTTGGAGAACCAGGCTATGGACGTCTCCAATCAGCTGGCCAG
GSTM3 : CTGAAGAAGAAAAGATTCGAGTGGACATCATAGAGAACCAAGTAATGGATTTCCGCACACAACTGATAAG

            360         *       380         *       400         *       420
GSTM1 : GATCTGCTACAATCCAGAATTTGAGAAACTGAAGCCAAAGTACTTGGAGGAACTCCCTGAAAAGCTAAAG
GSTM5 : ACTGTGCTATGACCCAGATTTTGAGAAACTGAAGCCAAAATACTTGGAGGAACTCCCTGAAAAGCTAAAG
GSTM2 : ACTCTGCTATGACCCAGATTTTGAGAAACTGAAACCAGAATACCTGCAGGCACTCCCTGAAATGCTGAAG
GSTM4 : AGTCTGCTACAGCCCTGACTTTGAGAAACTGAAGCCAGAATACTTGGAGGAACTTCCTACAATGATGCAG
GSTM3 : GCTCTGTTACAGCTCTGACCACGAAAAACTGAAGCCTCAGTACTTGGAAGAGCTACCTGGACAACTGAAA

                 *       440         *       460         *       480         *
GSTM1 : CTCTACTCAGAGTTTCTGGGGAAGCGGCCATGGTTTGCAGGAAACAAGATCACTTTTGTAGATTTTCTCG
GSTM5 : CTCTACTCAGAGTTTCTGGGGAAGCGGCCATGGTTTGCAGGAGACAAGATCACCTTTGTGGATTTCCTTG
GSTM2 : CTCTACTCACAGTTTCTGGGGAAGCAGCCATGGTTTCTTGGGGACAAGATCACCTTTGTGGATTTCATCG
GSTM4 : CACTTCTCACAGTTCCTGGGGAAGAGGCCATGGTTTGTTGGAGACAAGATCACCTTTGTAGATTTCCTCG
GSTM3 : CAATTCTCCATGTTTCTGGGGAAATTCTCATGGTTTGCCGGGGAAAAGCTCACCTTTGTGGATTTTCTCA

            500         *       520         *                   *       560
GSTM1 : TCTATGATGTCCTTGACCTCCACCGTATATTTGAGCCCAAGTGCTTGGACGCCTTCCCAAATCTGAAGGA
GSTM5 : CCTATGATGTCCTTGACATGAAGCGTATATTTGAGCCCAAGTGCTTGGACGCCTTCCTAAACTTGAAGGA
GSTM2 : CTTATGATGTCCTTGAGAGAAACCAAGTATTTGAGCCCAGCTGCCTGGATGCCTTCCCAAACCTGAAGGA
GSTM4 : CCTATGATGTCCTTGACCTCCACCGTATATTTGAGCCCAACTGCTTGGACGCCTTTCCAAATCTGAAGGA
GSTM3 : CCTATGATATCTTGGATCAGAACCGTATATTTGACCCCAAGTGCCTGGATGAGTTCCCAAACCTGAAGGC

                 *       580         *       600         *       620         *
GSTM1 : CTTCATCTCCCGCTTTGAGGGCTTGGAGAAGATCTCTGCCTACATGAAGTCCAGCCGCTTCCTCCCAAGA
GSTM5 : CTTCATCTCCCGCTTTGAGGGTTTGAAGAAGATCTCTGCCTACATGAAGTCCAGCCAATTCCTCCGAGGT
GSTM2 : CTTCATCTCCCGATTTGAGGGCTTGGAGAAGATCTCTGCCTACATGAAGTCCAGCCGCTTCCTCCCAAGA
GSTM4 : CTTCATCTCCCGCTTTGAGGGCTTGGAGAAGATCTCTGCCTACATGAAGTCCAGCCGCTTCCTCCCAAAA
GSTM3 : TTTCATGTGCCGTTTTGAGGCTTTGGAGAAAATCGCTGCCTACTTACAGTCTGATCAGTTCTGCAAGATG

            640         *       660         *
GSTM1 : CCTGTGTTCTCAAAGATGGCTGTCTGGGGCAACAAGTAG---------
GSTM5 : CTTTTGTTTGGAAAGTCAGCTACATGGAACAGCAAATAG---------
GSTM2 : CCTGTGTTCACAAAGATGGCTGTCTGGGGCAACAAGTAG---------
GSTM4 : CCTCTGTACACAAGGGTGGCTGTCTGGGGCAACAAGTAA---------
GSTM3 : CCCATCAACAACAAGATGGCCCAGTGGGGCAACAAGCCTGTATGCTGA
```

Figure 6.11: The alignment of coding sequence for the GSTM1 – 5 genes using ClustalW and representation within GeneDoc. Exons are alternatively coloured light blue and grey. Red boxes denote the position of SNPs dbSNP:1056806 and dbSNP:506008 at positions 72 GSTM1 and position 78 GSTM4, respectively. Dotted red boxes indicate nucleotide variants between genes which permitted that unique assignment of the two SNPs.

6.8.2.4 Rejected candidate SNPs

The identification of SNPs by mRNA BLAST alignment to genomic sequence can erroneously localise SNPs within highly homologous genes. Seventeen GSTM SNPs with *de novo* sequence coverage were rejected as candidate SNPs within this study because they could not be uniquely placed by genomic alignment within a chromosome specific ACeDB. Thirteen of these SNPs could not be uniquely assigned due to the failure of two of the five paralogous genes to generate 3'UTR sequence (to which the majority of ambiguous matches aligned). Detailed analysis of one of the remaining four rejected candidates, dbSNP: 3211191, which localised to exon 4 of GSTM1 and GSTM4 (blue diamond in figure 6.7), indicated that exact sequence alignment of 20 bp either side of the SNP failed to align uniquely assign it to either individual GSTM gene. There was 100% sequence homology extending for 58 bases 5' and 41 bases 3' of the SNP into intron 3 and intron 5, respectively. The observation of a single base difference between the two genes, in the present study, allowed assignment of the read and therefore the SNP to GSTM4. A second SNP that failed to localise to a unique position by sequence matching was dbSNP: 402505. The SNP was located within intron 3 of GSTM2 and 5. It had 100% homology both 5' and 3' of the SNP but nucleotide differences at 155 bases 5’ and 46 bases 3’ of the SNP, from the present study, permitted assignment of the SNP to GSTM5. The remaining two of the seventeen unlocalised SNPs were rejected on based upon similar arguments to those examples described above.

6.8.3 Effect of sequence variation upon gene structure

A novel SNP, identified within the *de novo* sequence of GSTT1, provided the opportunity to examine the effects allelic variation may have upon the translated amino acid sequence and protein structure. The protein encoded by GSTT1, like the GSTM family members, is a phase 2 enzyme that detoxifies carcinogenic metabolites, for example halomethanes, by conjugation of glutathione which changes the polarity of the metabolite making them more readily excreted (Mannervik *et al*., 1988).

It has previously been shown that the GSTT1 gene is absent from 38% of the population (Pemble *et al*., 1994). Analysis of the set of DNAs examined in this study indicated that only 1 of 8 CEPH DNAs tested, NA07017, failed to generate a PCR product from any primer pair of the 5 exons from the gene. The lower observed null percentage (13%) compared to that previously reported may relate to the small sample set tested.

The novel A110C SNP, within exon 3 of GSTT1, results in a first base substitution of amino acid 104 causing a non-synonymous, non-conservative change of a threonine to a proline (Thr104-Pro). Threonine is an aliphatic amino acid that has a neutral side chain whereas proline is an amino acid with a secondary group. Alignment of the protein domain containing the single nucleotide polymorphism was performed within PFAM (http://www.sanger.ac.uk/Software/Pfam/). Of the 751 proteins that have conserved homology to the domain encoding the Thr104 – Pro nucleotide variant none contain a proline at position 104. This suggests that the protein conformation resulting from the non-synonymous, non-conservative SNP is not usually assumed by GSTT1 or any closely related protein domain. A homologous protein with an elucidated 3-D structure was identified within PDB by sequence alignment to investigate what effect the SNP may have upon the conformation of GSTT1. 3LJR, whose 3-D structure as a dimer conjugated with glutathione substrate was elucidated by X-ray diffraction, is a glutathione S-transferase family member, GSTT2, that shows a 55% sequence homology with GSTT1 (figure 6.12). The substitution Thr104 to Pro in GSTT1 corresponds to Asp104 to Pro in 3LJR. Viewing the elucidated 3-D structure within ICMLite (http://www.molsoft.com/products/icmlite.htm, Abagyan *et al*., 1994) (with assistance from Robert Steward) shows Asp104 is at the C-terminal of an α-helix and shares a hydrogen bond with Trp101. This hydrogen bond cannot exist with a Pro substitution, since the Pro residue is cyclic with no free NH group. The loss of this hydrogen bond by an introduction of a Pro may destabilise the 3D conformation of the protein (Karvonen *et al*., 1998).

The substitution of a proline at amino acid position 104 may have a significant affect upon the function of the protein as Asp 104 (in 3LJR) conjugates with the glutathione substrate within the other subunit of the homodimer. The structural affect the proline substitution may also be important as an adjacent residue, Arg 107, interacts with the glutathione substrate on the same subunit (figure 6.13). Therefore a change in geometry caused by Thr104 – Pro may effect the conformation of active site binding residues (thereby effecting protein function). Further experimental studies would be required to determine whether the introduction of a Pro at amino acid position 104 would affect GSTT1 enzyme function when conjugating glutathione as a homodimer.

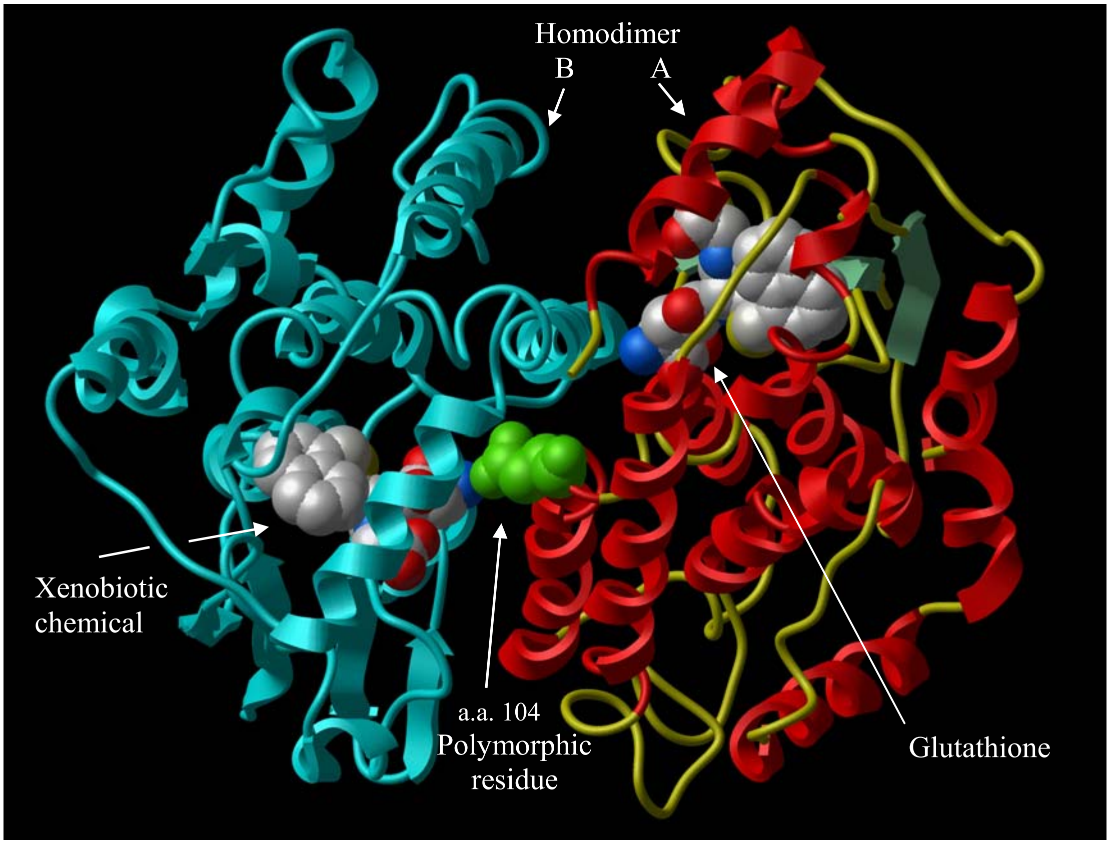


Figure 6.12: A 3-D representation within ICMLite of the homodimeric GST model, 3LJR. The model was used to interpret the effect that a novel cSNP within GSTT1 may have upon the structure of the protein. Shown are the two GST units that form the dimer (yellow/red and green) conjugated with glutathione and xenobiotic biochemicals. The polymorphic Asp104Pro residue is drawn in green at the C-terminal of an α-helix.

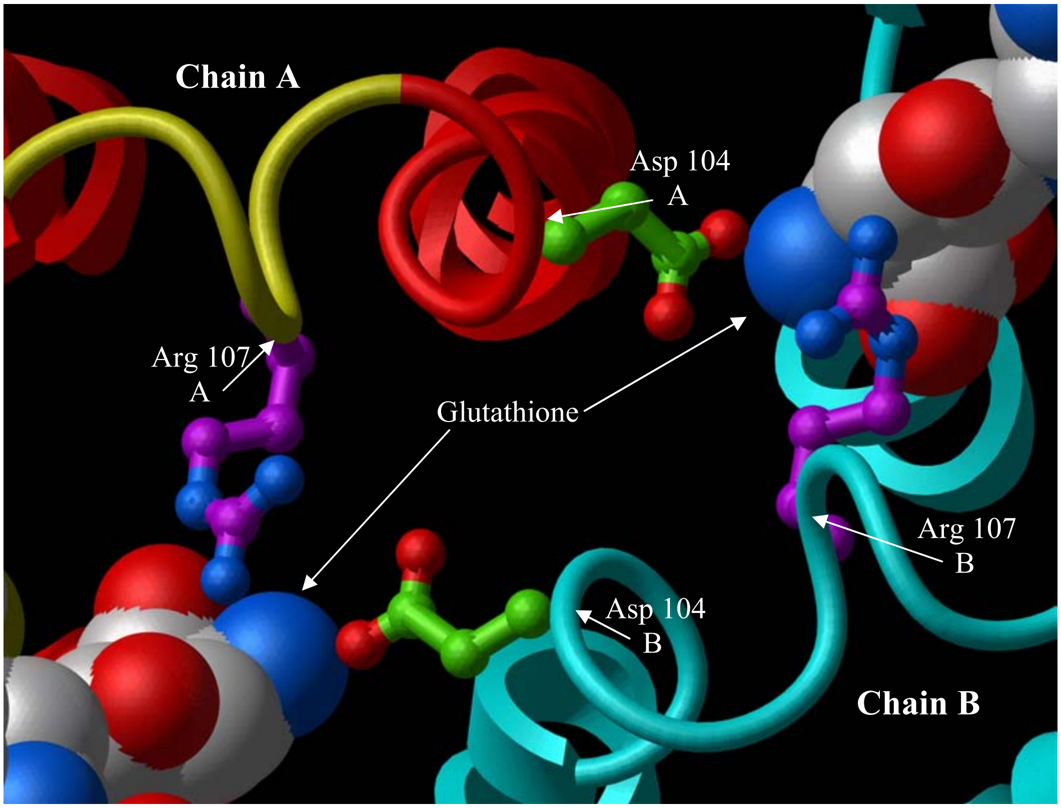


Figure 6.13: A vertical cross-sectional view of glutathione conjugating amino acid residues of 3LJR. Polymorphic amino acid residue, Asp 104, conjugates with the glutathione molecule on the opposite chain in the homodimer, chain B. Arg107 on chain A conjugates with glutathione residue within the same chain. The conformation of the dimmer means that hydrogen bonding conjugation is reciprocated between dimers.

6.9 Discussion

Novel sequence coverage has been generated for 77 (85%) exons from 12 genes that has identified 17 known and 9 novel SNPs. Exon specific primers were designed within highly homologous glutathione-S-transferase Mu 1 – 5 paralogous genes and seven other genes of medical interest to facilitate the unique assignment of known and identify novel SNPs.

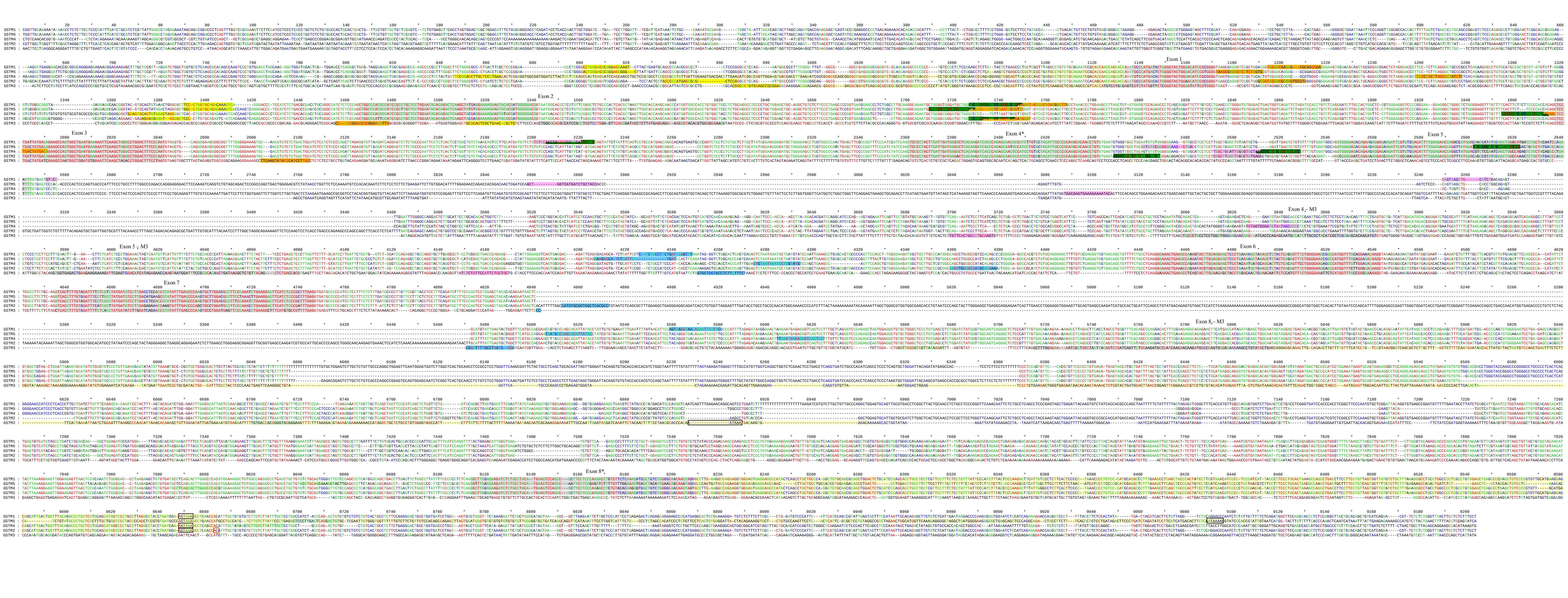
The unambiguous localisation of SNPs within genomic sequence is central to determining the biological effects of coding SNPs or to their use in the construction of haplotype maps. Two known SNPs, with multiple loci according to the Ensembl genome database, were uniquely localised with sequence contigs assembled from the GSTM exon specific PCR products. A further 2 known SNPs with multiple loci, whose placement was not resolved within *de novo* sequence contigs, were uniquely assigned by sequence alignment of extended flanking sequence within an ACeDB database.

Fifty five of the 291 known SNPs localising to the genes in this study were not detected within exon, and partial flanking intron, specific sequence contigs. The absence of these SNPs does not necessarily mean that they are false but may instead be attributed to the ethnicity of the population from which the SNP was derived being different to the population tested here. Alternatively, the number of chromosomes contained within the reads of the sequence assembly may not be sufficient to detect the SNP if it has a low minor allele frequency (in this study, it would be unlikely for a minor allele of around 1% or less to be detected).

Eight known SNPs covered by *de novo* sequence failed to align to genomic sequence within genes structures in respective ACeDB databases. This may be caused by nucleotide differences within draft sequence from which the SNPs were identified (for example, dbSNP: 2545753 within exon 6 of CYP2A6) which were not subsequently present in the finished sequence. The alignment of mRNA sequence to genomic sequence to aid SNP identification may also cause inaccuracies in SNP assignment if there are errors in original mRNA sequence (for example, dbSNP: 1061604 within exon 8 of CYP2A6). In addition, if the mRNA was derived from a closely related paralogous gene, it is possible the mRNA sequence may be misaligned to the genomic sequence of a highly homologous locus.

There are approximately 60,000 coding SNPs present within the human genome (ISMWG, 2001), corresponding to 1-2 per gene per individual. However, the effect of these SNPs upon the function of genes is largely unknown. Many of these cSNPs will result in synonymous amino acid changes due to codon redundancy or conservative non-synonymous changes by the substitution of an amino acid of similar properties. However, a proportion of cSNPs will result in non-synonymous non-conservative changes that, depending on the amino acid substitution, may cause changes in the structural integrity and/or biological function of the protein. As discussed in the previous chapter, the structure, function and interaction of the majority of human gene products is largely unknown. The determination and characterisation of protein structures and the networks in which they interact, together with the elucidation of genuine SNPs within coding features will contribute to our understanding of the metabolic differences between individuals.

This chapter describes the possible effect that a non-synonymous, non-conservative coding SNP, Thr104 – Pro, may have upon the structure and function of the GSTT1 protein. Since this work, discovery of the same SNP in a Swedish population has been published (Alexandrie *et al*., 2002). Experimental investigation by Alexandrie *et al.,* (2002) has confirmed that the variant does indeed have a functional consequence. The so-called GSTT1*B allele was reported as having a frequency of 0.05 in Swedish Saamis, the same as observed within the CEPH DNAs used here. An ELISA assay showed that a stable protein product was not produced by individuals who had previously been genotyped as non-conjugators, lacking a GSTT1 protein that could functionally conjugate methyl chloride. Western blot analysis was also used to determine that the functional protein was absent within erythrocyte lysates from non-conjugating individuals. The association of the GSTT1*B, allele in conjunction with the functional GSTT1*A or the null GSTT1*0 allele, may explain why individuals that were previously thought to produce functional copies of the GSTT1 protein are associated with the non-conjugating phenotype.


GSTM1
GSTM5
GSTM4
GSTM2
GSTM3
Exon 1
Exon 2
Exon 3
Exon 4*
Exon 5
Exon 4 – M3
Exon 5 – M3
Exon 6
Exon 7
Exon 8 – M3
Exon 8*