# Structure and organization of the human Theta-class glutathione S-transferase and D-dopachrome tautomerase gene complex

Marjorie COGGAN, Lel WHITBREAD, Angela WHITTINGTON and Philip BOARD[1]

Molecular Genetics Group, John Curtin School of Medical Research, Australian National University, P.O. Box 334, Canberra, ACT 2601, Australia

The structure and organization of the human Theta-class glutathione S-transferase (GST) genes have been determined. *GSTT1* and *GSTT2* are separated by approx. 50 kb. They have a similar structure, being composed of five exons with identical exon/intron boundaries. *GSTT1* is 8.1 kb in length, while *GSTT2* is only 3.7 kb. The *GSTT2* gene lies head-to-head with a gene encoding D-dopachrome tautomerase (DDCT), which extends over 8.5 kb and contains four exons. The sequence between *GSTT2* and *DDCT* may contain a bidirectional promoter. The *GSTT2* and *DDCT* genes have been duplicated in an inverted repeat. Sequence analysis of the duplicated *GSTT2* gene has identified an exon 2/intron 2 splice site abnormality and a premature translation stop signal at codon 196. These changes suggest that the duplicate gene is a pseudogene, and it has been named *GSTT2P*.

## INTRODUCTION

The glutathione S-transferases (GSTs) are a large family of proteins that catalyse the conjugation of reduced glutathione to a variety of electrophilic and hydrophobic compounds. Although many substrates are xenobiotics, some endogenously derived products of oxygen metabolism, such as the lipid hydroperoxides and the alkenals, have been shown to be substrates for particular GST isoenzymes [1,2]. In addition to their catalytic activities, some GSTs have been shown to bind a number of hydrophobic compounds, including bilirubin, haem and steroid hormones [3,4]. The role of GSTs in the metabolism of toxic substances has led to their implication in both susceptibility to carcinogens [5,6] and the development of resistance to drugs, pesticides, herbicides and antibiotics [1,7].

The GSTs are a diverse superfamily, and in mammals the cytosolic GSTs have been grouped into the Alpha, Mu, Pi, Sigma, Theta and Zeta classes [8–11]. There are multiple enzymes within each of the Alpha, Mu and Theta classes that are the products of distinct gene loci. In contrast, the variant isoenzymes in the human Pi and Zeta classes appear to be allelic.

To date, two human Theta-class isoenzymes, GSTT1-1 [12] and GSTT2-2 [13,14], have been identified. Although these two enzymes have a number of features in common, they share only 55% amino acid sequence identity. These isoenzymes represent the two Theta subclasses that have been identified in mammals [15]. The human enzyme GSTT1-1 is orthologous to the rat 5-5 [9] and mouse mGST T1-1 [16] enzymes, while the human enzyme GSTT2-2 is orthologous to rat yrs-yrs [17], rat 12-12 [9] and mouse mGST T2-2 [15].

Compared with the other classes, the Theta-class isoenzymes have a number of distinct characteristics, including their inability to bind to glutathione affinity matrices and their distinct substrate specificities. Dichloromethane and related compounds appear to be significant substrates for GSTT1-1, while GSTT2-2 is most active towards cumene hydroperoxide, ethacrynic acid and menaphthyl sulphate.

Although structures have been reported for the rat *rGSTT2* [17] and the mouse *mGSTT2* genes [15], almost nothing is known about the genomic organization of the human Theta-class GSTs or the factors that regulate their expression. The human *GSTT1* gene is frequently deleted [12], and around 16% of Europeans are homozygous for this deletion [18].

The present study describes the cloning, sequencing and structure of the human *GSTT2* gene and a previously unknown *GSTT2* pseudogene. During the course of this investigation, a gene encoding D-dopachrome tautomerase (DDCT) was identified in close proximity to the *GSTT2* gene. The *DDCT* gene is also duplicated in tandem with the *GSTT2* pseudogene. The *GSTT1* gene has been located approx. 50 kb away from the *GSTT2/DDCT* gene cluster, and has been shown to have an intron/exon structure similar to that of the *GSTT2* gene.

## EXPERIMENTAL

### Library screening

A human cosmid library [19] was screened by hybridization with a 767 bp cDNA fragment encoding human GSTT2 [14,20]. A number of positive clones were isolated and analysed by restriction endonuclease digestion to identify appropriate fragments for subcloning. Two *Bam*HI fragments (3.5 and 2.7 kb) and two *Sac*I fragments (0.8 and 3.3 kb) that hybridized with the 767 bp cDNA fragment used for the library screening were isolated from the cosmid pT2cos2.

### Cloning and sequencing

All fragments were subcloned into M13mp18 and M13mp19 for sequencing. Initial sequences were determined with vector-specific primers, and these were extended using GSTT2-sequence-specific oligonucleotide primers. Direct sequencing of cosmid

Abbreviations used: GST, glutathione S-transferase; DDCT, D-dopachrome tautomerase; MIF, macrophage migration inhibitory factor; EST, expressed sequence tag; RFLP, restriction fragment length polymorphism.

[1] To whom correspondence should be addressed (e-mail Philip.Board@anu.edu.au).

The nucleotide sequence data obtained in this study have been submitted to the GenBank, EMBL and DDBJ Nucleotide Sequence Databases with the accession numbers AN057172, AN057173, AN057174, AN057175 and AN057176.

**Table 1 Primer sequences used for PCR amplification of *GSTT2* and *GSTT2P* gene fragments**

| Primer | Sequence | Position identified |
|---|---|---|
| T2SeqEx2 | Forward 5′ GCAGATCAACAGCCTGG 3′ | GSTT2P |
| T2INT2B | Reverse 5′ CCATGGGGTATGGGAGG 3′ | Exon 2 splice junction |
| HTA5 | Forward 5′ GAACTGTTTGAGGGACGGC 3′ | GSTT2P |
| HTB5 | Reverse 5′ GAAGCAGCATAGCCTGATAG 3′ | Exon 5, codon 196 |
| HT4F | Forward 5′ GCCACTCATTGGGGTCC 3′ | GSTT2 |
| HTIPR | Reverse 5′ ATCAGCTCCTCCAGGGCC 3′ | Exon 4, codon 139 |
| Cos F | 5′ CAATTTAACTGTGATAAACTACCG 3′ | 5′ and 3′ ends of the multiple cloning |
| Cos R | 5′ CACGAGGCCCTTTCGTCTTC 3′ | site of the cosmid vector pCV001(19) |

clones was also carried out using a Thermosequenase cycle sequencing kit (Amersham). All exons and exon/intron boundaries were sequenced on both strands.

### Identification of polymorphisms

Several differences between the GSTT2 cDNA and gene were identified during the sequencing process. These variations were investigated by restriction endonuclease analysis of PCR products. Oligonucleotide primers were designed to amplify the exon 2/intron 2 junction, exon 4 and exon 5. A G-to-A transition in the first base of intron 2 was identified by digestion with *Bsp1286*I. Two variations can be identified in exon 4 by digestion with *Ava*I and *Nco*I, and one variant in exon 5 was identified by digestion with *Fok*I. The digested products were separated on 12 % (w/v) acrylamide gels in TBE buffer (89 mM Tris, 89 mM boric acid, 2.5 mM $Na_2EDTA$). The primers and sequence variations detected are shown in Table 1.

### PCR amplification

DNA amplification was performed in a volume of 20 $\mu$l using Advanced Biotechnologies buffer 4 [75 mM Tris/HCl, pH 9.0, 20 mM $(NH_4)_2SO_4$, 0.1 % (w/v) Tween] plus 1.5 mM $MgCl_2$, 4–6 pmol of each primer, 0.2 $\mu$M of each dNTP, 0.5 unit of *Taq* polymerase and 25 ng of DNA template. Amplification in an FTS-1 Thermal Sequencer (Corbett Research, Sydney, Australia) was typically at 95 °C for 20 s, annealing for 20 s and extension at 72 °C for 20 s. The annealing temperature varied with specific primer pairs: T2SeqEx2/T2INT2B, 56 °C; HTA5/HTB5, 59 °C; HT4F/HTIPR, 56 °C; CosF/CosR, 64 °C. PCR amplification of long DNA fragments was performed using standard amplification techniques with longer extension times and the 'Expand' long template PCR system (Boehringer Mannheim).

## RESULTS

A number of positive clones were identified by library screening and initially one cosmid, T2cos2, was subjected to further analysis. Restriction endonuclease digestion of the clone confirmed the presence of a previously identified strongly hybridizing 3.6 kb *Bam*HI fragment and identified appropriate fragments for subcloning. Using both the subcloned DNA and direct sequencing of the cosmid DNA, the entire sequence of the encoded gene was determined. The exon/intron boundaries were identified by comparison with the cDNA sequence and the mouse *GSTT2* gene [15]. The approximate start of transcription was estimated by analysis of cDNA clones identified in the EST (expressed sequence tag) database. Using the longest cDNA (accession no. W68102) as a 5′ reference point and the polyA addition signal as the 3′ end, the gene encoded in T2cos2 appears to be approx. 3.7 kb in length and is composed of five exons (Figure 1). Because the precise start of transcription is not clear, the bases in this figure have been numbered from the ATG start of translation.

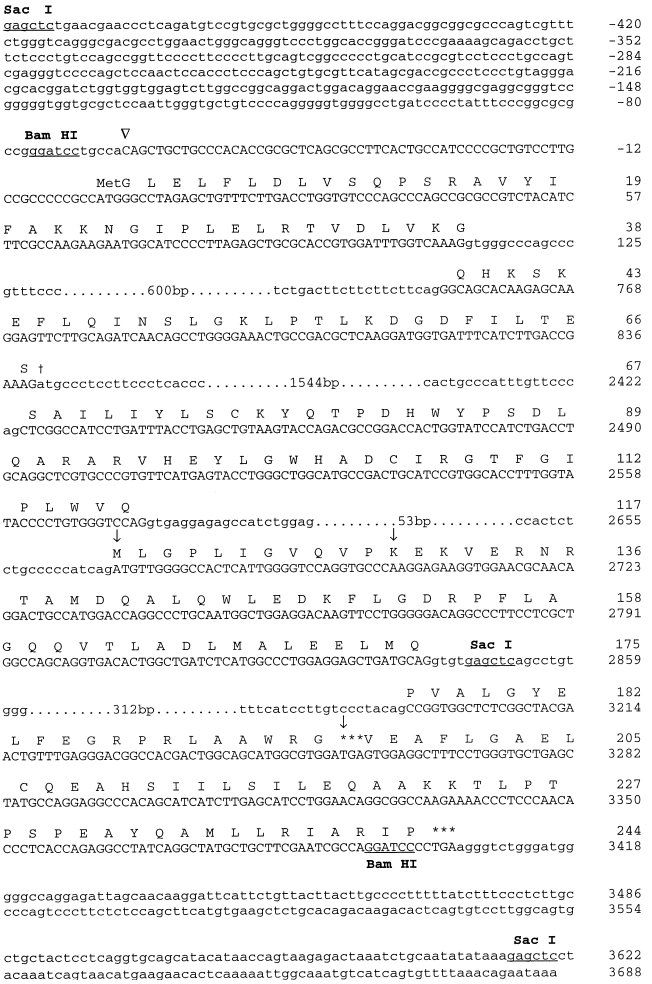


**Figure 1 Nucleotide sequence and intron/exon structure of *GSTT2P***

The nucleotides are numbered from the ATG translation start site to the poly(A) addition signal (AATAAA). The start of the longest known cDNA clone is marked (▽). A deviation from the splice site consensus is indicated (†). Deviations from the cDNA sequence [14], including Met-118, Ile-129 and Stop196, are each indicated (↓). The positions of the *Bam*HI and *Sac*I sites used for subcloning are underlined and indicated directly. Exon sequence is shown in upper case.

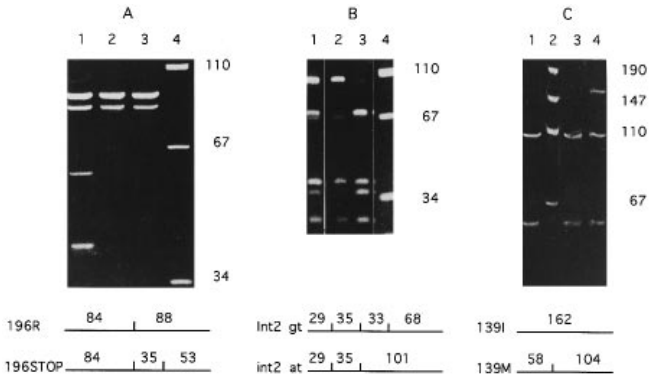


**Figure 2 PCR–RFLP analysis of polymorphic sites in the *GSTT2* and *GSTT2P* genes**

A schematic diagram of the predicted fragment sizes is shown beneath each panel. (**A**) *Fok*I digestion of a 172 bp fragment containing codon 196 from *GSTT2* and *GSTT2P*. Lane 1 contains DNA from an individual with both codon Arg-196 (196R) and codon Stop196 (196STOP). Lanes 2 and 3 contain DNA from individuals with codon 196R only. Lane 4 contains molecular length markers. (**B**) *Bsp1286*I digestion of a 165 bp fragment containing the exon 2/intron 2 junction from *GSTT2* and *GSTT2P*. Lane 1 contains DNA from an individual with both the normal (Int2 gt) and abnormal (int2 at) splice junctions. Lane 2 contains DNA from an individual with the abnormal splice junction only. Lane 3 contains DNA from an individual with the normal splice junction only. Lane 4 contains molecular length markers. (**C**) *Nco*I digestion of a 162 bp fragment containing codon 139 from both *GSTT2* and *GSTT2P*. Lanes 1 and 3 contain DNA from individuals with codon Met-139 (139M) in *GSTT2* and *GSTT2P*. Lane 2 contains molecular length markers. Lane 4 contains DNA from an individual with codon Ile-139 (139I) in *GSTT2* and codon 139M in *GSTT2P*. These results were obtained from analysis of human DNA and are similar to the results obtained from the analysis of T2cos2, T2cos8 and T2cos17. Positions of molecular size markers (bp) are indicated to the right of each gel.

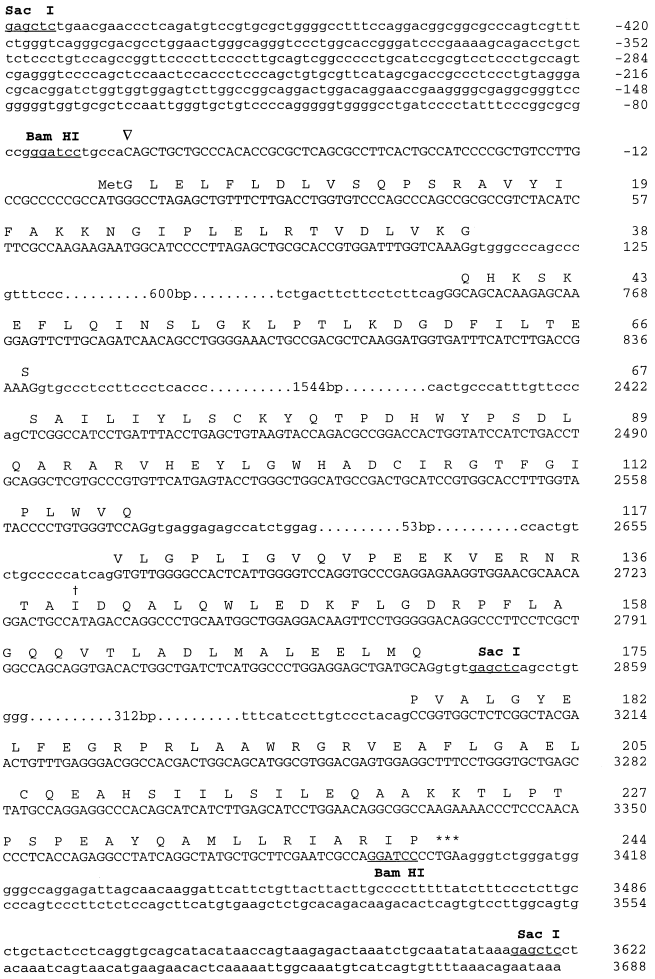


**Figure 3 Nucleotide sequence and intron/exon structure of *GSTT2***

The nucleotides are numbered from the ATG translation start site. The start of the longest known cDNA clone is marked (▽). A deviation from the cDNA sequence [14], Ile-139, is indicated (†). The positions of the *Bam*HI and *Sac*I sites also found in *GSTT2P* are underlined and indicated directly. Exon sequence is shown in upper case.

Analysis of the sequence identified a number of variations that differed from the cDNA and splice consensus sequences. These included a G toA transition at nt 2669 changing Val-118 to Met in exon 4, a G to A transition at nt 2702 changing Glu-129 to Lys in exon 4, a G to A transition at nt 2732 changing Met-139 to Ile in exon 4, a C to T transition at nt 3255 changing Arg-196 to a stop codon in exon 5, and a G to A transition at nt 841, the first base of intron 2. The identification of an in-frame stop codon and a possible splice site defect in the gene sequence suggested that the gene encoded in T2cos2 may not be active and may be a pseudogene. Since the C to T transition causing the Arg-196 to Stop substitution in exon 5 from T2cos2 also created a new *Fok*I site, it was possible to use PCR to amplify exon 5 DNA from additional cosmids and to use *Fok*I digestion to determine if they contained genes encoding either Arg-196 or Stop196 (Figure 2). Two additional cosmids (T2cos8 and T2cos17) were selected for further study. T2cos8 contains an exon 5 sequence that encodes an Arg at codon 196. In contrast, exon 5 DNA amplified from T2cos17 contained sequences for both Arg-196 and Stop196.

Sequencing of T2cos8 confirmed the exon/intron structure found in T2cos2. In this case the exon/intron boundaries all agree with the established gt–ag consensus [21], and are shown in Figure 3. The exon sequence of the gene encoded in T2cos8 agreed with the cDNA sequence [14], except for a G to A transition that results in a Met-139 → Ile substitution. It seems highly likely that the gene in T2cos8 is the normal *GSTT2* gene and that T2cos2 contains a previously undetected pseudogene, to be termed hereafter *GSTT2P*. The structures of both genes are shown schematically in Figure 4. Since T2cos17 contains exon 5 DNA encoding both Arg-196 and Stop196, it appears that the gene and pseudogene are not allelic and are in relatively close proximity. Further studies were undertaken to determine their arrangement.

Restriction enzyme analysis of the three cosmid clones showed that they all contained a hybridizing 3.6 kb *Bam*H1 fragment and hybridizing 0.77 kb and 3.3 kb *Sac*I fragments. This suggests that the pseudogene has been generated by a duplication and that some of the flanking sequence has also been duplicated.

To determine the orientation of the two copies of exon 5 in T2cos17, oligonucleotide primers were prepared from the cosmid sequence on each side of the human genomic DNA insert (Cos F and Cos R; Table 1). Long PCR amplification was performed using T2cos17 as a template with each of the cosmid primers paired with HTA5, an exon 5 primer from the sense strand of the *GSTT2* sequence (Table 1). Both cosmid primers produced an amplified product. These products (approx. 4 kb and 9 kb) were gel-purified and used as a template with primers for the specific amplification of exon 5 DNA (HTA5 and HTB5; Table 1). The amplified products were digested with *Fok*I in a manner similar to that shown in Figure 2, and the results indicated that the two

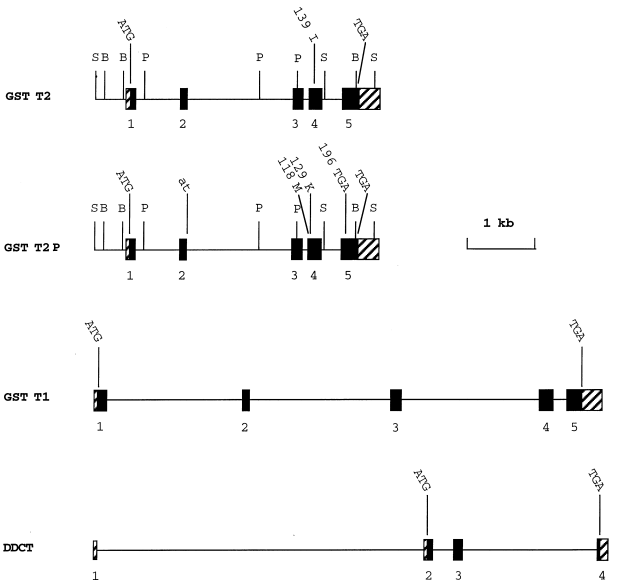


**Figure 4 Diagram of the *GSTT2*, *GSTT2P*, *GSTT1* and *DDCT* genes**

Exons are shown as boxes, non-coding regions are hatched and coding regions are filled. Translation start and stop signals and variations from the cDNA sequences and splice site consensus are indicated directly. Restriction sites are indicated as follows: B, *Bam*HI; P, *Pst*I; S, *Sac*I.

```
                                    →
...GCTGGGCTGGGACACCAGGTCAAGAAACAGCTCTAGGCCCATGGCGGGGGCGGCAAGG      58
   CGACCCGACCCTGTGGTCCAGTTCTTTGTCGAGATCCGGGTACCGCCCCCGCCGTTCC
   S  P  Q  S  V  L  D  L  F  L  E  L  G  Met      →

                                        GSTT2 Exon 1        ∇
ACAGCGGGGATGGCAGTGAAGGCGCTGAGCGCGGTGTGGGCAGCAGCTGtggcaggatccc     119
TGTCGCCCCTACCGTCACTTCCGCGACTCGCGCCACACCCGTCGTCGACaccgtcctaggg

ggcgcgccgcgggaaataggggatcaggccccacccccctggggacagcacccaattggagcgc     180
ccgcgcggccctttatcccctagtccggggtgggggaccccctgtcgtgggttaacctcgcg

accacccccggacccgcctcgccccttcggttcctgtccagtcctgccggccaagactcca     241
tggtgggggcctgggcggagcggggaagccaaggacaggtcaggacggccggttctgaggt

                                ←
ccaccagatccgtgcgtccctacagggagggcggtcgctatgaacgcacagctgggagggt     302
ggtggtctaggcacgcagggatgtccctcccgccagcgatacttgcgtgtcgaccctccca
                                                ←

ggagttggagctggggaccctcgacaggcagggaggacgcggatgcaggggccgactgca     363
cctcaacctcgaccctgggagctgtccgtccctcctgcgcctacgtccccggctgacgt

                                                * DDCT Exon 1
aggggaaggggaaccggctggacagggagaagcaggtctgcttttcggGATCCCGGTGCCA     424
tccccttcccctggccgacctgtccctgttcgtccagacgaaaagccCTAGGGCCACGGT

GGGACCCTGCCCAGTTCCAGGCGTCGCCCTGACCCAGAAACGACTGGGCGCCGCCGTCCTG     485
CCCTGGGACGGGTCAAGGTCCGCAGCGGGACTGGGTCTTTGCTGACCCGCGGCGGCAGGAC

GAAAGGCCCCAGCGCACGGACATCTGAGGgttcg......                           519
CTTTCCGGGGTCGCGTGCCTGTAGACTCCcaagc
```

**Figure 5 Nucleotide sequence of the intergenic region between the *GSTT2* and *DDCT* genes**

The start of the *GSTT2* coding sequence is indicated, and the start of the longest known *GSTT2* cDNA is marked (▽). The start of the longest known *DDCT* cDNA is marked (*). The predicted promoter regions between positions 26 and 276 on the forward strand and between positions 299 and 49 on the reverse strand are indicated (→, ←). Exon sequence is shown in upper case.

large amplified products each contained a copy of exon 5 from a different gene; the 4 kb product contained the translation stop signal at codon 196 and the 9 kb product contained an Arg codon at that position. These data confirm that the two genes are present in T2cos17 and indicate that they are in different orientations.

```
GATCCCGGTGCCAGGGACCCTGCCCAGTTCCAGGCGTCGCCCTGACCCAGAAACGACTGGGCGCCGCC     -5351

GTCCTGGAAAGGCCCCAGCGCACGGACATCTGAGGgttcgttcagagctctgttt.........5250      -45

                                                          MetP  F  L          4
...........cccagcttttcttccgccagAGCTGTTTCCGTTCCTCTGCCCGCCATGCCGTTCCTG        12

E  L  D  T  N  L  P  A  N  R  V  P  A  G  L  E  K  R  L  C  A  A  A        27
GAGCTGGACACGAATTTGCCCGCCAACCGAGTGCCCGCGGGGCTGGAGAAACGACTCTGCGCCGCCGC       80

 A  S  I  L  G  K  P  A  D                                                36
TGCCTCCATCCTGGGCAAACCTGCGGACgtaagcgtgggccgggcagc.........323.......       451

                         R  V  N  V  T  V  R  P  G  L  A  M  A  L  S       51
...acttttcatgtccccctcagCGCGTGAACGTGACGGTACGGCCGGGCCTGGCCATGGCGCTGAGC      516

G  S  T  E  P  C  A  Q  L  S  I  S  S  I  G  V  V  G  T  A  E  D  N       74
GGGTCCACCGAGCCCTGCGCGCAGCTGTCCATCTCCTCCATCGGCGTAGTGGGCACCGCCGAGGACAA      584

 R  S  H  S  A  H  F  F  E  F  L  T  K  E  L  A  L  G  Q  D  R           95
CCGCAGCCACAGCGCCCACTTCTTTGAGTTTCTCACCAAGGAGCTAGCCCTGGGCCAGGACCGgtgcg      652

                                                          I  L  I          98
taggggtagtagggg.........2061.........tttttctctgtctcctgaagGATACTTTAT      2757

 R  F  F  P  L  E  S  W  Q  I  G  K  I  G  T  V  M  T  F  L  ***         118
CCGCTTTTTCCCCTTGGAGTCCTGGCAGATTGGCAAGATAGGGACGGTCATGACTTTTTTATGAttgg      2825

gcacggagggatccagggcatctgtgaactggctgcttcttccagagagatctcttggcagagtgagg      2893

gcctggagataaccagctttggattatcccgcatgcaacattcctgtgatcacataatcctcttcttc      2961

atcctcatatgaaataaa                                                        2979
```

**Figure 6 Nucleotide sequence and intron/exon structure of *DDCT***

The nucleotides are numbered from the ATG translation start site. The sequence is listed from the start of the longest known cDNA sequence (accession no. U49785). Exon sequence is shown in upper case.

## DDCT

DDCT converts 2-carboxy-2,3-dihydroindole-5,6-quinone (D-dopachrome) into 5,6-dihydroxyindole, and is expressed widely in human tissues [22,23]. Additional sequence from the 5′ ends of the *GSTT2* and *GSTT2P* genes was obtained from T2cos8 and T2cos2, and it was noted that both 5′ flanking sequences contained the start of a gene encoding *DDCT*. In each case the *DDCT* gene is in the opposite orientation to its associated *GSTT2* gene. The intervening sequence between the *GSTT2* gene and its corresponding *DDCT* gene is shown in Figure 5. It is evident from this sequence that there are no TATA or CAAT boxes in the 5′ flanking regions of either gene. A search of this area for possible promoter sequences by application of the program Proscan: version 1.7 [24] (http://bimas.dcrt.nih.gov/molbio/proscan) predicted promoter regions between positions 26 and 276 on the forward strand and between positions 299 and 49 on the reverse strand. These overlapping regions contain a number of Sp1 sites that are commonly associated with promoters without TATA boxes.

Searches of the GenBank database also revealed that the cosmid T2cos17 overlaps with an unpublished human sequence from the clone BAC 322B1 (accession no. Z84718) obtained from chromosome 22 and an unpublished partial gene sequence for DDCT (accession no. Y11151). Comparison of these database sequences and sequence derived from T2cos17 with the *DDCT* cDNA sequence (accession no. U49785) allowed us to determine the structure and organization of the *DDCT* gene that is adjacent to *GSTT2*. The *DDCT* gene, illustrated in Figures 4 and 6, extends over 8.5 kb and contains four exons. Since this gene sequence matches the cDNA and the intron/exon junctions are in agreement with the gt–ag rule [21], it is probable that this is a functional gene. Because another *DDCT* gene can be found in association with the *GSTT2P* gene, it is clear that the duplication event that generated *GSTT2P* has included at least part of the associated *DDCT* gene. At this stage it is unclear whether the *DDCT* gene associated with *GSTT2P* is also a pseudogene.

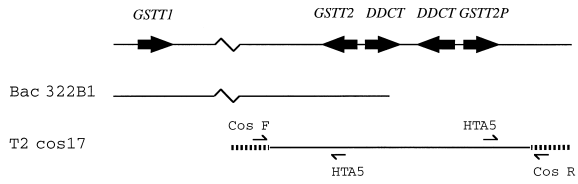


**Figure 7 Organization of the Theta-class *GST* and *DDCT* gene cluster**

The diagram shows the relative positions and orientation of the genes and is not drawn to scale. The distance between *GSTT1* and *GSTT2* is approx. 50 kb. The information used to determine this gene arrangement has been derived from the overlapping clones T2cos17 and BAC 322B1. The positions of flanking cosmid primers used to determine the orientation of the *GSTT2* and *GSTT2P* genes in T2cos17 are indicated.

### Organization of the *GSTT1*, *GSTT2* and *DDCT* genes and pseudogenes

We have previously shown that the genes encoding the two human Theta-class GST enzymes, *GSTT1* and *GSTT2*, are both located on chromosome 22q11.2 [14,25]. Examination of the BAC 322B1 sequence shows that *GSTT1* and *GSTT2* are separated by approx. 50 kb. Based on the longest available EST clone (accession no. AA300354), the *GSTT1* gene is composed of five exons and spans 8.1 kb between nt 7816 and nt 15916 of the BAC sequence. A schematic diagram of the *GSTT1* gene is shown in Figure 4. Information derived from the overlapping BAC 322B1 and T2cos17 clones has allowed the organization of the *GSTT1*, *GSTT2* and *DDCT* genes and pseudogene(s) to be determined. The relative position and orientation of each gene is summarized in Figure 7. Because of the presence of repetitive sequence in the intervening region, the complete sequence between the DDCT genes has proved to be difficult to determine. However, long PCR using a single-sense primer from exon 4 of *DDCT* generated a 2.5 kb fragment, confirming the proximity of the two genes.

### Polymorphism of *GSTT2* and *GSTT2P*

Because we found a difference between the sequence of the cDNA and the *GSTT2* gene that encoded an M139I substitution, we tested a sample of normal Australian European blood donors to determine if this substitution was the result of allelic variation. A 162 bp fragment from exon 4 was amplified with the primers HT4F and HTIPR (Table 1) and digested with *Nco*I. The G to A transition in codon 139 eliminates an *Nco*I site that can be detected by acrylamide electrophoresis of the digested PCR product (Figure 2). The results of the survey are shown in Table 2, and indicate that, with only four heterozygotes identified, the presence of the allele encoding isoleucine at codon 139 is relatively rare in the population sampled here.

We also tested the sample of blood donors for the presence of the *GSTT2P* gene by amplification of a 172 bp exon 5 fragment and digestion with *Fok*I (Figure 2). If both *GSTT2* and *GSTT2P* are present, the pattern of *Fok*I restriction fragments should contain equal amounts of exon 5 with and without the *Fok*I site that marks the presence of the premature stop codon. The results of this survey are shown in Table 2, and indicate that 28% of individuals had both the *GSTT2* gene and the *GSTT2P* gene, as indicated by the presence of the stop codon at position 196. The remaining 72% had only the *GSTT2* gene and showed no evidence for a gene with a stop codon at position 196. There are two possible explanations for these data. Either the gene duplication is polymorphic and is only present on some chromosomes; alternatively, the substitution creating the stop codon in *GSTT2P* may itself be polymorphic. To differentiate between these two possibilities, we determined the frequency of the exon 2/intron 2 splice junction substitution in the same blood donors by PCR/restriction fragment length polymorphism (RFLP) analysis (Figure 2). The results of this study are shown in Table 2, and indicate that, while 25% of the group had the *GSTT2* gene and at least one copy of a gene with the splice junction mutation, 74% had the *GSTT2* gene and presented no evidence for a gene with the splice mutation. In this case one individual (1%) presented a restriction fragment pattern indicating the presence of the *GSTT2P* gene and the absence of the *GSTT2* gene.

In 92% of the individuals studied there was a direct correlation between the presence of the exon 2/intron 2 splice site mutation and the premature stop codon in exon 5. However, despite the obvious linkage of the splice site mutation and the premature stop codon, the presence of discordant individuals supports an argument in favour of allelic variation in *GSTT2P*. In fact, the individual that appears to only have the *GSTT2P* gene, based on the splice site genotype, is apparently heterozygous for the Arg-196 → Stop mutation. This also supports the conclusion that the codon 196 variation in *GSTT2P* is polymorphic.

## DISCUSSION

The human *GSTT2* gene (3.7 kb) is intermediate in size between the rat *Yrs* gene (4 kb) [17] and the mouse *GSTT2* gene (3.1 kb) [15]. Thus the subfamily 2 Theta-class genes appear to be about the same size as the Pi-class genes and notably smaller than the

**Table 2 Restriction site polymorphisms in the *GSTT2* and *GSTT2P* genes**

| Site of variation | Restriction enzyme | Variation | Products (bp) | No. of individuals | Presence of restriction sites | | |
|---|---|---|---|---|---|---|---|
| | | | | | + | +/− | − |
| *GSTT2P* | | | | | | | |
| Exon 2/intron 2 splice junction | *Bsp1286*I | Normal splice: | 29, 33, 36, 68 | | | | |
| | | Abnormal splice: | 29, 35, 101 | 106 | 78 | 27 | 1 |
| Exon 5, residue 196 | *Fok*I | 196 R: | 84, 88 | | | | |
| | | 196 Stop: | 35, 53, 84 | 106 | 0 | 30 | 76 |
| *GSTT2* | | | | | | | |
| Exon 4, residue 139 | *Nco*I | 139M: | 58, 104 | | | | |
| | | 139I: | 162 | 101 | 97 | 4 | 0 |

human Alpha- and Mu-class genes [26–29]. In contrast, the human *GSTT1* gene identified in BAC 322B1 extends over 8.1 kb. The structural organization of the rat and mouse subfamily 2 genes appears to be conserved in the human *GSTT2* gene. All have five exons, and the positions of the exon/intron boundaries are conserved. The human *GSTT1* gene also consists of five exons, and the positions of the introns are conserved between the GSTT1 and GSTT2 subfamilies.

Previous *in situ* hybridization studies have mapped both the *GSTT1* and *GSTT2* genes to chromosome 22q11.2, suggesting that the two genes are in close proximity [14,25]. The results obtained from analysis of the BAC 322B1 and T2cos17 sequences have shown that *GSTT1* and the *GSTT2* genes are separated by approx. 50 kb. Given the differences in amino acid sequence between GSTT1 and GSTT2 (55 % identity), it appears that the 50 kb separation of the two genes may have restricted gene conversion events that tend to homogenize gene families.

The present study clearly indicates that the human *GSTT2* gene has been duplicated, giving rise to a pseudogene *GSTT2P* that is in the opposite orientation and in close proximity. The conservation of the restriction enzyme sites in and around the *GSTT2* and *GSTT2P* genes has obscured the presence of the gene duplication in the previous studies [28]. Pseudogenes have been previously identified in the Alpha-class gene cluster [30], and a Pi-class reverse-transcribed pseudogene has also been described [31].

The G to A transition at the first base of intron 2 of *GSTT2P* could result in defective splicing *in vivo*, as the junction no longer conforms to the splice site consensus. Examination of the human EST databases confirmed the presence of a cDNA which has not been spliced at the end of exon 2 and runs into the intron sequence (accession no. W68102). The identification of this abnormally spliced cDNA in a foetal heart library suggests that the duplicated pseudogene may be transcribed. A search of the same EST database has not identified any cDNA clones carrying the exon 5 stop codon. However, if the transcripts of this gene are abnormally spliced they may utilize a different poly(A) addition signal that may exclude exon 5 sequence from the mature mRNA. Even if a complete mRNA containing the exon 5 stop codon was transcribed from the duplicated gene, it is probable that the product, truncated by 49 residues from the C-terminus, would be inactive. The recently determined crystal structure of GSTT2-2 shows that the C-terminal residues are part of a helix that packs across the active site [32]. In studies to be published elsewhere, it has been shown that the mutagenic deletion of 23 C-terminal residues inactivates the GSTT2 enzyme (J. Flanagan and P. Board, unpublished work). Similarly, a C-terminal deletion of residues from the Alpha-class enzyme GSTA1-1 has been shown to severely reduce its activity [33]. Therefore, although there is some evidence that *GSTT2P* is transcribed, it is unlikely that a functional product is expressed.

The sequence analysis carried out in the present study has identified and determined the structure of a gene encoding DDCT lying head-to-head with the *GSTT2* gene. A similar gene occurs in the same relative orientation with respect to the *GSTT2P* gene. This indicates that the duplication that gave rise to *GSTT2P* was of sufficient size to include *DDCT*. We have not determined if both of the *DDCT* genes are functional. There are several sequences encoding DDCT in the EST database; however, at this stage we have not been able to determine if they are the products of either or both of the *DDCT* genes.

The proximity and orientation of the *GSTT2* and *DDCT* genes suggests that they may be under the control of a bidirectional promoter, such as those shown to occur within the mouse Surfeit locus [34,35]. The *Surf* genes are a closely linked cluster of unrelated genes that are arranged in head-to-head pairs. It has been suggested that the *Surf* genes may be regulated by *cis*-interactions, including the sharing of regulatory elements, antisense regulation and promoter occlusion [35]. Examples of other divergently transcribed genes include the Wilms tumour genes *WT1/Wit1* [36], the ataxia telangiectasia genes *ATM/E14* [37], the collagen genes *a1/a2* [38], the *Tap1/LMP2* genes [39] and the *BRCA1* and *NBR2* genes [40]. The Proscan search [24] of the DNA sequence between *GSTT2* and *DDCT* identified overlapping regions on each DNA strand that could function as a bidirectional promoter. Further studies are clearly required to determine if transcription of *GSTT2* and *DDCT* is co-ordinately regulated.

The conversion of D-dopachrome into 5,6-dihydroxyindole is the only known enzymic activity of the enzyme DDCT [22,23]. Since D-dopachrome is not naturally occurring, it is highly likely that the enzyme has another, yet to be discovered, function and may in the future be more appropriately renamed. Amino acid sequence alignments have shown a degree of similarity (33 % identity) between DDCT and the cytokine termed macrophage migration inhibitory factor (MIF). Since MIF has been shown to have DDCT activity, it is likely that DDCT may have a similar physiological role to that of MIF [41]. Interestingly, MIF was previously suggested to show sequence similarities to Mu- and Theta-class GSTs, and was reported to have GST activity [42,43]. Although an evolutionary relationship between MIF and the GSTs has been disputed [44], the occurrence of a closely related gene (*DDCT*) within the Theta-class gene cluster is, if nothing else, an interesting coincidence. Furthermore the metabolism of dopachrome has been linked to the GSTs in a previous study [45], where the Mu-class enzyme GSTM2-2 was shown to catalyse the conjugation of *o*-quinones such as dopachrome and adrenochrome to glutathione.

Approximately 16 % of Caucasians are homozygous for a deletion of the *GSTT1* gene [18]. The survey of polymorphisms in the *GSTT2* gene shown in Table 2 clearly indicates that *GSTT2* is not commonly deficient and that the deletion of *GSTT1* does not include *GSTT2*. The presence of a duplicated pseudogene makes formal genetic analysis of polymorphisms in the *GSTT2* and *GSTT2P* loci difficult. PCR–RFLP techniques amplify identical products from both loci, and it is difficult to distinguish reliably between homozygotes and heterozygotes based on DNA fragment staining intensity. It has been difficult to determine whether the duplication giving rise to *GSTT2P* is polymorphic or whether the sites within the *GSTT2P* gene are polymorphic. Since there are individuals in which the abnormal splice site at the start of intron 2 and the stop codon at position 196 are discordant, it appears most likely that the sites in the *GSTT2P* gene are polymorphic. A study of the relative hybridization intensity of the 3.5 kb *Bam*HI fragment failed to find any difference between individuals with and without the codon 196 stop sequence (results not shown). Thus it seems most likely that the *GSTT2P* gene occurs in all individuals and that the variant sites are polymorphic. During the present study one normal blood donor was identified with the *GSTT2P* gene, as indicated by the presence of the intron 2 splice defect, and showed no evidence for a normal gene. This individual may be deficient in GSTT2, and is under further investigation.

Although the Met-139 → Ile substitution in GSTT2 was found to be a rare polymorphism in Australian Europeans, we considered the possibility that it may influence the enzyme's function. Modelling of this substitution in the recently solved crystal structure of GSTT2 [32] indicated that residue 139 is located in helix 5 and occupies a hydrophobic pocket that is bounded by Phe-110, Tyr-181, Pro-176, Pro-113 and Leu-183. The sub-

stitution of isoleucine in this position is very unlikely to have a direct effect on the enzyme's function, however, subtle secondary effects are difficult to predict and cannot be excluded.

We are grateful to the Canberra Red Cross Blood Transfusion Service for providing the blood samples used in this study, and to Dr. G. Chelvanayagam for advice on the structural implications of the polymorphism in GSTT2-2.

## REFERENCES


1 Hayes, J. D. and Pulford, D. J. (1995) CRC Crit. Rev. Biochem. **30**, 445–600
2 Board, P. G. (1998) Biochem. J. **330**, 827–831
3 Bhargava, M. M., Listowsky, I. and Arias, I. M. (1978) J. Biol. Chem. **253**, 4112–4115
4 Homma, H. and Listowsky, I. (1985) Proc. Natl. Acad. Sci. U.S.A. **82**, 7165–7169
5 Board, P. G. (1981) FEBS Lett. **135**, 12–14
6 Seidegard, J., Pero, R. W., Miller, D. G. and Beattie, E. J. (1986) Carcinogenesis **7**, 751–753
7 Mannervik, B. and Danielson, U. H. (1988) CRC Crit. Rev. Biochem. **23**, 283–337
8 Mannervik, B., Awasthi, Y. C., Board, P. G., Hayes, J. D., Di Ilio, C., Ketterer, B., Listowsky, I., Morgenstern, R., Muramatsu, M., Pearson, W. R. et al. (1992) Biochem. J. **282**, 305–306
9 Meyer, D. J., Coles, B., Pemble, S. E., Gilmore, K. S., Fraser, G. M. and Ketterer, B. (1991) Biochem. J. **274**, 409–414
10 Meyer, D. J. and Thomas, M. (1995) Biochem. J. **311**, 739–742
11 Board, P. G., Baker, R. T., Chelvanayagam, G. and Jermiin, L. S. (1997) Biochem. J. **328**, 929–935
12 Pemble, S., Schroeder, K. R., Spencer, S. R., Meyer, D. J., Hallier, E., Bolt, H. M., Ketterer, B. and Taylor, J. B. (1994) Biochem. J. **300**, 271–276
13 Hussey, A. J. and Hayes, J. D. (1992) Biochem. J. **286**, 929–935
14 Tan, K. L., Webb, G. C., Baker, R. T. and Board, P. G. (1995) Genomics **25**, 381–387
15 Whittington, A. T., Webb, G. C., Baker, R. T. and Board, P. G. (1996) Genomics **33**, 105–111
16 Mainwaring, G. W., Nash, J., Davidson, M. and Green, T. (1996) Biochem. J. **314**, 445–448
17 Ogura, K., Nishiyama, T., Hiratsuka, A., Watabe, T. and Watabe, T. (1994) Biochem. Biophys. Res. Commun. **205**, 1250–1256
18 Chenevix-Trench, G., Young, J., Coggan, M. and Board, P. (1995) Carcinogenesis **16**, 1655–1657
19 Choo, K. H., Filby, G., Greco, S., Lau, Y.-F. and Kan, Y. W. (1986) Gene **46**, 277–286
20 Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) in Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
21 Breathnach, R. and Chambon, P. (1981) Annu. Rev. Biochem. **50**, 349–359
22 Odh, G., Hindemith, A., Rosengren, A.-M., Rosengren, E. and Rorsman, H. (1993) Biochem. Biophys. Res. Commun. **197**, 619–624
23 Nishihira, J., Fujinaga, M., Kuriyama, T., Suzuki, M., Sugimoto, H., Nakagawa, A., Tanaka, I. and Sakai, M. (1998) Biochem. Biophys Res. Commun. **243**, 538–544
24 Prestridge, D. S. (1995) J. Mol. Biol. **249**, 923–932
25 Webb, G., Vaska, V., Coggan, M. and Board, P. (1996) Genomics **33**, 121–123
26 Suzuki, T., Johnston, P. N. and Board, P. G. (1993) Genomics **18**, 680–686
27 Comstock, K. E., Johnson, K. J., Rifenbery, D. and Henner, W. D. (1993) J. Biol. Chem. **268**, 16958–16965
28 Pearson, W. R., Vorachek, W. R., Xu, S.-j., Berger, R., Hart, I., Vannais, D. and Patterson, D. (1993) Am. J. Hum. Genet. **53**, 220–233
29 Cowell, I. G., Dixon, K. H., Pemble, S. E., Ketterer, B. and Taylor, J. B. (1988) Biochem. J. **255**, 79–83
30 Klöne, A., Hussnätter, R. and Sies, H. (1992) Biochem. J. **285**, 925–928
31 Board, P. G., Webb, G. C. and Coggan, M. C. (1989) Ann. Hum. Genet. **53**, 205–213
32 Rossjohn, J., McKinstry, W. J., Oakley, A. J., Verger, D., Flanagan, J., Chelvanayagam, G., Tan, K.-L., Board, P. G. and Parker, M. W. (1998) Structure **6**, 309–322
33 Board, P. G. and Mannervik, B. (1991) Biochem. J. **275**, 171–174
34 Lennard, A. C. and Fried, M. (1991) Mol. Cell. Biol. **11**, 1281–1294
35 Colombo, P., Yon, J., Garson, K. and Fried, M. (1992) Proc. Natl. Acad. Sci. U.S.A. **89**, 6358–6362
36 Malik, K. T. A., Wallace, J. I., Ivins, S. M. and Brown, K. W. (1995) Oncogene **11**, 1589–1595
37 Byrd, P. J., Cooper, P. R., Stankovic, T., Kullar, H. S., Watts, G. D. J., Robinson, P. J. and Taylor, A. M. R. (1996) Hum. Mol. Genet. **5**, 1785–1791
38 Heikkila, P., Soininen, R. and Tryggvason, K. (1993) J. Biol. Chem. **268**, 24677–24682
39 Wright, K. L., White, L. C., Kelly, A., Beck, S., Trowsdale, J. and Ting, J. P. (1995) J. Exp. Med. **181**, 1459–1471
40 Xu, C.-F., Brown, M. A., Nicolai, H., Chambers, J. A., Griffiths, B. L. and Solomon, E. (1997) Hum. Mol. Genet. **6**, 1057–1062
41 Bjork, P., Aman, P., Hindemith, A., Odh, G., Jacobsson, L., Rosengren, E. and Rorsman, H. (1996) Eur. J. Haematol. **57**, 254–256
42 Blocki, F. A., Schlievert, P. M. and Wackett, L. P. (1992) Nature (London) **360**, 269–270
43 Blocki, F. A., Ellis, L. B. and Wackett, L. P. (1993) Protein Sci. **2**, 2095–2102
44 Pearson, W. R. (1994) Protein Sci. **3**, 525–527
45 Baez, S., Segura-Aquilar, J., Widersten, M., Johansson, A. S. and Mannervik, B. (1997) Biochem. J. **324**, 25–28


Received 14 April 1998/18 June 1998; accepted 8 July 1998