The Molecular Genetics and Evolution of the Glutathione *S*-transferases

by

Veronica Louise Ross

A thesis submitted for the degree of Doctor of Philosophy in the Australian National University

John Curtin School of Medical Research
Australian National University
Canberra

April 1994

For my sons, Simon and Patrick,
as a small recompense for the irrevocable changes
this thesis has wrought in their lives.

STATEMENT

This thesis describes the results of research undertaken in the Molecular Genetics Group of the John Curtin School of Medical Research, Australian National University, Canberra. This was accomplished under the supervision of Dr. Philip Board and Dr. Simon Easteal with an Australian National University Research Scholarship.

The results and analysis presented in this thesis are my own original work, except where otherwise acknowledged.

Veronica L. Ross

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my debt to my major supervisor, Dr. Philip Board, whose patience, as much as time and intellectual input, have contributed substantially to the culmination of this project. I would also like to thank Dr. Simon Easteal, for his substantial time and help, especially as he became head of a research group at this time. My thanks also to Dr. John Oakeshott for helpful suggestions. Dr. Graham Webb kindly allowed me to conduct the *in situ* hybridization experiments reported in Chapter 5 in his laboratory in the Queen Elizabeth Hospital, Adelaide. His guidance and aid through the procedures was critical to their success. I would also like to thank him for his generous friendship. Chapters 6 and 7 required the use of ANGIS, the Sydney University's on-line Australian National Genome System, and I am indebted for the extensive support of its staff, especially Dr. Alex Reisner and Carolyn Bucholtz.

Help and advice was always unstintingly provided by Drs. Taka Suzuki and Rohan Baker from the Molecular Genetics Group. I would like to thank members of both the Molecular and Human Genetics Groups, all of whom provided a friendly, happy and professional atmosphere in which to work, assistance whenever it was needed and advice when required. My thanks to Professor Sue Serjeantson, of the Human Genetics Group, who provided several pieces of sage council at critical times.

Marj Coggan, Senior Technical Officer in the Molecular Genetics Group, provided superb technical expertise, administrative support, lots of fun and a friendship that I hope will endure. Thanks must also go to fellow PHD students, Juleen Cavanaugh and Wendy Baker, for their special friendship, help and advice.

I would like to acknowledge my appreciation to my husband, John Magee, for his encouragement and support for the initiation of this adventure, and to my sons, Simon and Patrick, a significant percentage of their lives have revolved around this monumental project. It is difficult to thank my family sufficiently for their tremendous support, my mother, Louisa Ross, who toils for me when she comes "on holiday", my father, Rodger Ross and my brothers, Brian and Daniel, for being always willing, no matter what the request. Thanks also to my mother-in-law, Shirley Magee, for so often stepping into the breach.

This list could be endless, what I have done to deserve the friendship and support of so many people I can't fathom, but I thank them all.

ABSTRACT

The Glutathione *S*-transferases are of wide ranging interest for several reasons. GSTs are known to play a significant role in protecting the cell from a combination of xenobiotic and endogenous challenges, they have also been implicated in a range of binding functions, both for the purposes of detoxication and for storage of several important metabolic products. There are indications that some GSTs may also be involved in particular biosynthetic pathways. Studies have associated deficiency of GST with the incidence of cancer and the over-expression of GSTs with insecticide or herbicide resistance.

GSTs have been revealed as a large multigene family exhibiting a long evolutionary history, with activity reported from almost every species tested, indeed a multiplicity of GSTs is common in most species. This, in conjunction with an apparent constancy of the glutathione conjugating function is indicative of a significant metabolic role that is probably critical to survival. Additionally, as the variety of species represented by sequence data increases the evolutionary history of this diverse multigene family has a potential utility in molecular evolutionary studies.

Whilst previous studies in the Mu class had revealed a complex family of related enzymes there was a clear need for further characterisation and understanding of their genetic relationships. The intentions of this study were to characterise further members of the human Mu class, to map these chromosomally and to assemble the information available on the GST supergene family to gain some understanding of the evolutionary relationships within this family.

The characterisation was performed by the isolation and sequence analysis of cDNA clones, the production of expression constructs and the analysis of the subsequently purified recombinant proteins. *In situ* hybridization experiments were conducted to determine chromosomal location. Evolutionary relationships were analysed among all currently available full length cDNA sequences and specifically within the Alpha, Mu and Pi class sequences. This was achieved by the calculation of genetic distances, construction of phylogenetic trees and testing of the validity of those trees.

1. Genetic characterisation:

Several clones which hybridized to a human Mu class probe were isolated from testis cDNA library screens. One clone was found to be a copy of the

cDNA encoding GSTM2, previously considered to be a muscle specific enzyme. Two of the clones were of cDNAs encoding a new Mu class glutathione *S*-transferase. Both clones are incomplete and appear to result from alternate splicing. Comparison with the known sequence of the rat Mu class subunit 4 gene permitted analysis of the sequences. One clone is missing the sequence encoding exon 4 and the second clone is missing the sequence encoding exon 8. The complete sequence of the previously undescribed isoenzyme can be deduced from the two cDNA clones. Alternate splicing may play a role in regulation of expression of the isoenzyme. The new gene was called GSTM4, in line with the new nomenclature system (Mannervik *et al.* 1992).

2. Characterisation of Mu class isoenzymes:

The cloning of cDNAs encoding GSTM2 and a novel human Mu class GST provided an opportunity to characterise all of the human Mu class enzymes available at that time. By combining components from both clones, pGST-T and pT7.3 a complete cDNA was constructed and the encoded GSTM4 protein expressed in *E. Coli*. Similar constructs were made of other human Mu class GST isoenzymes. This permitted the characterisation and comparison of the purified isoenzyme for the first time. In general, the recombinant GSTM4-4 enzyme has relatively low activity with all the substrates tested and this may explain why it has not been previously detected. GSTM1, on the other hand, had significantly higher activity with *trans*-stilbene oxide than any of the other Mu class tested.

3. Chromosomal mapping:

The chromosomal localisation of the human Mu class glutathione *S*-transferase genes has been complicated by two factors, the total number of genes was unknown and there is a polymorphism which results from the presence or absence of the *GSTM1* gene. In view of the apparent contradictions in the literature and the complexity of the human Mu class, a study was initiated to examine the extent of hybridization between the characterised human Mu class genes and, by the use of *in situ* hybridisation, to localise the Mu class GST gene family, in subjects with and without the *GSTM1* gene. A probe derived from GSTM4 hybridizes to genomic DNA generating the same hybridisation pattern as that observed when using a GSTM1 probe. The same GSTM4 probe cross-hybridizes with GSTM1, GSTM2 and GSTM3 cDNAs. *In situ* hybridzation with the GSTM4 probe localised a major region of hybridisation on chromosome band 1p13. These data

indicate that the human Mu class gene family is largely clustered and not dispersed on different chromosomes. The identical *In situ* hybridization patterns in individuals with or without the *GSTM1* gene suggests that this locus is a component of the Mu class GST gene cluster.

4. Evolutionary analysis

Since this study commenced even the number of classes of GSTs has increased, with the discovery of the Theta class being prompted by the identification of previously unrecognised enzymes and the proposal of the Sigma class, currently containing only cephalopod lens crystallins. There has also been a rapid increase in the total number of genes and pseudogenes identified and cloned within each class. The evolution of this supergene family has not as yet been studied in detail. Alignment of the 72 full length GST sequences available was based on the recently reported 3D structure of three mammalian proteins. The phylogenetic trees constructed suggested that there may have been an early duplication of an ancestral gene, the descendants of each of the duplicate genes being found in two fan-like arrays (sectors A and B), separated by an internal branch. Several species are represented in the analysis by sequences in each sector of the tree. This analysis indicates that the original duplication occurred prior to the animal/plant split and the separation of the Alpha, Mu and Pi classes was prior to the vertebrate radiation.

5. Origin of the GSTM1 polymorphism:

A more detailed analysis of the human Mu class sequences provides corroborative evidence for earlier suggestions that gene conversion has occurred within the family. The GSTM1-null allele has a high frequency in most human populations and the evidence suggests that this is due to the absence of the *GSTM1* gene. There has been an assumption in the literature that this polymorphism is due to a deletion of the gene, however, there has also been a proposal that GSTM1 is the chimeric product of an unequal crossover between two pre-existing genes. Several models were tested for both an insertion or deletion origin for the GSTM1 polymorphism. Exon specific analysis of the sequences indicated that relationships among these sequences were complex and that neither the simple insertion or deletion models could fully explain these relationships. A complex insertion model has been evaluated, however, its very complexity makes it a less likely explanation. Each of the models makes some predictions for further testing.

PUBLICATIONS

Ross, V.L. and P.G. Board (1993). Molecular cloning and heterologous expression of an alternatively spliced human Mu class glutathione S-transferase transcript. *Biochem. J.* **294**:373-380.

Ross, V.L., P.G.Board and G.C.Webb (1993). Chromosomal location of the human Mu class Glutathione S-transferases. *Genomics* **18**:87-91.

Board, P.G., Ross, V.L., Coggan, M. and Suzuki, T. (1993) Polymorphic expression of human glutathione transferases. In: *Structure and function in the Glutathione transferases* Tew, K.D., Pickett, C.B., Mantle, T.J. and Hayes, J.D. (Eds.) CRC Press, Boca Raton, Florida.

Board, P.G., P.N. Johnston, V.L. Ross, G.C. Webb, M. Coggan and T. Suzuki (1991). Molecular genetics of the human glutathione S-transferases. In: H. Gelboin and L. Ernster (Eds.) *Xenobiotics and Cancer: Implications of Chemical Carcinogenesis and Cancer Chemotherapy*. Japan; Scientific Societies Press.

Board, P.G., M. Coggan, P.N. Johnston, V.L. Ross, T. Suzuki and G.C. Webb (1990) Genetic heterogeneity of the human glutathione transferases: a complex of gene families. *Pharmacology and Therapeutics*. **48**:357-369.

TABLE OF CONTENTS

CHAPTER 1 THE GLUTATHIONE *S*-TRANSFERASES — AN INTRODUCTION

1.1. CELLULAR DETOXICATION AND THE GLUTATHIONE *S*-TRANSFERASES

All living organisms are constantly exposed to compounds that are not only unnecessary to their continued existence but in many cases inimical to it. Many of these compounds are the products of intermediary cellular metabolism; some are a consequence of oxygen-based energy production. There are also many environmentally derived (xenobiotic) sources, which include micro-organisms and plants used as food. For example, as one source of potentially harmful compounds, it has been estimated that a "normal" American diet provides 1.5g of natural pesticides daily (Ames *et al.* 1990a). Relatively recently, an entirely new category of compounds has been generated by our industrial efforts (Davis 1987, 1989, Ames *et al.* 1990b). These compounds are now found in our food, air and water in increasing amounts. Irrespective of the generating source, many of these superfluous compounds are toxic and some are carcinogens. The necessity of removing these compounds from the cell has meant that multiple detoxification systems are found in all organisms.

Cellular detoxication can be carried out in several ways. Many potentially harmful xenobiotics are hydrophobic, having limited water solubility, and are therefore difficult to excrete. By increasing their water solubility it is possible to increase their excretion rate from the cell and ultimately from the body. Alternatively, electrophilic compounds, which otherwise can be highly reactive with cellular macromolecules, may be deactivated. Both these processes are catalysed by complex systems of enzymes.

The enzyme systems developed by organisms enabling them to deal with the toxic compounds that impinge on the cell may be divided into phase

I and phase II enzyme systems. Phase I enzymes catalyse transformation reactions such as hydrolysis, oxidation, and reduction; whereas phase II enzymes are primarily involved in conjugation reactions. Commonly, the biotransformation reaction which initiates detoxication introduces a functional moiety into the substrate and is mediated predominantly by the Phase I membrane bound cytochrome P450 enzyme system. It is then possible for Phase II enzymes, such as the Glutathione *S*-transferases, to conjugate a hydrophilic group to the functionalised substrate (Chasseaud 1979).

Instead of a specific enzyme for each molecular structure, detoxication systems contain families of enzymes with broad overlapping substrate specificities. If consideration is given to the number and variety of only the xenobiotic compounds needing detoxication it is not surprising that as a general rule, detoxication enzymes have a low specificity for their substrates. Many of these enzymes are capable of induction by some of their substrates. One such group of enzymes are the Glutathione *S*-transferases.

1.2 GENERAL CHARACTERISTICS OF GST ENZYMES

Glutathione *S*-transferase (**GST**) was first characterised separately as a glutathione (**GSH**) conjugating enzyme Booth Boyland and Sims (1961) and as a binding protein, named ligandin (Ketterer *et al.* 1967, Litwack *et al.* 1971). It was some years before ligandin and rat GST B were shown to be synonymous (Habig *et al.* 1974a). Since then many GST isoenzymes, mainly from humans and rats, have been isolated, purified and characterised and, in some cases, the nucleotide or amino acid sequences are known. The GSTs are a large family of multi-functional enzymes. Although both soluble and membrane bound GSTs have been characterised in several species, this study is primarily concerned with the soluble forms.

GST activity has been observed in the majority of aerobic species studied so far and frequently in multiple forms (Mannervik and Danielson 1988, Ketterer *et al.* 1988, Clark *et al.* 1989, Board *et al.* 1990 and section 1.6.1 below).

Multiple GST isoenzymes have been identified in all mammalian species investigated so far and they appear to be expressed in a developmentally regulated and tissue specific manner (Faulder *et al.* 1987, Strange *et al.* 1990, Board *et al.* 1990 and sections 1.6.2 & 3 below).

Mammalian GSTs have at least two distinct and well documented modes of action; catalysis and binding. These properties gave rise to the alternative names by which this enzyme was originally known and are discussed in section 1.4 below.

All cytosolic GST enzymes appear to be dimeric, consisting of combinations of identical or similar protein subunits. The only currently known exception occurs in *Tetrahymena,* which has been reported to express a monomeric form (Overbaugh *et al.* 1988). The only GST heterodimers observed *in vivo* appear to consist of subunits of a single class. Although it is possible to form inter-species heterodimers *in vitro*, these experiments have only been successful within a class (Mannervik *et al.* 1985).

1.3 NOMENCLATURE

The nomenclature of the soluble GSTs has been a subject of debate and concern among workers in the field for several years. In general, enzymes are named with reference to their substrates. Notwithstanding the fact that the substrates, and in particular the endogenous substrates, of GSTs are yet to be fully determined, several classification systems have been attempted. Each system has been based on differing criteria and this has been reflected in their terminology.

Initially, a classification attempt was made using substrate specificity as a determinant and each enzyme activity was named accordingly (e.g. alkyltranferase, aryltransferase, epoxidetransferase etc.; Boyland and Chasseaud 1969), but this was abandoned when overlapping substrate specificities were demonstrated. Kamisaka *et al.* (1975) divided the five human isoenzymes separated from liver, on the basis of isoelectric points,

whereas Habig *et al.* (1974b) proposed a classification of the rat cytosolic GSTs based on fractionation from a Carboxymethyl cellulose column. These systems have proved inadequate in coping with the diversity of GSTs that have been characterised subsequently.

There are now two major GST nomenclature systems in rats, both based on subunit composition: One is a simple numerical system, the subunits being numbered in order of characterisation, heterodimers reflecting subunit composition in their names, for example GST 3-4 (Mannervik and Jensson 1982, Jakoby *et al.* 1984). The alternative nomenclature developed from the identification of a binding protein named Y, as ligandin (Litwack *et al.* 1971) and then as GST (Habig *et al.* 1974a). Each enzyme is denoted Y with a letter designation for the individual subunit, for example Ya-Ya (Bass *et al.* 1977).

Board (1981a) made the first attempt to develop a nomenclature system predicated on genetic relationships following the observation that the multiple human isoenzymes separable on non-denaturing electrophoresis gels appeared to be the products of three genetic loci, termed GST1, GST2 and GST3. Other laboratories have provided further evidence in support of this system (Strange *et al.* 1984, Akiyama & Abe 1984, Laisney *et al.* 1984, Harada *et al.* 1987). Mannervik *et al.* (1983) also supported the division of the isoenzymes into three groups but used isoelectric points as a basis and designated them acidic, basic and near-neutral. Subsequently, Suzuki *et al.* (1987) added more loci, termed GST4, GST5 and GST6, to the genetic-based system. These, additional loci, showed some immunological relationship to the GST1 isoenzymes. By comparison of the biochemical and immunological characteristics (and in some cases, amino terminal sequences) of GSTs from both rats and humans, it was possible to generalise the system, providing confirmation that three major evolutionary classes exist (Mannervik *et al.* 1985). These are now designated Alpha, Mu and Pi, corresponding to the original three groups, GST2, GST1 and GST3 respectively, as described by Board (1981a).

It has been found possible to accommodate most mammalian cytosolic GSTs within this system. Recently however, a fourth class, Theta, has been added and both human and rat enzymes have been characterised (Hiratsuka *et al.* 1990, Meyer *et al.* 1991, Hussey and Hayes 1992) and two rat Theta class cDNAs have now been cloned and sequenced (Ogura *et al.* 1991, Pemble and Taylor 1992).

Sequence determination of a large number of GSTs from several mammalian species has led to widespread confirmation and acceptance of the class classification system. Nevertheless, the nomenclature of individual enzymes is still confusing. Recently, a large number of researchers in the field have proposed, and accepted, a new system of nomenclature for the human GSTs (Mannervik *et al.* 1992). It is based on primary structure similarities and a lower limit of 50% sequence similarity within a class is proposed. In the case of many of the known GSTs, sequence similarities are of the order of 70-90% within a class but between classes this similarity can be as low as 20-30% (Mannervik and Danielson 1988). The true utility of the new classification is in the incorporation of the individual isoenzymes and genetic loci into a single, logical system. Table 1.1 shows rat and mouse nomenclature systems as well as the new human nomenclature and some previous designations for different isoenzymes.

There are several possible difficulties with the new nomenclature; of prime importance is that the utility of the system may not extend significantly beyond mammalian species. Secondly, it is confined to cytosolic forms of GST, exceptions are the (membrane-bound) microsomal GSTs, which by most criteria fall outside the three well defined classes (Morgenstern *et al.* 1980, 1983, 1985, 1990, section 1.4.3.2). Lastly, some of those researchers working on non-human species may not find it advantageous to use similarity to the human isoenzymes as the base determinant, although the rat numerical system is compatible. These problems notwithstanding, the new

human-based

nomenclature system has considerable potential to simplify future discussions of the GSTs and will be used here. The species designations (eg HS = human) have been used in cross species comparisons.

Table 1.1 Different nomenclature systems for the GSTs in rats, mice and humans. Strict orthology is not implied and the species prefix is not used in this table.

New designations **human**	Previous designations **human**	**Class**	**rat**	**mouse**
GSTA1	e, GST2-type 1, Ha (subunit 1) αx-αx, B_1B_1	**Alpha**	subunit 1, Ya_1 Ya_2; ligandin*	M1, Ya_1Ya_2 subunit 4
GSTA2	g, GST2 - type 2 Ha (subunit 2) α_Y-α_Y, B_2B_2		subunit 2 Yc, AA	Ya_3
			subunit 8, Yk	
			subunit 10, Yl	
GSTM1a	μ, GST1-type2 Hb (subunit 4)	**Mu**	subunit 3 Yb_1, A	GT10, MIII subunit 1
GSTM1b	ψ, GST1-type1		subunit 4, Yb_2	GT2
GSTM2	GST4, muscle		subunit 6, Yn_1	
GSTM3				
GSTM4	GST-Tes			
	GST5, brain GST6			
GSTM5				
			Yb_4	GST9.3 subunit 2
GSTP1	π, GST3	**Pi**	subunit 7, Yp	MII, subunit 2
	GST q	**Theta**	subunit 5, E	
			subunit 5*	
			subunit 12	
Microsomal	microsomal GST	**Microsomal**	microsomal GST	

compiled from Mannervik 1985, Ketterer et al. 1988, Mannervik et al. 1992
* this term also includes the heterodimer GSTA1-A2.

1.4 MODES OF ACTION

Although very little is known of the specific modes of action of GSTs in other taxa, the mammalian GSTs are known to act in at least two distinct ways; firstly, they may serve as intracellular binding and transport proteins and, in addition, they have several catalytic activities, the most extensively studied being the conjugation of GSH to electrophilic compounds. A broad range of compounds are utilised as substrates in one or other mode of GST activity. However, as alternative functions of these versatile enzymes have been

suggested, it is possible that GSTs may also be significant players in other metabolic pathways.

1.4.1 GSTs as binding proteins

It is thought that GSTs may use their binding capacity to function as intracellular ligand storage and transport proteins, especially for organic compounds with limited solubility, thus acting as the intracellular equivalent of albumin in the circulation (Litwack *et al.* 1971, Årias *et al.* 1976, Tipping and Ketterer 1978, Bhargava *et al.* 1978, Wolkoff *et al.* 1979, Harvey and Beutler 1982). GSTs bind with high affinity to a variety of hydrophobic compounds. In rat liver for instance, where basal levels are ≅10% of soluble protein (Jakoby and Habig 1980), the concentration of GSTs and their efficiency of uptake of lipophilic compounds is capable of playing a significant role in intracellular transport (Tipping and Ketterer 1981). Compounds in this category include leukotriene C_4 and xenobiotics such as polycyclic aromatic hydrocarbons and dexamethasone (for review see Listowsky *et al.* 1988). More significantly, GSTs have also been implicated in the uptake of bilirubin by the liver (Wolkoff *et al.* 1979) and the intracellular transport of haem (Husby *et al.* 1981, Harvey and Beutler 1982).

Nebert (1990) has proposed that some drug metabolising enzymes, such as the GSTs, bind to small bio-organic oxygenated molecules (such as steroid (Maruyama and Listowsky 1984), and thyroid hormones, (Ishigaki *et al.* 1989)) that act as signals for growth, differentiation, virulence or tumour promotion. In this way the drug metabolising enzymes control the levels of these molecules within the cell and may provide a method of steady-state regulation.

It has been thought that some GSTs may act as non-histone DNA binding proteins (Bennett *et al.* 1982). Although there are several studies providing evidence for the presence of multiple GSTs in the nucleus (Bennett *et al.* 1985, 1986, Tan *et al.* 1988), there are as yet only limited indications of a specific DNA

binding role for GSTs. *In vitro,* Yb subunits (subunits 1 and 2) have demonstrated an ability to associate with DNA (Bennett *et al.* 1982) while *in vivo* Yb has been shown to associate with DNA-containing structures (Bennett *et al.* 1985, 1986). The significance of this association as well as its function and mechanisms are unclear, however, rather than a specific binding function, this association may facilitate peroxidase activity involving DNA hydroperoxides (section 1.4.2.2 below).

Covalent binding of some reactive electrophilic compounds, apparently leading to their immobilisation and inactivation, has also been observed (Ketterer *et al.* 1967, Jakoby 1976, Hayes and Wolf 1988). This may be an additional protective mechanism provided by some of the GST family of enzymes, although very little is known about this aspect of GST activity.

1.4.2 Catalytic activities of the GSTs

GSTs are major representatives of phase II enzymes and are known to catalyse at least three different types of reactions, all of which are GSH dependent. These include conjugation of GSH, hydroperoxide reduction and double bond isomerism. There has been very little study of the last of these three reactions, it being confined to isomerisation of $\Delta^{5\text{-}3}$-ketosteriods, and the significance of the involvement of GSTs in steroid metabolism is unclear. So far the Δ^5-androstene-3,17-dione isomerase activity has only been shown to be a significant property of the Alpha class GSTs (Benson *et al.* 1977, Ålin *et al.* 1985). A further catalytic activity suggested for GSTs is fatty acid ethyl ester synthesis (Bora *et al.* 1989a, b and 1991), however, this has not been confirmed by other workers (Suzuki *et al.* 1990, Sharma *et al.* 1991, Board *et al.* 1993).

1.4.2.1 Conjugation

The only general criterion for compounds acknowledged to act as substrates for GSH conjugation is that they either possess an electrophilic centre or give rise to an electrophilic compound. GSTs catalyse the attack of

GSH on the electrophilic centre of both endogenous and exogenous electrophilic agents such as epoxides, alkyl and aryl halides and α,β-unsaturated aldehydes and ketones. The resulting GSH-conjugate is then capable of further degradation, ultimately producing a mercapturic acid, which can be excreted in bile or urine. Transformation of substrates into more soluble mercapturic acid products markedly increases the efficiency of excretion of potentially toxic compounds (for reviews see Boyland and Chasseaud 1969, Habig *et al.* 1974, Mannervik 1985). Thus GSH conjugation is the initial step in the mercapturic acid pathway, considered one of the most important detoxification processes (Boyland and Chasseaud 1969, Chasseaud 1979).

Phase I reactions rarely affect water solubility and in many cases the activity of the substrate is likewise unaffected, however, in some instances the products can be even more reactive than the original substrate. In contrast, Phase II conjugated products are usually less reactive, more polar and therefore more easily excreted. However, there are exceptions and the product of conjugation between GSH and a haloalkane (in this case, 1,2-dibromoethane) forms an extremely reactive alkylating agent (Dekant *et al.* 1988, for a review see van Bladeren 1988).

The model substrate used in assaying GST activity is 1-chloro,2,4-dinitrobenzene (CDNB). This has been found useful in characterisation and monitoring GST activity in purification procedures. Various model substrates have been used to differentiate enzymes of the mammalian classes, for instance ethacrynic acid for Pi class, *trans*-stilbene oxide for Mu class and cumene hydroperoxide for the peroxidase activity with Alpha class. But it is the relative rates of activity rather than absolute values which charaterise the class differences. Many of the model substrates used for *in vitro* studies to characterise GSTs, have no known biological significance.

The endogenous substrates of the GSTs are still relatively poorly characterised, but in line with the suggestion that GSH is a true antioxidant,

some of the acknowledged substrates are the products of oxidative stress (Mannervik and Danielson 1988). At present, xenobiotic substrates have been better documented and it is possible that naturally occurring xenobiotics have been, and are, a major source of substrates for this family of enzymes.

Because the same, or very similar substrates can be generated from several different sources, it is possible for substrates, such as lipid oxidation products (Ames 1983) and amino acid pyrolysis products (Saito *et al.* 1983) to be derived from both endogenous and exogenous sources. Another example, acrolein, is found as a metabolite of several compounds in cells, as a degradation product of the cytostatic drug cyclophosphamide, and also as an environmental pollutant. Acrolein is known to cause single-stranded breaks in DNA and many GSTs show conjugating activity and apparently also covalent binding activity with this compound (Berhane and Mannervik 1990).

1.4.2.2 Peroxidation

In addition to the conjugation activity of all GSTs, enzymes of the Alpha and Theta classes particularly, have been shown to also have peroxidase activity with lipid peroxides, linoleate and arachidonate hydroperoxides (Ketterer *et al.* 1988). Stenberg *et al.* (1991) have also reported rat GST8-8 having low, but significant activity towards hydrogen peroxide. It is the free fatty acids rather than the gyceryl esters which have been shown to be good substrates for the cytosolic GSTs (Ketterer *et al.* 1988). This has led to suggestions that GSTs may be part of a system for the protection from, and repair of, free radical damage to membrane phospholipids (Lawrence and Burk 1976, Burk *et al.* 1980, van Kuijk *et al.* 1987). Both the human Pi class GSTP1-1 and the rat 7-7 have been shown to have more peroxidase activity with linoleate and arachodonate hydroperoxides than with the model substrate cumene hydroperoxide (Ketterer *et al.* 1988). On the other hand, only in the case of GSTM3 has it been possible to attribute any peroxidase activity to the Mu class and this is limited (Campbell *et al.* 1990, Chapter 4).

* It should be noted that a complex pattern of expression is also charactertistic of such enzyme systems, providing each tissue with its own combination of enzymes (section 1.6.2).

Thymine hydroperoxide has been shown to be a substrate for GSH peroxidase activity, as has peroxidised DNA, which can be a product of free radical damage (Christopherson 1969). More recently, GST peroxidase activity with these substrates has been confirmed with the isolation of a Theta class nuclear enzyme shown to have significantly higher activity with DNA hydroperoxides than any other known GST (Tan *et al.* 1988). Ketterer *et al.* (1987) have suggested a role for GSTs in DNA repair mechanisms and this may be a reason for the association of GSTs with DNA structures.

1.4.3 Class substrate specificities

1.4.3.1 Cytosolic GSTs

The ability of GSTs to catalyse reactions toward a large number of structurally diverse substrates is a general characteristic of detoxication enzyme systems and is a consequence of a combination of two factors. Firstly, the presence in a cell of a multiplicity of isoenzymes, each of which has a broad but optimal substrate selectivity. Secondly, the striking tolerance of each of these enzymes for both the type of electrophilic functional group and the structure of the molecules it is capable of accepting.*

The Alpha class can act as an example. Rat GST isoenzyme subunit 1-1 shows high activity towards CDNB and N-acetyl-p-benzoquinone imine, intermediate activity with 4-hydroxynon-2-enal and lower, but still significant, activity with ethacrynic acid, *trans* -4-phenyl-3-buten-2-one (*t*-PBO) and 1,2-Dichloro 4-nitro benzene (DCNB) (Ketterer *et al.* 1988, Mannervik and Danielson 1988). However, Alpha class enzymes in general exhibit not only conjugating activity but peroxidase activity and binding capabilities. Many of the products of lipid peroxidation are α,β-unsaturated compounds which are highly cytotoxic. One such group are activated alkenes such as 4-hydroxyalkenals and although most isoenzymes tested had some activity, the Alpha class rat GST 8-8 showed by far the highest activity (Danielson *et al.* 1987).

Of the two human Alpha class homodimeric enzymes assayed, GSTA1-1 had approximately twice the activity compared to GSTA2-2 (Mannervik *et al.* 1985).

Many xenobiotic substrates are natural in origin and include, for example, the mycotoxin derivative, aflatoxin B_1-1,2-oxide. While it does not appear to be a particularly good substrate for rat Alpha class GSTs 1-1 and 2-2, a mouse Alpha class GST is reported to be considerably more active (Ramsdell and Eaton 1990, Hayes *et al.* 1992). Although this may be taken as a timely warning on the inherent problems of such cross-species comparisons, a novel rat Alpha class isoenzyme related to rat GST1-1, termed Ya_3, has recently been observed to have high activity with the aflatoxin derivative (McLellan *et al.* 1991).

Conversely, many isoenzymes of the GST supergene family can also show marked substrate specificities. Highly reactive and carcinogenic compounds such as 7,8-styrene oxide which is example of a xenobiotic epoxide, are efficiently metabolised by human Mu class GSTM1-1 (Warholm *et al.* 1983). However, in the case of another epoxide, the same enzyme uses only *trans*-stilbene oxide as a substrate and activity with the *cis*- isomer is very low (Seidegård *et al.* 1985, Seidegård *et al.* 1987). Similarly, Robertson *et al.* (1986) reported 7,8-dihydrodiol-9,10-oxybenzo(a)pyrene with an R,S,S,R absolute stereochemical configuration is significantly more effectively conjugated by the human class Pi enzyme, but the corresponding enantiomers are equally active substrates for both the Pi and Mu enzymes tested. Class specificity is also seen in peroxidase activity where the thymine hydroperoxide, 5-hydroperoxymethyl-uracil, has been shown to be a substrate for Alpha GSTs. However, it is a novel Theta class enzyme 5*-5*, appropriately purified from the nuclear fraction, for which the highest activity for catalysing the detoxication of peroxidised DNA, has been reported (Tan *et al.* 1988).

The possibility of a role for GSTs in biosynthesis has been raised by the evidence that human Alpha class GSTs, can mediate the conversion of

prostaglandin (**PG**) H_2 to PG $F_{2\alpha}$ and PG E_2 whereas GSTA2-2 catalysed the isomerisation of the same precursor primarily to PG D_2 and PG E_2 Burgess *et al.* 1989). Although this has yet to be demonstrated *in vivo* it points to the possibility of the isoenzyme preferences of the GSTs extending beyond substrate specificities to product selectivities. There are only 11 amino acid differences between the two isoenzymes, some of which presumably code for this specificity. It was postulated that this product specificity may form the basis of a feedback system of regulation for the prostaglandins (Burgess *et al.* 1989) but the observed inter-individual variation in the levels of expression in these two isoenzymes (Board *et al.* 1981a, Strange *et al.* 1984) may be a complicating factor in this proposition.

Mu class GSTs have also been implicated in a synthesis pathway, in this case, synthesis of leukotriene C_4 (Mannervik *et al.* 1984, Tsuchida *et al.* 1987, Abramovitz *et al.* 1987). Although the product, *in vitro* ,was shown to be the same as that which is produced *in vivo,* it has been suggested that a specific enzyme, leukotriene C synthase, is probably responsible for the reaction *in vivo* (Söderström *et al.* 1988).

GSTs are clearly capable of several other activities besides conjugation and some of their versatility and effectiveness may depend on the interaction of these activities. One example of this is the ability of GSTs to both conjugate and covalently bind acrolein as mentioned above. Whilst lipid hydroperoxides that escape the initial peroxidase catalysis by GSTs may produce cytotoxic hydroxyalkenals among their decomposition products, these are also known conjugation substrates of GSTs (Slater 1984, Danielson *et al.* 1987, Stenberg *et al.* 1992). Thus GSTs are able to act at more than one point, on differing substrates, in the cell's protection system against free radical damage of membrane phospholipids. Table 1.2 summarises some of the compounds known as substrates of the different classes of GST enzymes (for reviews see Ketterer *et al.* 1988 and Mannervik and Danielson 1988).

Table 1.2 Some of the known substrate preferences of the Glutathione *S*-transferase classes. The designation "GST" has not been used in each name in this table due to space limitations. HS = *Homo sapiens*, MM = *Mus musculus*, RN = *Rattus norvegicus*. Units are µmol/min/mg unless otherwise stated.

Substrates	Alpha		Mu		Pi		Theta
binding							
bilirubin	HSA1 & A2						
haeme	HSA1 & A2						
steroid hormones	HSA1 & A2						
thyroid hormones	HSA1 & A2						
leukotriene C_4	HSA1 & A2						
conjugation							
1-Chloro-2,4-dintro benzene	HSA1	55	HSM1	90[d]	HSP1	72[d]	
	HSA2	60	MM 3	148[i]	MM 7	119[k]	
	MM 1	19[k]					
1,2-Dintro-4-nitrobenzene			MM 3	4[k]			
epoxides eg *trans*--stilbene oxide			HSM1				
N-Acetyl-*p*-benzo quinone imine	RN 1	24[d]	RN 3	6[d]	RN 7	60[d]	
	RN 2	48[d]	RN 4	3[d]			
ethacrynic acid	RN 2	2[d]	HSM2	2[f]	HSP1	2[d]	
	RN 8	18[d]	RN 4	1[d]	RN 7	4[d]	
					MM 7	1[k]	
trans, trans-deca-2,4-dienal			HSM2	0.52[h]			
trans-non-2-enal			HSM1	1.48[h]			
D^5-ketosteroid isomerase activity							
eg D^5-androstene-3,17-dione	HSA1 & A2	0.23[d]	MM 3	0.04[k]	MM 7	0.14[i]	
	MM 1	0.07[d]					
		0.04[k]					
peroxidase activity							
cumene hydroperoxide	HSA1	5[b]	HSM3	1[g]			RN 5 13[d]
	HSA2	16[b]					
	RN 1	2[d]					
	RN 2	3[d]					
	RN 8	12[d]					
	MM 1	11[k]					
t-butyl hydroperoxide	HSA1	1[b]					
	HSA2	8[b]					
lipid hydroperoxides	HSA1	5[d]			RN 7	2[d]	RN 5 5[d]
eg linoleate hydroperoxide	HSA2	2[d]					
	RN 1	3[d]					
	RN 2	2[d]					
DNA hydroperoxide			HSM1	82[a]	P1	9[a]	RN 5 500
			RN 3	18[a]	RN 7	5[a]	RN 5* 1500
			RN 4	32[a]			
			RN 6a	9[a]			
prostaglandin H_2	HSA1[c]						
	HSA2[c]						
4-hydroxynon-2-enal	RN 1	3[e]	RN 3	3[e]	MM 7	3[k]	
	RN 8	170[e]	RN 4	7[e]			
	MM 1	1[k]	MM 3	6[k]			

[a] Tan *et al.* 1988 nmol/min/mg.
[b] Chow *et al.* 1988.
[c] Burgess *et al.* 1989 nmol/min/mg.
[d] Ketterer *et al.* 1988.
[e] Mannervik *et al.* 1985.
[f] Singhal *et al.*
[g] Campbell *et al.* 1991.
[h] Chapter 4
[k] Warholm *et al.* 1986

Theta class has only been recently defined as enzymes of this class have a very low activity, or no activity with CDNB. They are also characterised by being labile and of low abundance, possessing a high K_m for GSH and in

purification procedures they fail to bind to glutathione affinity columns (Habig *et al.* 1974a, Hiratsuka *et al.* 1990, Ogura *et al.* 1991 and Meyer *et al.* 1991). This is probably why most Theta enzymes have gone unnoticed for so long and may point to there being more, as yet unreported, forms of GST with differing substrate specificities. Rat GST5-5 has higher conjugating activity with several model substrates, 1,2-epoxy-3(*p*-nitrophenoxy)propane, *p*-nitrophenethyl bromide, *p*-nitrobenzyl chloride and dichloromethane than those of other known rat GSTs (Meyer *et al.* 1991). Reported activity of this enzyme with cumene hydroperoxide is also very high, however, previous reports of high levels of activity with DNA hydroperoxide seem to have been due to contamination (Meyer *et al.* 1991), confirming that rat GST5-5 is not the high activity DNA hydroperoxide Theta enzyme, 5*-5*, reported by Tan *et al.* (1988). Isoenzyme Yrs-Yrs (subunit 12-12, Meyer *et al.* 1991), is reported to have high activity towards reactive sulphate esters (Hiratsuka *et al.* 1990).

1.4.3.2 Microsomal GSTs

Many of the compounds known as conjugation substrates for the cytosolic GSTs are also good substrates for the microsomal GSTs and they are apparently more effective catalysts of the toxic environmental contaminants, polyhalogenated hydrocarbons (Morgenstern and DePierre, 1983, 1988, Morgenstern *et al.* 1990). Peroxidase activity towards fatty acid hydroperoxides is also significant (Ketterer *et al.* 1988) and may form the basis of an '*in situ*' membrane protection system against the damage caused by free radicals. Microsomal GSTs appear to possess binding capabilities and it has been suggested that this functions to protect against inhibitory effects of these substrates on other enzymes embedded in the membrane (Morganstern and Depierre 1988). Microsomal GSTs are only slow catalysts of leukotriene C_4 synthesis (Morganstern and Depierre 1988) however, initial indications are that the enzyme, leukotriene C_4 synthase, recently purified from mitochondria may be a protein related to microsomal GST (Penrose *et al.* 1992).

1.5 GENE STRUCTURE AND REGULATION

Genomic clones from rat GST genes have been isolated and characterised (Telakowski-Hopkins *et al.* 1986; Okuda *et al.* 1987; Lai *et al.* 1988; Morton *et al.* 1990), including representatives from each of the Alpha, Mu and Pi classes. There are marked differences in gene structure between the three classes with gene sizes of 11 kb for Alpha class subunit 1 (Telakowski-Hopkins *et al.* 1986), 5 kb for the Mu class subunit 2 gene and 3 kb for the Pi class subunit 7 (Okuda *et al.* 1987; Muramatsu *et al.* 1987). The genes encoding rat subunits 1 (Alpha class) and 7 (Pi class) have seven exons while the Mu class gene has eight exons.

The demonstration of common structural and catalytic properties among GSTs from rats, mice and humans led to the mammalian class classification system (Mannervik *et al.* 1985). More recently sequence determination of an increasing number of GSTs has provided widespread confirmation that isoenzymes from within a class exhibit greater similarities to isoenzymes which come from the same class but a different species than to isoenzymes of different classes within the same species. Thus within a class protein coding regions show approximately 70% to 90% sequence similarity, however, the 5′ and 3′ untranslated regions can be very different.

Genomic data from mouse (Daniel *et al.* 1987), rat (Telakowski-Hopkins *et al.* 1986) and human (Board *et al.* 1991, Klone *et al.* 1992, Rohrhanz *et al.* 1992) Alpha class genes show close conservation of intron-exon structure. Some intron sequences also show a significant conservation between the mouse and rat genes (Daniel *et al.* 1987).

The partial human Mu class clones also show both intron-exon structure and splice site conservation when compared with the rat Mu gene (Taylor *et al.* 1991). As with the Alpha class sequences, there is also some intron sequence similarity. This is particularly striking among the human sequences,

but also in comparison with the rat Mu class gene and has led to suggestions of gene conversion events (Taylor *et al.* 1991). Comstock *et al.* (1993) has very recently published the genomic sequence for GSTM4 confirming the intron sequence similarity reported by Taylor *et al.* (1991). Both the rat and human Pi class genes have been isolated and again there is significant gene structure and sequence similarity to the coding region (Okuda et al. 1987, Cowell *et al.* 1988).

Evidence from Southern blots of genomic digests with probes from different classes demonstrates the existence of several more hybridising fragments than are represented by known genomic or cDNA sequences at present (Rothkopf *et al.* 1986; Board and Webb 1987; DeJong *et al.* 1988a; Seidegård *et al.* 1988; Pearson *et al.* 1988). Several pseudogenes have been reported in humans (Board *et al.* 1990, Suzuki *et al.* 1993) and processed pseudogenes have been characterised both in rats (Okuda *et al.* 1987) and in humans (Board *et al.* 1989).

1.5.1 Gene regulation

The majority of GST genes analysed to date appear to be inducible and several chemical compounds have been shown to elevate the cytoplasmic levels of GSTs. Evidence, thus far suggests this regulation is primarily by augmenting the transcriptional rates of the respective genes (Ding and Pickett 1985). GST-inducing compounds include phenobarbital in rat (Ding and Pickett 1985), 2(3)-tert-butyl-4-hydroxyanisole (BHA) and disulfiram (bis(diethyl-dithiocarbamyl)disulfide) in mouse (Pearson *et al.* 1983, Wattenberg 1978, McLellen 1991). The effect of BHA, a dietary antioxidant, can be dramatic. Where mice were fed both the carcinogen, dimethylbenzanthracene, and 1% BHA, no tumours were found, whereas 55% of the control mice, receiving only the carcinogen, developed tumours (Pearson *et al.* 1988).

Telakowski-Hopkins *et al.* (1988) identified two *cis*-acting regulatory elements by progressive deletions from the promoter region of a CAT-GST

promoter construct of the rat subunit 1 gene. One was required for maximum basal level expression, and a second element was required for inducible expression. A specific Xenobiotic Response Element (XRE) has been implicated in induction of expression by planar aromatic compounds and an Antioxidant Response Element (ARE), responsive to electrophilic compounds, has been identified (Rushmore *et al.* 1990; Rushmore and Pickett 1990). The core XRE sequence has also been found in the 5′ flanking region of the cytochrome P450IA1 gene (Rushmore and Pickett 1990) and a similar structure has been found in the enhancer of the rat Pi class gene (Okuda *et al.* 1990). Similar XRE and ARE sequences have been found in a mouse Alpha class gene, however, the XRE sequence was not found to be active (Friling *et al.* 1990). Induction via the rat XRE appears to be dependent upon cytochrome P450 metabolism of planar aromatic compounds and the nuclear presence of the *Ah* or aryl hydroxylase, receptor (also called the dioxin receptor, Hapgood *et al.* 1989). In contrast, the ARE recognition sequence and the HNF1 appear to be *Ah* receptor independent (Friling *et al.* 1990; Rushmore *et al.* 1990).

Despite the sequence similarity between the rat and human Pi class isoenzymes there appears to be very little similarity in the regulation of these two genes. Expression of rat subunit 7 is highly inducible and currently the most useful marker available for hepatocellular carcinogenesis in rats (Satoh *et al.* 1985). Conversely, the human Pi subunit shows evidence of only a slight increase in the level of expression in similar circumstances (Dixon *et al.* 1989). Evidence has also been presented suggesting that during chemical hepatocarcinogenesis, expression of the rat Pi gene may, in part, be regulated by the product of the *c-jun* oncogene (Sakai *et al.* 1990) and that there is a significant increase in expression of GSTP1 in several types of tumour (Satoh *et al.* 1985, Di Ilio *et al.* 1988, Awasthi *et al.* 1988).

More recently several candidate *trans*-acting proteins have been implicated in the transcriptional regulation of the rat Alpha class gene. Two of

these, Hepatocyte Nuclear Factors 1 and 4 (HNF1 and 4) contribute to basal level transcription and liver specific expression, respectively (Paulson *et al.* 1990).

Evidence is mounting of the influence of hormones on GST expression and this appears to be not only organ specific but also subunit specific. Sex differences in subunit composition within specific tissues have been reported in rats (McLellan and Hayes 1987, Igarashi *et al.* 1987, Listowsky *et al.* 1990) and regulation of the expression of the major hepatic Pi class GST by testosterone has also been observed in mice, with ten-fold higher levels of this enzyme in males compared to females (Hatayama *et al.* 1986). Further evidence for this role has been reported by Moscow *et al.* (1988a and b), who, while investigating multi-drug resistance in breast cancer cells, observed an inverse association between human GSTP1 expression and the estrogen receptor concentration of cells. Chang *et al.* (1987) demonstrated that rat GST subunit 3 transcription is repressed by androgens and therefore proposed that GST may play a role in the regulation of growth in androgen sensitive organs. Since a rat Mu class (GST subunit 3) isoenzyme has been found to co-localise with U-snRNPs (a family of small nuclear RNA/protein complexes which have been implicated in mRNA splicing) and both are found within interchromatinic nuclear domains (Bennet *et al.* 1986), this regulation of growth by androgens may be mediated by modification of RNA production or processing of specific genes (Chang *et al.* 1987). Whilst the mechanism, function and significance of the interaction between sex hormones and GSTs is not clear, the implication of these enzymes in growth regulation further extends the role of GSTs and warrants further attention.

1.5.2 Protein structure

As mentioned above all mammalian cytosolic GSTs appear to be dimeric, consisting of identical subunits or subunits from the same class. Predictions of the protein tertiary structure from circular dichroism suggested

that the GSTs have an α-helix/β-sheet type structure (Kamisaka *et al.* 1973, Warholm *et al.* 1983). Persson *et al.* (1988) who based their deductions on hydropathy profiles and Chou-Fasman computer predictions of primary sequences, suggested a similar structure for representatives of each of the three major classes. Both these techniques have their limitations and for a review of the difficulties involved in using sequence data to predict secondary structure see Lesk *et al.* (1992) or Rost *et al.* (1993). Nonetheless, the recently solved 3D structures do substantially corroborate the earlier predictions as both the porcine and human Pi class, enzymes, the rat Mu class subunit 3 enzyme and the Alpha class GSTA1-1 are considered to be composed of two major domains, an alternating α-helix/β-sheet structure and a helical domain (Reinemer *et al.* 1991, 1992, Ji *et al.* 1992, Sinning *et al.* 1993). The N-terminal α/β-domain 1 is considered to have some structural similarity to both bovine Se-dependent glutathione peroxidase (Ladenstein *et al.* 1979) and T4 thioredoxin (Holmgren *et al.* 1975), all three enzymes being dependent on GSH.

Single subunits appear to be inactive and so it was assumed that the active site lay in the cleft between the subunits. But investigations into the kinetics of both homo- and heterodimeric enzymes suggested substantially kinetically independent functions for each of the subunits in a dimer (Tahir and Mannervik 1986). The G-site has a high specificity which is demonstrated by the very few compounds which can substitute for GSH or act as inhibitors at this site (Adang *et al.* 1988, 1989, 1991, Andersson 1991, Danielson and Mannervik 1985). Solution of the 3D structures show the active site faces the cavity formed at the interface of the monomers, each monomer having an active site at opposite ends of the V-shaped crevice which spans the dimer interface (Reinemer *et al.* 1991, 1992, Ji *et al.* 1992, Sinning *et al.* 1993). The GSH binding site (G-site) is lined with polar residues from domain 1 (which includes the N-terminal portion and in the Alpha class also 30 residues of the C-terminal). Many of the residues involved in the G-site are conserved in the other two classes but the specific interactions differ slightly. The residues considered important in GSH binding

and are conserved in the three enzymes whose structures have been solved are Tyr9, Gln67, Asp101 and Glu104 (numbered on HSGSTA1 and according to Sinning *et al.* 1993). When bound to S-benzyl-glutathione, the Alpha class HSGSTA1 is non-covalently bound to the benzyl group in three different regions. The residues in those regions that are conserved in the three classes of GSTs are Gly14 and Glu104 of a possible 15 residues involved. The broad range of second substrates catalysed by GSTs makes it a difficult task to predict residues which may be important for this function across the three classes, indeed, it would be surprising if there was such conservation. One aspect of the crystallographic studies to be highlighted is the considerable overall three dimensional similarity found among each of the three classes. Thus the folding is considered similar and each subunit is built of two of domains that interact to form similar monomers, with approximately the same regions interacting to form the dimer (Sinning *et al.* 1993). Studies of the three representative enzymes show that the active sites are in spatially equivalent positions. However, the level of fit seen in a comparison of the the structures of individual domains is higher than when whole subunits are compared because of small, but significant, rotations between the domains (Sinning *et al.* 1993).

A more detailed insight into the mechanism of GST catalysis has been provided more by site-directed mutagenesis than from the 3D structures. GSTs catalyse the nucleophilic attack of the sulphur atom of GSH on electrophilic groups in a given substrate. It is thought that the activation of the sulphur is precipitated by the removal of the proton from the SH group by a base, Tyr9 in GSTA1-1. Removal of the OH group (Tyr9) reduces activity in Alpha to less than 10% and in Mu and Pi enzymes to below 1% (Stenberg *et al.* 1991, Liu *et al.* 1992, Kolm *et al.* 1992). Gly14 (in GSTA1-1) is considered crucial as it is present in an unusual conformation with its peptide oxygen pointing away from Tyr9. This prevents close contact of the thiolate and the hydroxyl group of Tyr9 and subsequent destabilisation.

These findings are in contrast to the demonstrated flexibility of the second substrate binding site (H-site) which is made up of residues from the C-terminal domain 2 and is highly hydrophobic. The hydrophobic pocket in GSTA1-1 would accomodate about 10 atoms, so larger compounds would presumably protrude into the solvent channel between the monomers.

1.5.2.1 Posttranslational modification

Although the the majority of regulatory mechanisms described for GSTs so far are pre-transcriptional there are some reports of posttranslational modifications. One such posttranslational modification results in the blocking of the N-terminus, preventing amino acid sequencing by Edman degradation and there is evidence that the amino group of the initiator methionine residue is acetylated in the rat 8-8 (Yk-Yk) isoenzyme (Lin *et al.* 1989) which may cause the blocking effect. Whether N-terminal blocking is generally caused by this type of modification is yet to be demonstrated. The specific function of this modification is also not apparent as blocked enzymes, when expressed in *E. coli*, and therefore without the modification, show no detectable alteration of function (Board and Pierce 1987). It is of interest to note that of those subunits in which the N-terminal has been determined directly from the protein, all have been reported as missing the initiating methionine.

In vitro studies have suggested that it is possible for rat Alpha class GSTs 1-1, 2-2 and 1-2 to be phosphorylated by protein kinase C (Taniguchi and Pyerin 1989) and that Mu class GSTs 3-3, 3-4, 4-4 could be methylated (Johnson *et al.* 1990), but the biological significance of these modifications needs clarification.

There is some evidence that expression of the human Pi class enzyme GSTP1-1 may be partially regulated by a post-transcriptional mechanism (Morrow *et al.* 1992). Studies by several groups in a range of tissues have provided evidence of at least two human Pi class isoenzymes (Marcus *et al.* 1978, Suzuki *et al.* 1987, Tsuchida *et al.* 1990, Singh *et al.* 1988), the only readily observable differences between the forms being minor differences in p*I*. Kuzmich *et al.* (1991) have provided evidence of glycosylation in both the rat

and human Pi class isoenzymes and suggest this may be a factor in the microheterogeneity observed in these isoenzymes. However, Aceto *et al.* (1989) found two immunologically similar Pi class subunits which were of different sizes in human testis. It is possible that these may represent the presence of multiple Pi class subunits showing restricted tissue specificity, but evidence from Southern blots has indicated a limited number of hybridising bands, all but one of which (a very weakly hybridising band) can be accounted for by GSTP1 itself or a charaterised pseudogene (Board *et al.* 1989, section 1.7.3).

1.6 DISTRIBUTION

1.6.1 Taxonomic distribution

The co-substrate of the GSTs, GSH, is confined essentially to aerobic organisms and is considered to have evolved as a reducing agent in an oxidative environment (Fahey 1977, Meister 1983). It is unlikely then that the GSTs would be present without their co-substrate in anaerobic species (Mannervik and Danielson 1988). Given that limitation, however, GST enzymes or GST enzyme activity have been found in every mammalian species studied and it has been demonstrated that this enzyme activity extends throughout the major aerobic phyla. Examples have been described in algae, yeast and some bacteria (Lau *et al.* 1980, Tamaki *et al.* 1991), the trematode *Schistosoma,* and many insect species (Smith *et al.* 1986, Balloul *et al.* 1987, Taylor *et al.* 1988, Toung *et al.* 1990) although chicken is the only representative of the birds at this point (Yeung and Gidari 1980, Liu *et al.* 1991, Chang *et al.* 1992). GSTs are also found in plants, some showing both constitutive GST activity and increased expression of GSTs associated with the administration of some insecticides and herbicides (Lammoureux and Frear 1979, More *et al.* 1986). For a detailed review of non-vertebrate GST distribution and enzymology see Clark (1989). Cross-reactivity to polyclonal antibodies raised

1 It is possible that the failure to find GST activity in some species may be due to the use of unsuitable substrates and it is clear that the Theta class mammalian enzymes were not recognised earlier, partially because of their lack of activity with CDNB.

against the rat microsomal GST has been reported in most mammalian species examined but no reactivity was detected in chicken, toad or pike (Morgenstern and Depierre 1988).

It should be noted that most studies on non-mammalian species have used CDNB assays for GST detection, thus reports of comparative activity levels and even isoenzyme variation should be treated with some caution.[1] Species distribution is discussed in more detail in Chapter 6.

1.6.2 Tissue distribution

Where GST is found, it usually exists in multiple forms. In humans, GSTs can be found in most tissues, including heart, skin, brain, pancreas, and testis (Suzuki *et al.* 1987, Del Boccio *et al.* 1987a). Levels of overall expression vary from high values in the liver and testis to very low levels in the non-lactating mammary gland and negligible levels in epidydimal sperm (Meyer *et al.* 1984). The liver, site of many detoxication reactions, is the richest source of GSTs, both in terms of numbers of isoenzymes and in absolute amounts (Ketterer *et al.* 1988).

It has been observed that each tissue has a unique isoelectric focussing profile of GST isoenzymes (Mannervik *et al* 1983, Strange et *al.* 1984, 1990, Laisney et *al.* 1984, Suzuki *et al.* 1987). The relationship between isoenzyme specificities and substrate distribution must play a role in this striking difference in isoenzyme pattern from one tissue to another, however, its full significance is not clearly understood. The extent of the variation observed in isoenzyme expression has been enhanced by immunohisto-chemical techniques which show variation in expression between different cell types within specific tissue (Redick *et al.* 1982, Mannervik *et al.* 1990). Even within the cell itself a different suite of isoenzymes has been observed in the nucleus compared to the cytoplasm (Tan *et al.* 1988).

In mammals the Alpha class enzymes are predominantly expressed in the liver and kidney and at a much lower, and more variable rate in many other tissues (Meyer *et al.* 1984, Suzuki *et al.* 1987). The Pi class enzymes are the

most abundant form found in many tissues including placenta, erythrocytes, testis, muscle and heart but only in a relatively small amount in the liver (Suzuki *et al.* 1987). In humans the Mu class contains several distinct isoenzymes and these all appear to show very specific tissue expression. The microsomal form of GST has been shown to localise on the endoplasmic reticulum and outer mitochondrial membranes and is found in several orders of magnitude more abundance in the liver compared to any other organ (Morgenstern and Depierre 1988).

1.6.3 Developmental variation

Developmental specificity of GST expression has been well documented in humans (Fryer *et al.* 1986, Faulder *et al.* 1987, Strange *et al.* 1990) and rodents (Hatayama *et al.* 1986). A detailed study of the amphibian *Bufo bufo* has shown a complex developmental pattern of expression by a range of GST isoenzymes (Del Boccio *et al.* 1987b). In the adult liver of both rats and humans, Alpha and Mu class isoenzymes predominate, but in the foetal liver, the Pi class isoenzymes are dominant and only low Alpha class activity is detectable (Faulder *et al.* 1987, Strange *et al.* 1990). In human foetal liver Mu class activity was detected only after 30 weeks gestation (Warholm *et al.* 1981, Guthenberg *et al.* 1986, Fryer *et al.* 1986, Pacifici *et al.* 1986). The Alpha class isoenzymes expressed in human adult kidney do not appear until 10 weeks after birth (Faulder *et al.* 1987, Strange *et al.* 1990).

1.6.4 GSTs and cancer

It is thought that the carcinogenic process may be initiated with covalent damage to DNA by electrophilic metabolites or xenobiotics. In mammalian cancers, GSTs are of interest for several reasons. Firstly, GSTs, as a family of multifunctional enzymes involved in the detoxication of electrophilic compounds, play a role that is perceived to be protective against such damage. However, under certain conditions, GSTs may be involved in the

promotion of cancer. In one such situation, as has already been mentioned, certain compounds are, in fact, made more reactive by conjugation with GSH, thus providing a possible source of carcinogenic compounds (section 1.4.2 and van Bladeren 1988).

Secondly, genetic differences in the levels of expression of Phase I and Phase II enzymes have been associated with an increased risk of cancer, in particular, the genetically determined deficiency of GSTM1 has been suggested as a risk factor, providing increased sensitivity to particular chemical carcinogens (Tsuchida and Sato 1992, Board 1981a and b). Supporting this, studies by Seidegård *et al.* (1986, 1990) and Hayashi *et al.* (1992) have found an increased frequency of the GSTM1 deficiency in lung cancer patients. Further, correlations between GSTM1 deficiency and increased sister chromatid exchange have been reported (Wiencke *et al.* 1990, van Poppel *et al.* 1992). Thus induction of cytogenetic damage by compounds detoxified by GSTM1, when present, may be one explanation of the observed susceptibility to cancer. These differences in sensitivity may also be of importance to the incidence of other cancers, such as stomach and colon adenocarcinomas (Strange *et al.* 1991) and skin cancers (Heagerty *et al.* 1994). The correlation between GSTM1 deficiency and cancer susceptibility has not been found in all studies, the frequency of the GSTM1-null is not significantl different between patients with breast cancer and controls (Shea *et al.* 1990, Forrester *et al.* 1990), in addition, Zhong *et al.* (1993) also failed to find a correlation in bladder cancer and Brockmoller *et al.* (1993) failed to find a correlation in lung cancer.

Conversely, a great deal of interest has been focused on GSTs because they have been shown to conjugate some cytotoxic drugs. There have now been many studies reporting elevated levels of Pi class GSTs in tumours such as metastatic liver tumours originating from the stomach, colon and gall bladder cancers (Satoh *et al.* 1985) and in drug resistant cell lines (Wang *et al.* 1985, 1989, Batist *et al.* 1986, Shea *et al.* 1988, 1990, Smith *et al.* 1989, Wolf *et al.* 1990, Tew *et al.* 1990, Tsuchida and Sato 1992).

1.7 GENETIC HETEROGENEITY OF THE HUMAN GSTS

From the earliest studies of the GSTs it has been evident that there is considerable heterogeneity and it has been established that multiple isoenzymes are expressed in mammals. In humans, variation in the occurrence of different GST isoenzymes have been observed between populations, individuals (Board 1981a and b, Strange *et al.* 1984), stages of development (Strange *et al.* 1985), different tissues (Suzuki *et al.* 1987), cell types within a tissue (Redick *et al.* 1982, Mannervik *et al.* 1990) and even between organelles (Tan *et al.* 1988). Some of this observed heterogeneity may be accounted for by the dimeric nature of the enzymes, which allows the formation of either homodimers or heterodimers from the individual subunits within the classes or by suggested posttranslational modifications such as glycosylation (Kuzmich *et al.* 1991). Moreover it has been suggested that deficiencies of GSTs may be associated with an increased risk of cancer or, conversely, many studies have reported over expression of GSTs in tumours and drug resistant cell lines (section 1.6.4).

Despite the implications of these reports and the considerable level of research in the area, the mechanisms regulating the complexity of tissue and developmental expression, substrate induction and the requirements for the multiplicity of enzymes are still to be fully explained.

1.7.1 Alpha class

The human Alpha class consists of several basic (p*I* 8.3-8.9) isoenzymes and these are found primarily in the liver and kidney although there are minor levels of expression in other tissues (Board 1981a, Strange *et al.* 1984, Suzuki *et al.* 1987). It is the Alpha class enzymes that are blocked at the amino terminal preventing sequencing by Edman degradation.

Kamisaka *et al.* (1975) purified a range of basic enzymes from human liver and suggested that they may be the products of post-translational modification of a single subunit. This proposition was supported by later work

(Laisney *et al.* 1984, Strange *et al.* 1984). However, Board (1981a) suggested the three major forms of the basic GSTs may be the products of alleles at a single genetic locus. The molecular cloning of the cDNAs of these isoenzymes has confirmed that they are very similar, with only 11 amino acid differences and that they are actually the homo- and heterodimeric combinations of two similar subunits (GSTA1- and GSTA2) encoded by separate genes (Tu and Quian 1986, Board and Webb 1987, Rhoads *et al.* 1987, Suzuki *et al.* 1993). Although the genes encoding the two isoenzymes are believed to be closely linked (Board and Webb 1987), the factors responsible for the differential expression of the two isoenzymes between individuals are not fully understood, and whether these factors are genetic, environmental or a combination is still to be determined.

There is evidence of other Alpha class genes in human genomic Southern blots (Board and Webb 1987) and analysis of genomic clones indicates there are several exon sequences with similarities to those reported for GSTA1 and 2 (Board *et al.* 1991), some of which have now been reported to be pseudogenes (Suzuki *et al.* 1993). In addition, Del Boccio *et al.* (1987a) have characterised a very basic isoenzyme from human skin which shows a high level of similarity to rat Alpha class subunit 2.

Although both individual and developmental variation in the relative expression of GSTA1-1 and GSTA2-2 have been reported (Board 1981a, Strange *et al.* 1985) deficiencies of the Alpha class genes in humans are very rare, with only a single reported case of the complete absence of Alpha class GSTs (Board *et al.* 1990a). Because of the complete absence of immunologically detectable GSTA1 or GSTA2 in this case, a simple point mutation was unlikely, and a gene deletion or major rearrangement was the more probable explanation. Very little is understood of the consequences of a major deficiency of the Alpha class GSTs, but because of its importance in the hepatic binding of bilirubin and other organic anions (Bhargava *et al.* 1978, Simons and Vander Jagt 1980), a deficiency may lead to such symptoms as

hyperbilirubinaemia. Interestingly, an individual with familial hyperbilirubinaemia (Rotor's Syndrome) exhibiting less than 0.2% of normal liver glutathione transferase activity has been reported (Adachi and Yamamoto 1987). This severe deficiency implies an absence of GSTA1 and 2 activity since GSTM1 is frequently deficient, and GSTP1 is not normally expressed in significant amounts in adult liver. The potential role of a GST deficiency in this rare disorder needs to be studied further, however, Southern blotting experiments have not indicated the presence of a major rearrangement in the Alpha class genes in rotor's syndrome patients (P.G. Board and Y. Adachi pers. comm.) suggesting Alpha class GST deficiency does not play a significant part in its aetiology and the presence of the deficiency in this patient may have been a coincidence. The rarity of reported individuals with Alpha class GST deficiency implies that the role of Alpha class GSTs may well be fundamental to survival.

1.7.2 Mu class

The human Mu class is the most numerous reported so far and there is an expanding array of distinct isoenzymes, many of which have been characterised (Warholm *et al.* 1983, Vander Jagt *et al.* 1985, Singh *et al.* 1987, Suzuki *et al.* 1987,1991, Ross and Board 1993). At the beginning of this study only one human Mu class cDNA had been cloned. Currently there are five characterised and cloned human Mu class isoenzymes: GSTM1, from liver (Board 1981a, Suzuki *et al.* 1987, DeJong *et al.*1988, Seidegård *et al.* 1988); GSTM2, from muscle (Laisney *et al.* 1984, Suzuki *et al.* 1987, Vorachek *et al.* 1991) ; GSTM3, from brain (Suzuki *et al.* 1987, Campbell *et al.* 1990); GSTM4, from testis, characterised in Chapters 3 and 4 (corresponding to a partial genomic clone known as GST*mu2*; Taylor *et al.* 1991) and GSTM5 from brain (Listowsky *et al.* 1993). Additional isoenzymes have been characterised, GST5 reported from brain tissue and GST6 found in a wide range of tissues (Suzuki *et al.* 1987, 1991).

GSTM1 is polymorphic with three common alleles; GSTM1a, GSTM1b and GSTM1-null (Board 1981a and b). The products of GSTM1a and 1b differ only in one amino acid in the coding region which leads to a replacement of Asn_{172} (Seidegård *et al.* 1988) with Lys_{172} (DeJong *et al.*1988) and results in a charge difference. It has been demonstrated that the null phenotype is the result of an absence of any immunologically detectable GSTM1 protein (Suzuki *et al.* 1987). Subsequently evidence has been presented of the lack of an 8kb *Eco*RI fragment containing the gene in GSTM1-null individuals (Seidegård *et al.* 1988, Board *et al.* 1990). The relationship between the Mu class GSTM1-null allele and susceptibility to carcinogens has been discussed in section 1.6.4.

The Mu class GSTM1-null allele is the most common deficiency documented in the GST gene family as approximately 40% of individuals have the null phenotype, demonstrating no detectable expression of GSTM1 isoenzyme. Indeed the null allele is the most common GSTM1 allele in all racial groups studied to date. A majority of population groups conform to this frequency for the null allele and it has reached fixation in some small pacific populations where population bottlenecks and founder affects are thought to have occurred (Table 1.3).

Table 1.3 Frequency of the Mu class GSTM1-null allele in different human populations.

population group	sample locality	sample number	gene frequency	authors
Indian	Malaysia	43	0.560	Board 1981a#
Japanese	Japan	168	0.691	Harada *et al.* 1987#
Chinese	Malaysia	96	0.765	Board 1981a#
European	Australia	40	0.816	Board 1981a#
European	England	49	0.640	Strange *et al.* 1984#
European	France	56	0.647	Laisney *et al.* 1984#
European	Sweden	248	0.725	Seidegård & Pero 1985+
Micronesian	Kiribati	37	1.00	Board *et al.* 1990*
Melanesian (Tolai)	Papua New Guinea	49	0.795	Board *et al.* 1990*
Polynesian	Cook Islands	49	0.904	Board *et al.* 1990*

#Determined by phenotyping of liver samples.
+Determined by measuring GSTM1 activity with *t* SBO in monocytes.
*Determined by restriction fragment deletion

1.7.3 Pi class

The Pi class appears to exhibit the least variation of any of the classes. There is only one well characterised enzyme,in humans ,GSTP1-1, and in rats, subunit 7-7, although sequence polymorphisms have been reported, in rat two cDNAs have been described (Sugowa *et al.* 1985, Pemble *et al.* 1986), these have only 2 nucleotides differing between them and this may represent a strain difference rather than the existence of two separate genes (Pemble *et al.* 1986). Two human Pi class GST cDNAs have also been characterised which differ by 2 amino acids ($I_{105}\rightarrow V, A_{114}\rightarrow V$; Kano *et al.* 1987, Board *et al.* 1989). Ahmad *et al.* (1990) reported a protein sequence that contained only one of the substituted residues, but a nucleotide sequence is not available.

Evidence has also been presented of human Pi class isoenzymes with minor differences in p*I*, in erythrocytes (Marcus *et al.* 1978, Suzuki *et al.* 1987), in muscle (Singh *et al.* 1988) and in heart and aorta (Tsuchida *et al.* 1990). It is possible that the variation may be due to some type of posttranslational modification, such as glycosylation (Kuzmich *et al.* 1991). The observation of two different sized subunits from human testis which cross-react with GSTP1-1 antiserum (Aceto *et al.* 1989) has been suggested as indicative of multiple isoenzymes. It seems unlikely however, that the variation is caused by multiple genes as hybridising Southern blots of *Hind*III digested human genomic DNA with a probe derived from GSTP1 indicates four hybridising fragments, three of which can be accounted for by either GSTP1 itself or a truncated processed pseudogene (Board *et al.* 1992). There is one weakly hybridising band that remains unaccounted for. This band does not hybridise to a 5′ Pi class cDNA probe and given that in this class both the sequence similarity between the rat and human coding regions and the intron/exon structure of the genes is well conserved, it is unlikely that this weak band is a closely related gene (Board *et al.* 1992). Bora *et al.* (1989a, b and 1991) characterised a human fatty acid ethyl ester synthase, reporting that it had more than 95% sequence similarity

with GSTP1 and also had GST activity. However, the reported sequence similarity is so high that this gene would be expected to hybridise with the GSTP1 probe almost as efficiently as GSP1 itself. Genomic Southern blot and PCR studies have not identified the fatty acid ethyl ester synthase gene and it may have been a cloning artifact by Bora *et al.*(1991) (Board *et al* 1989, 1993).

In addition to the apparent variation in the p*I* of the Pi class isoenzymes, Scott and Wright (1980) in a study of an Alaskan population, including 156 Eskimos, observed up to a six fold difference in the Pi class GST activity of erythrocytes from different individuals. Furthermore, family studies showed that although the activity levels of husbands and wives were not correlated with each other, it was possible to correlate those of the children with the mean levels of their parents. There is also some indication of a difference in the levels of activity between the Eskimo population and the general blood donor population. This suggests that there is a heritable variation in activity in the Eskimo population. A study of British blood donors did not identify similar variation in that population (Strange *et al.* 1982).

There is only a single reported example of a deficiency of GSTP1-1. In that case the erythrocyte isoenzyme activity was deficient, but unfortunately it was not determined whether expression in other tissues was also involved. Furthermore it was not possible to determine whether there was a genetic cause (Beutler *et al.* 1988).

It is clear that there exists considerable individual variation in isoenzyme activity within each of the major classes of GST, yet complete deficiency of any but the GSTM1-1 isoenzyme is vanishingly rare. Evaluation of the potential health risks associated with the variation of GST expression is not possible at present but given the evidence that isoenzymes have differing activities with specific substrates, it could be assumed that any reduction in the expression of an isoenzyme may be accompanied by a similar decrease in protection of the individual. Without deficient individuals to assess it is not surprising that

conclusive evidence has not as yet been provided of the range of functions of this gene family or its full significance.

1.7.4 Chromosomal location

Gene mapping experiments have assigned the human Pi class *GSTP1* gene to a locus at 11q13 (Suzuki and Board 1984, Moscow *et al.* 1988, Board *et al.* 1989) and have also revealed a second, weaker zone of hybridisation at 12q13-14 (Board *et al.* 1989). Further study has shown the chromosome 12 locus to be a truncated processed pseudogene (Board *et al.* 1992). *In situ* hybridisation with a GSTA1 cDNA probe has indicated that the human Alpha class genes are clustered at 6p12 (Board and Webb 1987). The human Theta class genes have yet to be characterised or mapped.

In comparison to the other GST classes, the chromosomal locations of the human Mu class GSTs have only recently been resolved ,having been the subject of some discussion in the literature. The Mu class constitutes a complex problem in chromosomal location because of the number of genes in the class; there are at least five closely related genes in this family which may cross-hybridise, and this is in addition to one established polymorphism, at the GSTM1 locus. The total number of cross-hybridizing genes in the human Mu class is not certain.

The GSTM1 locus has been reported on chromosome 1, using *in situ* hybridisation (DeJong *et al.* 1988), and on chromosome 13, using human/rodent somatic cell hybrids (DeJong *et al.* 1991). In this latter study, DeJong *et al.* (1991) also assigned additional Mu class loci to chromosomes 1 and 6 whereas Islam *et al.* (1989), again using human/rodent somatic cell hybrids, identified a Mu class locus on chromosome 3. The complexity of the chromosomal mapping of Mu class genes is discussed further in Chapter 5.

1.8 EVOLUTION

In the last few years the application of molecular biological techniques to elucidate the GST gene structure, regulation and function has begun to reveal the immense complexity of this supergene family. This burgeoning of GST amino acid and nucleotide sequence data has enabled intra- and inter-species comparisons amongst a wide range of GSTs from a variety of sources. Not only has there been a general increase in the sequences available but also the significant recognition of the new Theta class and characterisation of further isoenzymes belonging to this class. There are now some seventy two sequences from a wide range of species available for study.

However, there has been little study of the earliest origins of the GST gene family. The lack of evidence concerning GSTs themselves can be somewhat compensated by the parallel work on their primary substrate, GSH and on other GSH metabolism enzymes. Since initial observations of the extensive taxonomic distribution of GSH, considerable progress has been made in understanding both the metabolism of GSH and its occurrence in prokaryotes. Despite some continued controversies over relationships within the prokaryotes, a consensus is emerging of GSH metabolism arising in two specific groups of prokaryotes, leading to its incorporation into eukaryotes as an integral part of the endosymbioses that have led to mitochondria and chloroplasts. Different enzymes involved in GSH metabolism probably appeared at different stages in the evolutionary process (Fahey and Sunquist 1991). It has been further postulated that the inclusion of a GST in the enzyme complement of the symbiont may have increased its tolerance of the host environment (Pemble and Taylor 1992). In a study of *Entamoeba histolytica,* a eukaryote that contains neither organelle (Hasegawa and Hashimoto 1993), no evidence could be found for GSH or several of the main enzymes of GSH metabolism, including GSTs (Fahey *et al.* 1987).

Prokaryotes have been taxonomically divided into two kingdoms, the archeabacteria and the eubacteria (Fox *et al.* 1980, Woese 1987; for alternative views on the monophylogeny of the archaebacteria see Lake 1991 and Benachenhou-Lahfa 1993). GSH is found in two major groups of the eubacteria, the purple bacteria (in particular the alpha group) and the cyanobacteria. It is worthy of note that cyanobacteria are capable of oxygen photosynthesis (related to the chloroplast ancestor) and the alpha subgroup of the purple bacteria are considered to have given rise to eukaryotic mitochondria, the site of aerobic respiration. Thus an association seems to exist between oxygen-dependent metabolism and the production of GSH. GSTs, however, have not yet been detected in either group, this may well be due to both limited study and the need for alternative substrates.

None of the genes of GSH synthesis or metabolism have been found in the genomes of mitochondria (Anderson *et al.* 1981, Cantatore and Saccone 1987) although recently a rat isoenzyme, apparently related to the Theta class, has been purified and partially sequenced from the matrix of the mitochondria (Harris *et al.* 1991). There is no evidence that this is encoded in the genome of the organelle (Harris *et al.* 1991). But Klapheck *et al.* (1987) have localised part of a GSH synthetase to the chloroplast genome, thus implying GSH, at least, may be synthesised in the organelle itself. In plants, therefore, it may have been possible for the genes of GSH synthesis and metabolism to have been descended from the genomes of either chloroplasts or mitochondria or perhaps both.

Among the non-photosynthetic purple bacteria (*E. coli* and some *Pseudomonas*, *Salmonella* and *Proteus* species) low levels of GST activity with CDNB as substrate have been observed (Lau *et al.* 1980, Kerklaan *et al.* 1985, Di Ilio *et al.* 1988, Sundquist and Fahey 1989). The evidence, clearly incomplete at this point, implies that the first GST would seem to have been made by a non-photosynthetic purple bacteria. Reference has been made to similarities in structure between GSTs and glutaredoxin/thioltransferase,

present in non-photosynthetic purple bacteria (section 1.5.2 and below). This gene apparently represents a plausible precursor to GSTs in this group.

Interestingly, the GSH-dependent dichloromethane halogenase of *Methylobacterium* sp. is considered a GST enzyme (LaRoche and Leisinger 1990). This form is highly inducible but not active with CDNB as a substrate, two traits that have been found in some other GSTs. Although this enzyme exists as a hexamer, an unusual arrangement in GSTs, the subunits are 37.4 kDa, a similar size to the dimeric GST enzymes purified from the non-vertebrate eukaryotes, whereas mammalian GSTs are dimers of 23-28 kDa subunits.

Thus glutathione-dependent enzymes apparently evolved in aerobic organisms to facilitate inactivation of toxic products of oxygen metabolism (Fahey 1977, Mannervik 1986). Further selection may have been prompted through exposure to naturally occurring xenobiotics. Thus the diversity of detoxification enzymes seen in organisms today may well be a reflection of the diversity of naturally occurring xenobiotics. For example, the reason that predators of plants evolved such defences is to counter a diverse and ever-changing array of plant toxins in an changing environment (Rosenthal and Janzen 1979, Ames *et al.* 1990). That such defences are usually general, rather than specific for each compound makes good evolutionary sense and saves an organism from the need to expand their genome by wholesale multiplication of genes.

GST enzymes or GST activity have been found in a wide range of species, from bacteria, through representatives of the plant kingdom, fish and mammals (Mannervik and Dannielson 1988, Ketterer *et al* 1988, Clark *et al.* 1989, section 1.6.1, Chapter 6.1.3.1). Most surveys of non-mammalian species have used CDNB assays for GST detection, thus reports of isoenzyme variation and even comparative activity levels should be treated with some caution, since other substrates may be more relevant. The recent characterisation of the Theta class in mammals GSTs is an example.

At the initiation of this study the number of cDNA transcripts, genomic or amino acid sequences available for study only numbered approximately twenty, the vast majority of which were mammalian, either rat, mouse or human. However, comparison of these and the other characteristics of the isoenzymes led to several propositions on the evolutionary relationships of the GSTs. Firstly, the demonstration of conserved regions existing between the Alpha, Mu and Pi classes suggested the possibility of a shared, common ancestor (Mannervik 1985). Secondly, gene structures and sequences encoding enzymes from different species within a class were shown to have significantly greater similarities than those from different classes obtained from the one species (Mannervik and Danielson 1988). There is a clear implication then that the separation into these three distinct classes occurred before the separation of mammalian species.

Maize was the first plant to yield a GST sequence and both cDNA transcripts and genomic data are available (Shah *et al.* 1986, Moore *et al.* 1986 Grove *et al.* 1988). Sequence similarities between maize and mammalian GSTs have been considered minimal, but Mannervik and Danielson (1988) suggested that it was sufficient to assume a common evolutionary origin. The maize sequences, however, did not fit into any of the three established mammalian classes, implying that the Alpha, Mu and Pi classes may have emerged after the divergence of plants and animals. Comparison of the available amino acid sequences of the three GST classes in rats with the two known human Alpha class and two maize sequences suggested that a total of 43% of the amino acids were under evolutionary constraint (Rhoads *et al.* 1987). 12 of the amino acids were conserved over all the sequences, 24 additional residues belonged to the same side chain groupings and a further 59 residues were conserved in 3 out of the 4 categories. Considering the evolutionary distance between the species this level of conservation would seem significant (Rhoads *et al.* 1987).

Dipteran GSTs share some regions of similarity with the maize GST (Toung *et al.* 1990; Board *et al*, 1993b). Similarities of the Theta class sequences with both maize and the dipteran cDNAs have been suggested (Pemble and Taylor 1992). To further confuse the relationships, cDNAs from *Schistosoma japonica* and *Fasciola hepatica*, two parasitic trematodes, clearly belong to the Mu class (Smith *et al.* 1986, Paccicio, unpublished GENBANK accession number m943434).

Limited but detectable sequence similarities between regions of thioltransferase or glutaredoxin and Mu class GSTs have previously been noted. Mannervik and Danielson (1988) have suggested a distant evolutionary link amongst these proteins. Furthermore, studies of higher order structure such as polypeptide chain folding, have led to possible ancestral relationships between glutaredoxin, thioredoxin, and selenium-dependent glutathione peroxidase being postulated. Mannervik *et al.* (1990) have suggested that perhaps these glutathione dependant enzymes share a common ancestor with the GSTs. Rather than a common origin, this similarity may be a reflection of the traits they hold in common, such as a glutathione binding site, that each appears to show glutathione-induced binding site conformational changes, and that each has a broader than is usual substrate specificity, including some of the products of oxidative metabolism. Sinning *et al* (1993) has suggested that the structural fit of domain 1 of mammalian GSTs with T4 glutaredoxin in particular, but also thioredoxin and glutathione peroxidase, warrants describing this as a canonical GSH binding domain.

1.9 AIMS OF THIS STUDY

The importance of the Phase II enzymes, and the GSTs in particular, in protecting the cell from a combination of xenobiotic and endogenous challenges is now well documented. A combination of the relatively long evolutionary history, the ubiquitous nature, the multiplicity of GSTs in most species, in conjunction with the apparent constancy of function and

complexity of the expression pattern in the GST gene family is indicative of a significant metabolic role, that is probably critical to survival.

This study was undertaken to extend our knowledge of (1) the molecular genetics of the human Mu class GSTs and (2) the evolutionary relationships of the GST superfamily of genes. At the initiation of these studies only one human Mu class cDNA was cloned, though four Mu class isoenzymes had been partially characterised. Whilst these previous studies had revealed a complex family of related enzymes there was a clear need for further characterisation and an understanding of their genetic relationships. In confirmation of this need, since the study commenced even the number of the classes of GSTs has increased, with the discovery of the Theta class being prompted by the identification of previously unrecognised enzymes. There has also been a rapid increase in the total number of genes and pseudogenes identified and cloned within each class. Knowledge of the regulation of these genes is still under intense examination and the evolution of this supergene family has not as yet been studied in any detail.

CHAPTER 2 MATERIALS AND METHODS FOR RECOMBINANT DNA TECHNIQUES

The materials and methods described below are those general techniques used throughout the research reported here. Techniques specific to only one aspect of this research are described in the relevant chapters.

2.1 REAGENTS AND MATERIALS

Enzymes, the buffers commercially supplied for them and the sources for both are given in Table 2.1. Reagents required in the experiments described here, along with the most commonly used suppliers, are listed in Table 2.2. All reagents were of analytical or A grade quality.

2.1.1 Bacterial strains, bacteriophage and plasmids

E. coli strain Y1090 (genotype - *sup F, hsdR, araD*139, Δ*lon*,.Δ*lac*, U169, *rspL*,.*trp*C22::Tn*10(tet^r)* pMC9: Young and Davis 1983) was used as the host for replication of λgt11 bacteriophage. Cultures were grown on either NZCYM (Sambrook *et al.* 1989) or low salt LB broth, a modification of L-broth (Lennox 1955) with only 1g NaCl per litre. Selection within stock cultures of Y1090 was maintained by the addition of ampicillin (100mg/ml).

E.coli strain TG1 (genotype—*K12, D (lac-pro), supE, thi, hsdD5/F'tra36, proA$^+$B$^+$, lacI^Q, lacZDM15*: Amersham, Melbourne, Australia) was used as a host for M13 mp18 and 19 (Messing and Viera, 1982) and for pUC18 (Norrander *et al.* 1983). L-broth or 2YT broth (Sambrook *et al.* 1989) were used for cultures of TG1 harbouring M13 or pUC vectors. TG1 stock cultures were maintained under selection on minimal media (Miller 1972). Media plates were 1·2% (w/v) agar or agarose and top agar/agarose contained 0·6% (w/v) agar/agarose.

Table 2.1 The sources for enzymes and their buffers.

enzyme	source
Alkaline Phosphatase	Boehringer-Mannheim
Deoxyribonulease1 grade II (DNase)	Boehringer-Mannheim
T4 DNA Ligase	Bresatec or Pharmacia
DNA Polymerase 1 (Klenow fragment)	N E B
Polynucleotide Kinase	Pharmacia
Restriction Endonucleases and 10X buffer	Boehringer-Mannheim, Progen, Pharmacia or Promega
Ribonuclease A type 1-As (RNase)	Sigma
Taq (*Thermus aquaticus*) DNA Polymerase and 10x Buffer	Promega

Pharmacia Aust. Sydney, Australia: Progen industries QLD., Australia: Promega Madison WI. USA: N E B: New England Biolabs. Beverly MA. USA.

2.2 STANDARD LABORATORY PROCEDURES

Sterilisation, if required, was performed by autoclaving at 121°C/100kpa for 15 min, or by filtration through 0·2µm membranes. Distilled deionised water was used for all solutions and disposable plastic labware was autoclaved before use.

All recombinant DNA procedures were performed using C1 biological containment conditions as stipulated by the Australian Recombinant Monitoring Committee and radioactive substances were handled in accordance with the Australian National University Radiation Handbook.

2.3 METHODS

2.3.1 Routine preparative methods

Nucleic acid samples were purified by phenol/chloroform extraction and subsequent ethanol precipitation in standard procedures (Sambrook *et al*. 1989). If volume was limiting in the precipitation step then ethanol was replaced by 0·6 volumes of isopropanol (Sambrook *et al*. 1989).

Alternatively, Geneclean II was used, which involves the binding of nucleic acids to ground glass powder in the presence of high concentrations of NaI, removal of impurities by successive ethanol washes and finally elution in water.

Table 2.2 Commonly Used Reagents and their suppliers.

Reagent	Source
Acrylamide/Bisacrylamide	Bio-Rad Lab.
Agar	Difco Lab.
Agarose, Type II	Sigma
Ampicillin	Sigma
ATP	Sigma
Black and White film, T667	Polaroid Corp.
b-mercaptoethanol	BDH
2',3' Deoxynucleotide Triphosphates (dATP, dCTP, dGTP, dTTP)	Boehringer-Mannheim
2',3' Dideoxynucleotide Triphosphates (ddATP, ddCTP, ddGTP, ddTTP)	Boehringer-Mannheim
DTT	Sigma
Ethidium Bromide	Sigma
Geneclean II	Bio101
Genescreen Plus (nylon) membrane	DuPont
IPTG	Sigma
λ DNA cut with HindIII	Bresatec
M13mp18 and 19	Boehringer-Mannheim
Nitrocellulose membranes (Transblot or Hybond-C)	Bio-Rad or Amersham
Radio chemicals:	
(α-^{32}P)dATP (3000Ci/mmol)	Amersham
(γ-^{32}P)dATP (>5000Ci/mmol)	Bresatec
(^{35}S)dATP (1000Ci/mmol)	Amersham
Sequencing system-T7 DNA Polymerase	Amersham
TEMED	Sigma
Tryptone	Difco Lab
Urea	Merck
X-gal	Boehringer-Mannheim
X-ray film Rx or NC II	Fuji
Yeast Extract	Difco-BRL

Amersham, Sydney, Australia: Bio-Rad Laboratories, Richmond CA., USA: BDH Chemicals Aust., Kilsyth, Victoria, Australia: Boehringer-Mannheim Aust., Sydney, Australia: Bio101, La Jolla CA., USA: Bresatec, Adelaide, Australia: Difco Laboratories, Detroit MI., USA.: DuPont Aust., Sydney, Australia: Fuji Photo Film Co. Japan: Merck, Damsadt, West Germany: Polaroid Corporation, Cambridge MA., USA: Sigma, St. Louis MO., USA.

Oligonucleotide primers were synthesised by the A.N.U. Biomolecular Resource Facility on a 3-column Applied Biosystems 380B Oligonucleotide Synthesizer (Applied Biosystems, Foster city CA., USA). The

primers were supplied in ammonia solution and dryed in a vacuum centrifuge, resuspended in 200µl H_2O, re-dried and finally resuspended in 200µl H_2O. Primer solutions were then quantified, stored at -20°C and diluted as required.

When required, nucleic acid concentration was estimated by measuring the spectrophotometric absorbance of solutions at 260nm (Sambrook *et al.* 1989) on either a Gilford 2600 or a Cary 1 UV-visible spectrophotometer.

2.3.2 DNA preparation

2.3.2.1 Preparation of Lambda DNA

Lambda bacteriophage DNA was prepared from either small scale 50ml cultures or larger scale 500ml cultures grown by standard procedures (Sambrook *et al.* 1989). Plaques were picked into 1ml of SM solution (Sambrook *et al.* 1989) containing 10µl of chloroform and allowed to elute overnight. An aliquot of the phage eluant, diluted to ~$1{\cdot}6X10^8$ pfu in SM, was pre-incubated with 0·1ml exponential *E. coli* Y1090 host cells (~$8X10^9$ cells) in 1ml of media containing 10mM $MgSO_4$ for 30 minutes at 37°C. This was used to inoculate 50ml of media, also containing 10mM $MgSO_4$, before shaking, at 37°C overnight. Phage DNA was purified by polyethylene glycol precipitation, essentially by the method of Yamamoto *et al.* (1970) with modifications as described in the manual "Promega Protocols and Applications Guide".

2.3.2.2 Preparation of Plasmid DNA

Small scale preparations of plasmid DNA were prepared from 1ml of an L-broth/ampicillin culture after overnight shaking incubation at 37°C, without amplification. The mini-prep boiling method of Holmes and Quigley (1981) was used for the preparation of plasmid DNA for all purposes.

2.3.2.3 Preparation of M13 bacteriophage DNA

Double stranded replicative form (RF) DNA of M13 was prepared by the mini-prep boiling method of Holmes and Quigley (1981) in the same way as plasmid DNA, often from the cell pellet left from a single stranded DNA (ssDNA) preparation. ssDNA derived from M13 DNA was prepared using a polyethylene glycol precipitation method described by Sanger *et al.* (1980).

2.3.3 Gel electrophoresis

2.3.3.1 Agarose gels

Rapid analyses of DNA samples at all stages of this research were conducted in mini gels (60X60X5mm) of 0·8-1·2% (w/v) agarose in TBE buffer (90mM Tris-HCl pH 8·2/90mM borate/1mM EDTA) at 16·6V/cm. Both the gel and TBE running buffer contained 0·5µg/ml ethidium bromide fluorescent intercalating dye. DNA samples were then visualised with U.V. irradiation and photographed with a Wratten filter using a Polaroid MP4 Land camera (Polaroid Corp. Cambridge, MA. USA).

2.3.3.2 Acrylamide gels

Denaturing gels for sequencing were 6% polyacrylamide, 0·75% Ammonium persulphate, 0·03% TEMED, 8M Urea and were made up in TBE buffer. A 40% stock solution of acrylamide/N,N' methylene-bis-acrylamide (19:1 w/v) was prepared by deionisation with mixed bed resin and stored at 4°C. Ammonium persulphate at 10% (w/v) was made periodically and frozen in 1ml aliquots. Gels were pre-run for ~20 minutes prior to sample loading and were electrophoresed at 30-33mA. To allow uniform migration of samples across the gel an aluminium heat dispersal plate was attached to one glass plate. A "sharks tooth" comb was used to form the loading wells. Sequencing gels were fixed in 10% methanol; 10% acetic acid solution (v/v) for 10-15 minutes, transferred to paper (Whatman 3MM) and dried on the paper at 80°C under vacuum. DNA detection was carried out by autoradiography and was performed at room temperature for 16 to 36 hours.

2.3.4 Restriction endonuclease digest of DNA

Restriction endonuclease (**RE**) digestions were performed under similar conditions to those described by Farrel *et al.* (1981) using low medium or high salt buffers. Alternatively, DNA (approximately 1μg) was digested with 10-20 units of the appropriate restriction endonuclease using the buffers supplied by the manufacturer, in a total volume of 20μl. Multiple digestions were generally performed simultaneously in a compatible buffer. The reactions were terminated by the addition of gel loading buffer (0·25% bromophenol blue, 30% glycerol, 1mM EDTA) for electrophoresis. When the DNA was to be used in further reactions termination was by heating to 65°C for 20 minutes or by phenol chloroform extraction. Bacteriophage DNA was heated to 65°C for 5 minutes after the addition of loading buffer and prior to electrophoresis to dissociate annealed compatible ends.

2.3.5 Subcloning

The pUC18 plasmid and M13 bacteriophage series were the vectors of choice for subcloning, general DNA manipulation and DNA sequencing because of the ease of detecting recombinants and versatility of the polylinker cloning sites. Preparation of vector, which involved RE digestion and removal of 5′ phosphates with alkaline phosphatase to prevent recircularisation (when using only one RE for digestion), was typically carried out using 1μg of vector DNA as described by Ausubel *et al.* (1987). This was followed by phenol chloroform extraction and ethanol precipitation. Prepared vectors were resuspended in TE buffer (10mM Tris-HCl, 1mM EDTA, pH8) and stored at -20°C.

Insert fragments for subcloning were generally cleaved with two REs so that orientation of the insert could be directed. After digestion target fragments were electrophoretically separated and isolated from agarose gel using Geneclean II.

Ligations were carried out in 50mM Tris-HCl, pH7·6, 10mM $MgCl_2$, 5% (w/v) PEG_{6000}, 1mM DTT, 1mM ATP by combining 40ng of prepared vector with approximately 80ng of the insert DNA and 0·3-1 unit of T4 DNA ligase. Cohesive ends require the lesser amount of enzyme due to higher efficiency, whereas blunt-ended ligations are known to be less efficient.

2.3.6 Preparation and transformation of competent cells

Competent cells were prepared by the method of Dagert and Ehrlich (1977) incorporating modifications described in the "Manual for the M13 Cloning/Sequencing System" (Pharmacia, Uppsala, Sweden). This involved pelleting of an exponentially growing culture then resuspension, first in 1/5th of the growth volume with 10mM NaOAc, pH 5·6, 50mM $MnCl_2$, 5mM NaCl, followed by a second resuspension in 1/50th of the growth volume with 10mM NaOAc pH5·6, 70mM $CaCl_2$, 5mM $MnCl_2$, with 5% (w/v) glycerol. The competent cells were then snap frozen on dry ice and stored at –70°C. Cells prepared in this way were transformed by either plasmid or M13 RF DNA as described by Yanisch-Perron *et al.* (1985) except that either L-broth or 2YT was used. In general, half of the ligation mixture described above was used in a transformation.

2.3.7 Identification of recombinants

M13 and pUC plasmids have been constructed to take advantage of α complementation which provides a colour indicator simplifying identification of recombinants (Ullman *et al.* 1967, Gronenborn and Messing 1978). An appropriate host (*E. coli* TG1) carries on its F′ episome a defective β-galactosidase gene whereas the M13 and pUC plasmids contain the missing regulatory region and the coding region for the first 146 amino acids of the β-galactosidase gene *(lacZ)*. When grown in the presence of IPTG and X-GAL, transformed host cells containing non-recombinant vectors produce a *lac*$^+$ phenotype identified by blue

colonies/plaques, due to complementation of the host's β-galactosidase gene deficiency. Insertion of recombinant DNA in the *lac* gene of the vector causes inactivation, seen as the clear or white plaques/colonies of the host.

Recombinant M13 phage ssDNA was examined for inserts by comparison with non-recombinant ssDNA using direct gel electrophoresis (DIGE) of the phage culture supernatant (Messing 1983). When required, clones with opposing insert orientations were detected by the "C-test" (Messing 1983).

2.3.8 Preparation of radioactively labelled probes

Oligonucleotides were 5'-end-labelled by a polynucleotide kinase forward reaction as described by Richardson (1965) employing (γ-^{32}P)dATP.

2.3.9 DNA sequence analysis

DNA sequencing was carried out by the dideoxy chain termination method of Sanger *et al.* (1977) in M13 phage mp18 or mp19 cloning vectors, following the modifications in the technique suggested by Messing (1983). A T7 DNA polymerase sequencing kit (Amersham) was used for the majority of the sequencing reported here. Reaction conditions were those recommended by the kit manufacturer, except that the labelling reaction was generally carried out over 3 minutes instead of the recommended 5 minutes. Reaction times and DNA concentrations were varied according to the distance from the primer required to be sequenced. Sequencing reactions were resolved on 6% polyacrylamide gels according to the methods described in section 2.3.3.2 and following Sambrook *et al.* (1989). All sequences were obtained on both strands and all restriction sites used in subcloning were overlapped.

Autoradiographs of gels were read manually for the most part, however some were entered directly into the Staden (1984) programmes in the Amersham/Staden package on an ASI "AT" personal computer using a

Sonic Digitizer (Graf/Bar CD-7, Science Accessories Corporation). The sequences were then transferred to the PCGENE package (Intelligenetics, Mountain View, CA. USA) where they were stored and analysed. The majority of sequences, however, were entered directly into PCGENE.

CHAPTER 3 CLONING OF AN ALTERNATIVELY SPLICED MU CLASS GLUTATHIONE *S*-TRANSFERASE TRANSCRIPT

3.1 INTRODUCTION

Consideration of the currently available human cDNA transcripts (at least seven human subunits) together with evidence from Southern blots of genomic digests (Rothkopf *et al.* 1986, Board and Webb 1987, DeJong *et al.* 1988a, Seidegård *et al.* 1988) indicates that our knowledge of the considerable heterogeneity amongst this supergene family is still incomplete.

The human Mu class is currently the most extensively reported. A number of distinct isoenzymes have been characterised or sequenced (Board *et al.* 1990a, Chapter 1.7). When this study commenced the cDNA encoding only one of the human Mu class isoenzymes had been cloned. Currently there are five characterised and cloned human Mu class isoenzymes: GSTM1 (Board 1981a, Warholm *et al.* 1983, Vander Jagt *et al.* 1985, Soma *et al.* 1986, Suzuki *et al.* 1987, DeJong *et al.*1988a, Seidegård *et al.* 1988); GSTM2 (Laisney *et al.* 1984, Suzuki *et al.* 1987, Board *et al.* 1988, Vorachek *et al.* 1991) GSTM3 (Campbell *et al.* 1990), GSTM4, characterised here and GSTM5 (Listowsky *et al.* 1993). Several additional isoenzymes have been characterised, including GST5 and GST6 (Suzuki *et al.* 1987, 1991) and it is not yet clear if these correspond to the GSTM3, M4 or M5 cDNAs. In addition, Tsuchida *et al.* (1990) purified five Mu class subunits from human heart and aorta, two of these appear to be previously undescribed variants and it is possible that they may be products of distinct genes.

GSTM1, the most extensively studied Mu class isoenzyme, is polymorphic with three alleles; GSTM1a, GSTM1b and GSTM1-null (Board 1981a and b). It has been shown that the null phenotype is the result of the absence of the *GSTM1* gene (Seidegård*et al.* 1988). The GSTM1a and GSTM1b alleles differ

For example, the human testes library has $1.0X10^6$ independent clones with an average size of 1.7kb (0.7 to 3.3kb).

by a single nucleotide that results in an Asn_{172} to Lys_{172} substitution, causing a charge difference (Seidegård *et al*. 1988, DeJong *et al*.1988a).

The varying patterns of GST expression described in Chapter 1.6.2 suggest that their regulatory mechanisms may be complex. Although the mechanisms described so far are largely pre-transcriptional (Daniel *et al*. 1989, Okuda *et al*. 1989, Rushmore and Pickett 1990 and Chapter 1.5) there is some evidence that expression of the Pi class enzyme GSTP1-1 may be regulated by a post-transcriptional mechanism (Morrow *et al*. 1992) and post-translational modification of several rat GSTs has been demonstrated *in vivo* (Chapter 1.5).

3.2 AIM

Evidence has accumulated from several studies in humans (Suzuki *et al*. 1987, 1991, Tsuchida *et al*. 1990, Taylor *et al*. 1990, 1991) that there are further Mu class isoenzymes to be genetically described. In an attempt to gain a greater understanding of the extent of genetic diversity of the human Mu class GSTs an investigation into the Mu class GSTs expressed in several human tissues was undertaken.

3.3 MATERIALS

The human tissue specific cDNA libraries used were supplied by Clontech (Palo Alto, CA, USA). The libraries were constructed using λgt11 as a vector with cDNA transcript fragments inserted at the *Eco* R1 endonuclease restriction site. Antibody raised against both denatured GSTM1 and GSTM2 was the gift of Dr. Philip Board (Molecular Genetics Group, JSCMR, A.N.U.) and goat anti-rabbit second antibody was supplied by Sigma. Circular nitrocellulose filters were supplied Schleicher and Schuell (Dassel, Germany).

3.4 METHODS

3.4.1 Immuno screening of human cDNA expression libraries

Three human cDNA libraries derived from brain, foetal heart and testis tissues were screened using antibody raised against both GSTM1 and GSTM2. Cross-reacting antigen was detected by a rabbit primary antiserum and a goat anti-rabbit IgG second antibody coupled to alkaline phosphatase, essentially as previously described by Board *et al.* (1988).

E. coli (Y1090) infected with each cDNA library were plated at a density of 50,000 plaques per 150mm NZCYM (Chapter 2.1.2) plate, using procedures described by Sambrook *et al.* (1989). After 4 hours of incubation at 42°C plates were over-lain with nitrocellulose filters previously soaked in 10mM IPTG and the incubation continued for a further 4 hours. The filters were then removed from the plates and additional protein binding sites were blocked by soaking in 50mM NaCl, 50mM Tris, pH7·5, 5% (w/v) milk powder. The filters were subsequently incubated with the primary antibody in the. blocking buffer.

Areas of the master plates corresponding to positive signals were picked into 500µl of SM buffer (Sambrook *et al.* 1989) with 10µl $CHCl_3$ for overnight elution. The initial positives were then replated at a lower density (~200pfu/90mm plate) to allow individual plaque identification. These plates were then rescreened in the same manner as the master plates and single positive plaques were picked into 500lµl of SM buffer as described.

3.4.2 Radioactive probe screening of human cDNA libraries

The human cDNA libraries derived from brain, foetal heart and testis tissues were also screened using a synthetic oligonucleotide probe. Libraries were plated out at the same density, using the procedures as described for antibody screening but without IPTG filters. The plates were incubated for 16 hours and the resulting plaque lifts, on nitrocellulose filters, were prepared by

denaturation and bonding of the adsorbed phage particles according to the method of Benton and Davis (1977). The detection of positive clones was carried out by hybridisation to a radioactively labelled probe (Chapter 2.3.8) (GT1 2B:5'AGGAGACTGCTATCATGCCCATGATACTG3') based on the known sequence of GSTM1b (Seidegård *et al.* 1988). Hybridisation was carried out at 65°C for 16 hours in a 5X Denhardts (0·02% Ficoll, 0·02% polyvinylpyrrolidone, 0·02% bovine serum albumen) solution and filters were washed in 2X SSC (0·15M NaCl, 0·015 M Na.citrate, pH 7·6) at room temperature and then, with several changes, in 2X SSC containing 0·1%(w/v) SDS at 60°C, as described by Sambrook *et al.* (1989).

Areas of the master plates corresponding to positive hybridisation signals were replated at the lower density as described in section 3.4.1 and rescreened by successive hybridisations with the original oligonucleotide probe. Those hybridising plaques that were clearly separated were used to prepare liquid lysate stocks (Chapter 2.3.2.1).

3.5 RESULTS

3.5.1 Characterisation of the positive clones from cDNA library screens

Of the 5×10^5 plaques screened from each of the heart, brain and testis λgt11 cDNA libraries only five plaques, all from the testis library, showed initial positive signals with Mu class GST antibody. Four of the five remained positive on immuno-rescreening but did not hybridise to a radioactively labelled oligonucleotide (GT1 2B) based on the human GSTM1 sequence. Due to the failure of the immuno-screened positive plaques to hybridise to the radioactively labelled probe it was decided the libraries should be rescreen using radioactively labelled probes.

Initial screening of the heart, testis and brain cDNA libraries by radioactively labelled oligonucleotides derived from the sequence of GSTM1 provided one heart, eighteen brain and twenty four testis positive

hybridisation signals. The single positive derived from the heart library failed to rescreen. Eight of the original positives from the testis library produced strong signals, while the remainder showed weaker signals. Five of the initial brain positives rescreened but with signals somewhat weaker than the testis positives, the remaining rescreens produced signals that were weaker again.

Eight positive testis clones were purified further, however, the inserts from two of these were judged too small to be full length transcripts. Sequence analysis showed that two colonies had no similarity to known GST sequences despite the hybridisation signals, the four remaining clones were identified with similarities to GST Mu class sequences.

Although twelve positives from the brain library screen were selected for further study, the majority failed to release inserts by digestion with *Eco*RI, despite numerous attempts at both digestion and DNA purification, using several different protocols. Such difficulty has also been reported by Hayes *et al.* (1992). Of the positive clones that yielded inserts, one was judged to be the wrong size and four remain to be fully characterised, but when hybridised to the partial GSTM4 clone, pGST-T, the brain library clones showed positive signals significantly weaker than those from the remaining testis clones.

3.5.2 Characterisation of the new GST cDNA clones

The sequencing of cross hybridising cDNA clones from the testis library has identified four clones with homology to GST Mu class sequences. Two clones encoded GSTM2 (formerly GST4) which has been described by Board *et al.* 1988 and cloned while this study was in progress, by Vorachek *et al.* (1991). In comparison with the published nucleotide sequence (Vorachek *et al.* 1991) the GSTM2 sequence had a single T to C substitution at base 222 (numbered from the initiating ATG) in the coding region and, in the 3′ non-coding region, a single base deletion and a single base insertion (Figure 3.1). These may be the result of natural polymorphisms, the coding region

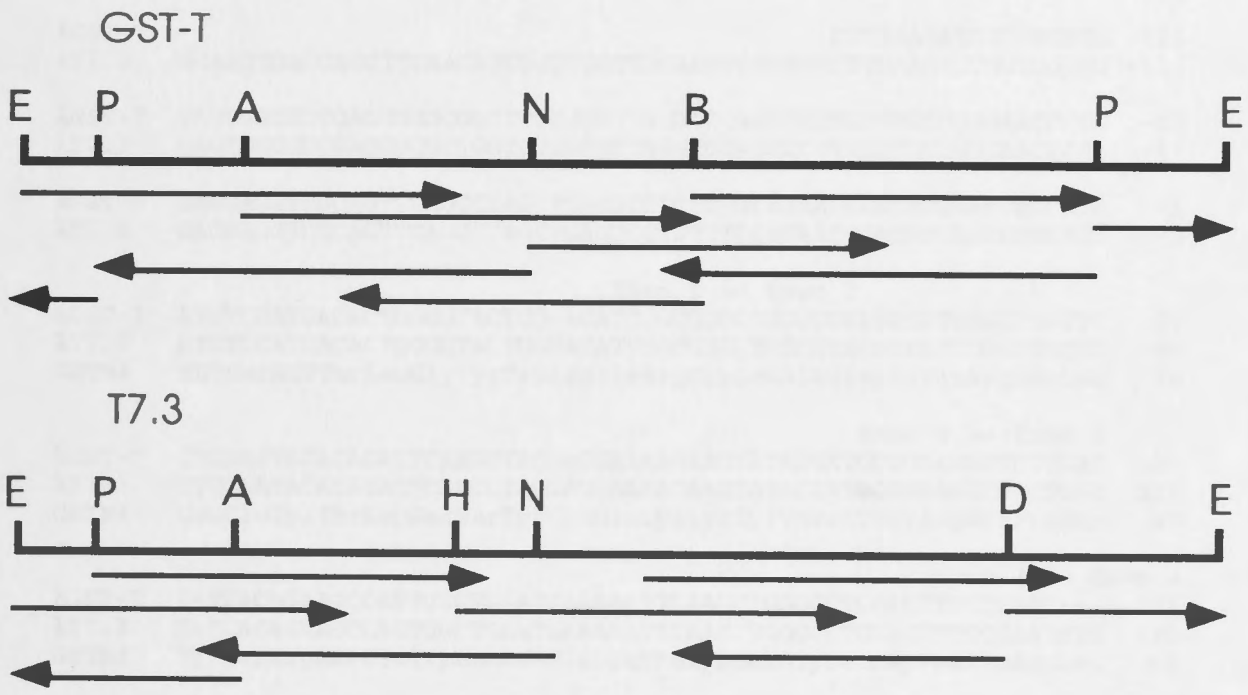
substitution is silent and does not alter the amino acid sequence, while the 3′ non-coding insertion and deletion are within 15 bases of each other and thus minimise any possible differences (Figure 3.1).

Figure 3.1 Nucleotide sequence of GSTM2. Nucleotides that differ from the published sequence are underlined in **bold**. There are 46 nucleotides missing at the 5′ end.

```
CGCGGGCTGG CCCATTCCAT CCGCCTGCTC CTGGAATACA CAGACTCAAG   50
CTACGAGGAA AAGAAGTACA CGATGGGGGA CGCTCCTGAT TATGACAGAA  100
GCCAGTGGCT GAATGAAAAA TTCAAGCTGG GCCTGGACTT TCCCAATCTG  150
CCCTACTTGA TTGATGGGAC TCACAAGATC ACCCAGAGCA ACGCCATCCT  200
GCGGTACATT GCCCGCAAGC ACAACCTGTG CGGGGAATCA GAAAAGGAGC  250
AGATTCGCGA AGACATTTTG GAGAACCAGT TTATGGACAG CCGTATGCAG  300
CTGGCCAAAC TCTGCTATGA CCCAGATTTT GAGAAACTGA AACCAGAATA  350
CCTGCAGGCA CTCCCTGAAA TGCTGAAGCT CTACTCACAG TTTCTGGGGA  400
AGCAGCCATG GTTTCTTGGG GACAAGATCA CCTTTGTGGA TTTCATCGCT  450
TATGATGTCC TTGAGAGAAA CCAAGTATTT GAGCCCAGCT GCCTGGATGC  500
CTTCCCAAAC CTGAAGGACT TCATCTCCCG ATTTGAGGGC TTGGAGAAGA  550
TCTCTGCCTA CATGAAGTCC AGCCGCTTCC TCCCAAGACC TGTGTTCACA  600
AAGATGGCTG TCTGGGGCAA CAAGTAGGGC CTTGAAGGCA GGAGGTGGGA  650
GTGAGGAGCC CATACTCAGC CTGCTGCCCA GGCTGTGCAG CGCAGCTGGA  700
CTCTGCATCC CAGCACCTGC CTCCTCGTTC CTTTCTCCTG TTTATTCCCA  750
TCTTTACTCC CAAGACTTCA TTGTCCCTCT TCACTCCCCC TAAACCCCTG  800
TCCCATGCAG GCCCTTTGAA GCCTCAGCTA CCCACTATCC TTCGTGAACA  850
TCCCCTCCCA TCATTACCCT TCCCTGCACT AAAGCCAGCC TGACCTTCCT  900
TCCTGTTAGT GGTTGTGTCT GCTTTAAAGG -CCTGCCTGG CCCCTCGCCT  950
GTGGAGCTCA GCCCCGAGCT GTCCCCGTGT TGCATGAAGG AGCAGCATTG 1000
ACTGGTTTAC AGGCCCTGCT CCTGCAGCAT GGTCCCTGCC TTAGGCCTAC 1050
CTGATGGAAG TAAAGCCTCA ACCAC                            1074
```

Two other clones were clearly related and encode transcripts of a new GST gene. The sequencing strategy for both clones is shown in Figure 3.2. The first clone sequenced (designated λGST-T) is 1158 nucleotides long with a 95 nucleotide poly-A tail extension (Figure 3.3.) This clone contains a 5′ prime non-coding region of 147 nucleotides and a 3′ prime non-coding region of 462 nucleotides. When the sequence was aligned with other known Mu class genes it was found to have an 82 nucleotide deletion which corresponds to exon 4. The deletion is excised exactly at the splice junctions expected from comparison with rat and human Mu class genes (Lai *et al.* 1988, Taylor *et al.* 1991). The deletion of exon 4 causes a shift in reading frame resulting in 36 missense amino acids before a stop codon.

Figure 3.2 A schematic drawing showing restriction sites and indicating the sequencing strategy of the GSTM4 cDNA clones, (a). λgst-t and (b). λGST-T7.3. E=*Eco*RI, P=*Pst*I, N=*Ncd*, H=*Hph*I, D=*Dra*I.



The other clone (λT7·3) isolated from the same testis library is 1323 nucleotides long and also finishes with a poly-A Tail (Figure 3.3). It has a 5′ non-coding region of 270 nucleotides and was found to contain the exon 4 sequence deleted from λGST-T. However, the homology of the coding region stops at the 3′ boundary of exon 7 and continues with unrecognisable sequence (Figure 3.3). This sequence may be intronic in origin. However, now that the gene sequence has been published (Comstock *et al.* 1993) the evidence suggests that if this is the case, then the unrecognised sequence does not come from the intron directly following exon 7. The junction between the coding and non-recognisable regions is again precisely at the predicted splice site, and has resulted in the omission of exon 8, which, based on λGST-T, contains sequence for 23 amino acids of the carboxy terminal region as well as the 3′ untranslated region before the poly-A tail. The presence of possible intronic sequence results in the substitution of 6 alternative amino acids before an inframe stop codon and a putative 463 3′ untranslated region including a clear poly-adenylation signal (AATAAA) and a poly-A tail. This alternative 3′ non-coding region is within 1 bp of the length of the 3′ noncoding region of GSTM4, but has no similarity to a Mu class sequence.

Figure 3.3 Alignment of the nucleotide sequences of two cDNA clones encoding the GSTM4 subunit. The predicted splice sites are shown above the sequence. The six substituted residues at the new truncated C-terminal of the protein encoded by λT7·3 are shown below the GSTM4 sequence at the start of exon 8.

```
λT7.3                                          GGGGCTGAACACTCGGAGGTGGCGGTGGATCT -241

λT7.3    TACTCCTTCCAGCCAGTGAGGATCCAGCAACCTGCTCCGTGCCTCCCGCGCCTGTTGGTT -181

λGST-T                                                  CTTGAAGATCGGCGGGCG -121
λT7.3    GGAAGTGACGACCTTGAAGATCGGCCGGTTGGAAGTGACGACCTTGAAGATCGGCGGGCG -121

λGST-T   CAGCGGGGCCGAGGGGGCGGGTCTGGCGCTAGGTCCAGCCCCTGCGTGCCGGGAACCCCA  -61
λT7.3    CAGCGGGGCCGAGGGGGCGGGTCTGGCGCTAGGTCCAGCCCCTGCGTGCCGGGAACCCCA  -61

λGST-T   GAGGAGGTCGCAGTTCAGCCCAGCTGAGGCCTGTCTGCAGAATCGACACCAACCAGCATC   -1
λT7.3    GAGGAGGTCGCAGTTCAGCCCAGCTGAGGCCTGTCTGCAGAATCGACACCAACCAGCATC   -1

                                        Exon 1 >< Exon 2
λGST-T   ATGTCCATGACACTGGGGTACTGGGACATCCGCGGGCTGGCCCACGCCATCCGCCTGCTC   60
λT7.3    ATGTCCATGACACTGGGGTACTGGGACATCCGCGGGCTGGCCCACGCCATCCGCCTGCTC   60
GSTM4    METSerMETThrLeuGlyTyrTrpAspIleArgGlyLeuAlaHisAlaIleArgLeuLeu   20

                                                      Exon 2 >< Exon 3
λGST-T   CTGGAATACACAGACTCAAGCTACGAGGAAAAGAAGTATACGATGGGGGACGCTCCTGAC  120
λT7.3    CTGGAATACACAGACTCAAGCTACGAGGAAAAGAAGTATACGATGGGGGACGCTCCTGAC  120
GSTM4    LeuGluTyrThrAspSerSerTyrGluGluLysLysTyrThrMETGlyAspAlaProAsp   40

                                                           Exon 3 >< Exon 4
λGST-T   TATGACAGAAGCCAGTGGCTGAATGAAAAATTCAAGCTGGGCCTGGACTTTCCCAAT---  177
λT7.3    TATGACAGAAGCCAGTGGCTGAATGAAAAATTCAAGCTGGGCCTGGACTTTCCCAATCTG  180
GSTM4    TyrAspArgSerGlnTrpLeuAsnGluLysPheLysLeuGlyLeuAspPheProAsnLeu   60

λGST-T   ------------------------------------------------------------
λT7.3    CCCTACTTGATTGATGGGGCTCACAAGATCACCCAGAGCAACGCCATCCTGTGCTACATT  240
GSTM4    ProTyrLeuIleAspGlyAlaHisLysIleThrGlnSerAsnAlaIleLeuCysTyrIle   80

                    Exon 4 >< Exon 5
λGST-T   -------------------GTGGGGAGACAGAAGAGGAGAAGATTCGTGTGGACATTTTG  218
λT7.3    GCCCGCAAGCACAACCTGTGTGGGGAGACAGAAGAGGAGAAGATTCGTGTGGACATTTTG  300
GSTM4    AlaArgLysHisAsnLeuCysGlyGluThrGluGluLysIleArgValAspIleLeu     100

                                                                Exon 5 >
λGST-T   GAGAACCAGGCTATGGACGTCTCCAATCAGCTGGCCAGAGTCTGCTACAGCCCTGACTTT  278
λT7.3    GAGAACCAGGCTATGGACGTCTCCAATCAGCTGGCCAGAGTCTGCTACAGCCCTGACTTT  360
GSTM4    GluAsnGlnAlaMETAspValSerAsnGlnLeuAlaArgValCysTyrSerProAspPhe  120

         < Exon 6
λGST-T   GAGAAACTGAAGCCAGAATACTTGGAGGAACTTCCTACAATGATGCAGCACTTCTCACAG  338
λT7.3    GAGAAACTGAAGCCAGAATACTTGGAGGAACTTCCTACAATGATGCAGCACTTCTCACAG  420
GSTM4    GluLysLeuLysProGluTyrLeuGluGluLeuProThrMETMETGlnHisPheSerGln  140

                                        Exon 6 >< exon 7
λGST-T   TTCCTGGGGAAGAGGCCATGGTTTGTTGGAGACAAGATCACCTTTGTAGATTTCCTCGCC  398
λT7.3    TTCCTGGGGAAGAGGCCATGGTTTGTTGGAGACAAGATCACCTTTGTAGATTTCCTCGCC  480
GSTM4    PheLeuGlyLysArgProTrpPheValGlyAspLysIleThrPheValAspPheLeuAla  160

λGST-T   TATGATGTCCTTGACCTCCACCGTATATTTGAGCCCAACTGCTTGGACGCCTTCCCAAAT  458
λT7.3    TATGATGTCCTTGACCTCCACCGTATATTTGAGCCCAACTGCTTGGACGCCTTCCCAAAT  540
GSTM4    TyrAspValLeuAspLeuHisArgIlePheGluProAsnCysLeuAspAlaPheProAsn  180

                                        Exon 7 >< Exon 8
λGST-T   CTGAAGGACTTCATCTCCCGCTTTGAGGGCTTGGAGAAGATCTCTGCCTACATGAAGTCC  518
λT7.3    CTGAAGGACTTCATCTCCCGCTTTGAGGTTTCCTGTGGCATAATGTGATGGTCAATTTTC  600
GSTM4    LeuLysAspPheIleSerArgPheGluGlyLeuGluLysIleSerAlaTyrMETLysSer  200
                                        ValSerCysGlyIleMET***

λGST-T   AGCCGCTTCCTCCCAAAACCTCTGTACACAAGGGTGGCTGTCTGGGGCAACAAGTAATGC  578
λT7.3    TGCATCAACTTGACTGGGCTAAGGGATGCTCAGATGGCAGGTAAAATCATTGTGCTTGTG  660
GSTM4    SerArgPheLeuProLysProLeuTyrThrArgValAlaValTrpGlyAsnLys***     218

λGST-T   CTTGAAGGCCAGGAGGTGGGAGTGAGGAGCCCATACTCAGCCTGCTGCCCAGGCTGTGCA  638
λT7.3    AGGGTGTTTCCAGAAGAGATTTGCCTTTGAATCAGAAGACAGCAAAGATTTCCTTCAGCA  720

λGST-T   GCGCAGCTGGACTCTGCATCCCAGCACCTGCCTCCTCGTTCCTTTCTCCTGTTTATTCCC  698
λT7.3    ATGAAGGAGGCATCCACCAAACTGTCAGGCCCAGAGAGAAGAAAAGACAGGAAGGGTGAA  780

λGST-T   ATCTTTACCCCCAAGACTTTATTGGGCCTCTTCACTTCCCCTAAACCCCTGTCCCATGCA  758
λT7.3    TTTGACCTCTCTGACTGGGACATCCATCTCTGCCTATCCTGGGACCTCCACACTCCTGGT  840

λGST-T   GGCCCTTTGAAGCCTCAGCTACCCACTTTCCTTCATGAACATCCCCCTCCCAACACTACC  818
λT7.3    TCTCTGGCCTTCAGACTTGATCAGGGACTAACACCATCGCCTCCCACCCCCACCTTTGTT  900

λGST-T   CTTCCCTGCACTAAAGCCAGCCTGACCTTCCTTCCTGTTAGTGGTTGTATCTGCTTTGAA  878
λT7.3    CTGAGGCCTTTAGCCTCTGAATGATACCACTGGCTTTCCTGCTTCTCTATCCTGCAGTCG  960

λGST-T   GGCCTACCTGGCCCCTCGCCTGTGGAGCTCAGCCCTGAGCTGTCCCCGTGTTGCATGACA  938
λT7.3    GCAGATCATGGGACTTCTTCACTCCAAAATTGTGTGAGCCAATTCCCATAACAGATAGAT 1020

λGST-T   GCATTGACTGGTTTACAGGCCCTGCTCCTGCAGCATGGCCCCTGCCTTAGGCCTACCTGA  998
λT7.3    AAATTTATAAATAAACACACAAATTTCCTACA

λGST-T   TCAAAATAAAGCCTCAGCCACA
```

Apart from the deleted region encoding exon 4 and the substituted exon 8, the sequences of λGST-T and λT7·3 are identical and it appears likely that each clone represents a separate transcript of the same Mu class gene. The λT7·3 clone has a relatively long 5′ non-coding region that extends to –271 nucleotides. Recent sequencing of the gene confirms that this is a true copy of the transcript and not a cloning artefact, although λT7.3 begins 5 nucleotides before that suggested by Comstock *et al.* (1993).

3.5.3 Features of the encoded GST subunit

The complete amino acid sequence of the isoenzyme subunit can be deduced by combining sequences from λGST-T and λT7·3 and is compared in Figure 3.4 with other human Mu class sequences. After discussion with members of the GST nomenclature group (Mannervik *et al.* 1992) it is appropriate that this new enzyme be termed GSTM4 in agreement with the new nomenclature system.

GSTM4 consists of a peptide of 218 amino acids and is identical in length to GSTM1 and GSTM2. The deduced sequence of GSTM4 indicates that it has a molecular weight of 25,561 Da. This is not substantially different from the deduced molecular weights of GSTM1 (25,697 Da) and GSTM2 (25,744 Da), but is somewhat smaller than GSTM3 (26,688 Da) which is 7 amino acids longer. Overall, the amino acid sequence of GSTM4 has 86% similarity with GSTM1, 81% similarity with GSTM2 and with GSTM5 but only 65% similarity with GSTM3 (using Kimura's corrected distances, see Chapter 6.3.4 and Table 6.5).

3.6 DISCUSSION

3.6.1 Evidence for a new mu class locus

There is evidence that the human Mu class is the most extensive of all the classes in the human GST gene family. The isoenzymes GSTM1-1 and GSTM2-2 have been well characterised (Warholm *et al.* 1983, Vander Jagt *et al.* 1985, Soma *et al.* 1986, Suzuki *et al.* 1987, Board *et al.* 1988,

Figure 3.4 Comparison of the alignment of the deduced amino acid sequence of GSTM4 with those of the other known human Mu class isoenzymes. Identical residues are identified (*).

```
HSGSTM4    M----SMTLG YWDIRGLAHA IRLLLEYTDS SYEEKKYTMG DAPDYDRSQW
HSGSTM1a   M----PMILG YWDIRGLAHA IRLLLEYTDS SYEEKKYTMG DAPDYDRSQW
HSGSTM1b   M----PMILG YWDIRGLAHA IRLLLEYTDS SYEEKKYTMG DAPDYDRSQW
HSGSTM2    M----PMTLG YWNIRGLAHS IRLLLEYTDS SYEEKKYTMG DAPDYDRSQW
HSGSTM3    MSCESSMVLG YWDIRGLAHA IRLLLEFTDT SYEEKRYTCG EAPDYDRSQW
HSGSTM5    M----PMTLG YWDIRGLAHA IRLLLEYTDS SYVEKKYTMG DAPDYDRSQW
           *     * ** ** ******  ****** **  ** ** ** *  *********

HSGSTM4    LNEKFKLGLD FPNLPYLIDG AHKITQSNAI LCYIARKHNL CGETEEEKIR
HSGSTM1a   LNEKFKLGLD FPNLPYLIDG AHKITQSNAI LCYIARKHNL CGETEEEKIR
HSGSTM1b   LNEKFKLGLD FPNLPYLIDG AHKITQSNAI LCYIARKHNL CGETEEEKIR
HSGSTM2    LNEKFKLGLD FPNLPYLIDG THKITQSNAI LRYIARKHNL CGESEKEQIR
HSGSTM3    LDVKFKLDLD FPNLPYLLDG KNKITQSNAI LRYIARKHNM CGETEEEKIR
HSGSTM5    LNEKFKLGLD FPNLPYLIDG AHKITQSNAI LRYIARKHNL CGETEEEKIR
           *  **** ** ******* **   ******** * *******  *** * * **

HSGSTM4    VDILENQAMD VSNQLARVCY SPDFEKLKPE YLEELPTMMQ HFSQFLGKRP
HSGSTM1a   VDILENQTMD NHMQLGMICY NPEFEKLKPK YLEELPEKLK LYSEFLGKRP
HSGSTM1b   VDILENQTMD NHMQLGMICY NPEFEKLKPK YLEELPEKLK LYSEFLGKRP
HSGSTM2    EDILENQFMD SRMQLAKLCY DPDFEKLKPE YLQALPEMLK LYSQFLGKQP
HSGSTM3    VDIIENQVMD FRTQLIRLCY SSDHEKLKPQ YLEELPGQLK QFSMFLWKFS
HSGSTM5    VDILENQVMD NHMELVRLCY DPDFEKLKPK YLEELPEKLK LYSEFLGKRP
            ****** **     *   **     *****  **  **       * ** *

HSGSTM4    WFVGDKITFV DFLAYDVLDL HRIFEPNCLD AFPNLKDFIS RFEGLEKISA
HSGSTM1a   WFAGNKITFV DFLVYDVLDL HRIFEPKCLD AFPNLKDFIS RFEGLEKISA
HSGSTM1b   WFAGNKITFV DFLVYDVLDL HRIFEPNCLD AFPNLKDFIS RFEGLEKISA
HSGSTM2    WFLGDKITFV DFIAYDVLER NQVFEPSCLD AFPNLKDFIS RFEGLEKISA
HSGSTM3    WFAGEKLTFV DFLTYDILDQ NRIFDPKCLD EFPNLKAFMC RFEALEKIAA
HSGSTM5    WFAGDKITFV DFLAYDVLDM KRIFEPKCLD AFLNLKDFIS RFEGLKKISA
           **   * *** **  ** *      * * ***  * *** *   *** * ** *

HSGSTM4    YMKSSRFLPK PLYTRVAVWG NK
HSGSTM1a   YMKSSRFLPR PVFSKMAVWG NK
HSGSTM1b   YMKSSRFLPR PVFSKMAVWG NK
HSGSTM2    YMKSSRFLPR PVFTKMAVWG NK
HSGSTM3    YLQSDQFCKM PINNKMAQWG NKPVC
HSGSTM5    YMKSSQFLRG LLFGKSATWN SK
           *  *  *          * *   *
```

Soma *et al.* 1986, Suzuki *et al.* 1987, Board *et al.* 1988, Seidegård *et al.* 1988, DeJong *et al.* 1988a, Vorachek *et al.* 1991). In addition, Campbell *et al.* (1990) have cloned and characterised a distinct Mu class isoenzyme (GSTM3-3) from brain and testis. Other studies by Tsuchida *et al.* 1990, Hussey *et al.* (1991), Singhal *et al.* (1991) and Suzuki *et al.* (1991) have provided limited amino terminal sequence and structural data suggesting that there are additional Mu class GSTs expressed in human tissues. In this study, of the four Mu class cDNA clones characterised, two were copies of GSTM2 which has been

described by Board *et al.* (1988), now cloned by Vorachek *et al.* (1991) and was previously thought to be muscle specific. Two additional clones have been shown to represent transcripts of a novel gene to be known as *GSTM4*. The nucleotides encoding exons 3, 4 and 5 from the GSTM4 cDNAs show complete identity with the partial genomic sequence termed *mu2* by Taylor *et al.* (1991). This homology suggests that the partial gene cloned by Taylor *et al.* (1991) represents the gene encoding λGST-T and λT7·3 and these clones do not represent an allelic variant of a previously described locus. Northern blots have confirmed that this gene is transcribed in testis, (Taylor *et al.* 1991) but further studies are required to determine if it is transcribed in other tissues. The deduced amino acid sequence of the GST subunit encoded by λGST-T and λT7·3 shows greatest similarity with GSTM1 and it is clearly of the Mu class, it is therefore evident that λGST-T and λT7·3 cDNAs represent transcripts encoding a new Mu class subunit. Concurrent studies by Comstock *et al.* (1993) also identified the GSTM4 gene.

3.6.2 Splicing of the GST transcript

Notably, both GSTM4 cDNA clones show evidence of alternative splicing at different sites. The excision of exon 4 in λGST-T changes the reading frame so that no meaningful GST sequence would be translated after the end of exon 3. This gross truncation and change in the C-terminal sequence is likely to result in a catalytically inactive protein, as has been demonstrated with a similar deletion in the inactivation of an Alpha class GST (Board and Mannervik 1991). In λT7·3 the sequence encoding exon 8 has been replaced by another sequence of unknown, but possibly intronic, origin. The inclusion of this new sequence in the subunit would substitute 6 new amino acids and a new C-terminal in the place of 23 amino acids lost by the removal of exon 8. It is highly likely that a subunit derived from this transcript would also lack GST catalytic activity as the truncation of GSTA1-1 (formally GST2) by 12 residues results in a substantial loss of activity (Board and Mannervik 1991). Although

both transcripts are unlikely to encode catalytically active GSTs, they may still encode proteins with other properties such as ligand binding.

No other cases of alternative splicing have been observed in this laboratory in other studies of GST cDNAs from a variety of tissues. In addition there is no published record of alternative splicing in this large family of genes. However, Comstock *et al.* (1993) reported the isolation and analysis of two incomplete but overlapping cDNA clones of GSTM4, but in this instance the break points were unclear. cDNA clones of other GST genes isolated in the current study from the same library, including two GSTM2 clones, have not shown similar evidence of alternative splicing, suggesting that it is not a tissue specific phenomenon but may be a property of the gene. Porter and Mintz (1991) have recently reported extensive alternative splicing of transcripts of the mouse tyrosinase gene in several tissues suggesting in that case that the events were gene specific. There is a growing body of evidence suggesting that alternative splicing of transcripts provides an important mechanism that can lead either to the production of multiple protein isoforms from a single pre-mRNA or as a means of gene regulation (Smith *et al.* 1989, Maniatis 1991, McKown 1992). Thus the differing isoforms generated often show either developmental or, in many cases, tissue specificity (for reviews see Andreadis *et al.* 1987, Smith *et al.* 1989, McKown 1992).

The mechanisms of alternative splicing are also capable of rapid response to environmental stimuli. It is possible that alternative splicing provides a mechanism allowing the cell to test, generally without detriment, novel proteins. Because it is post-transcriptional, genome information is not necessarily irreversibly altered and all the possibilities inherent in the gene remain available. Whilst it can be seen that alternative splicing increases the coding capacity of the genome, unfortunately, the selective advantages of encoding multiple isoforms in a single gene are not yet clear. Nonetheless, given the number of genes now known to be regulated by alternative

splicing, this is clearly a significant alternative to the regulation of promoter activity (Andreadis *et al.* 1987).

It has been postulated that alternative isoforms, and presumably the mechanisms providing such new diversity, have been maintained through evolution because of the selective advantage of the diversity itself (Andreadis *et al.* 1987). It should be noted that we are currently aware of only those cases of regulation by alternative splicing in which the RNA products can be relatively easily observed. It is possible that the actual level of alternative splicing of transcripts may be much higher. RNA products may have a rapid turnover (lack stability) or may be present at low levels, in which case they may either go unnoticed or be discarded as unimportant or artefacts of the purification process. Although many of the processes regulating constitutive mRNA splicing are now understood and several different types of alternative splicing are recognised, the identification of each of these mechanisms has not yet led to an understanding of the factors regulating alternative splicing. Importantly, there is still much to be learnt about the role and full significance of alternative splicing in gene regulation.

Most GST genes studied so far appear to be predominantly regulated by the initiation of transcription (Daniel *et al.* 1989, Okuda *et al.* 1989, Rushmore and Pickett 1990). Morrow *et al.* (1992) have suggested that the *GSTP1* gene may be regulated by post-transcriptional mechanisms. It is possible that the alternative splicing of *GSTM4* transcripts represents a form of gene regulation not previously noted in the GST gene family. This regulation could take several forms; in both the *su(w^a)* (suppressor of white apricot) gene in *Drosophila,* and the *ras* oncogene two alternative transcripts are produced. In both cases it is the minor transcript which is the functional form within the cell. In the case of *ras,* a majority of the primary transcript is processed with an extra exon containing stop codons, thus producing a truncated and apparently non-functional protein. The existence of the truncated transcript was only described from mutant studies, as standard RNA

blots and protection assays failed due to its extreme instability (Cohen *et al.* 1989). The opposite is true of the non-functional major transcript in *su(w^a)* which is the more stable, although formed by incomplete splicing. In this system it is the functional, minor form, that is present at such low levels as to be difficult to detect (Chou *et al.* 1987, Zachar *et al.* 1987). The regulatory mechanism involved in the alternative splicing of *ras* is unclear but evidence has been put forward that in *su(w^a)* the function of the alternative splicing is a form of autoregulation (Zachar *et al.* 1987).

Additionally, studies have shown that cells may permit a limited amount of transcription of genes not normally required in that cell type, this has been termed illegitimate transcription (Chelly *et al.* 1989, Fonknechten *et al.* 1992). Whether this represents a necessary requirement of the cell, providing selective advantages, or simply a lack of absolute control in regulation of transcription which is not detrimental remains a moot point. It is not clear if normal splicing always takes place after illegitimate transcription. Although normal splicing is commonly reported, Roberts *et al.* (1993) have reported the inclusion of a novel exon in 50% f the transcripts obtained. from illegitimate transcription.

It may be possible that GST-T and T7·3 represent an inherent variability in the splicing process. Both illegitimate transcription and the proposed inherent variability in the splicing process may be initial stages in a continuum of permissive transcription that eventually leads to regulated alternative or even constitutive splicing of the mRNA transcript of a novel protein. Production of new protein isoforms within a cell can then proceed without removing the old ones or the requirement for permanent change to the nucleotide sequence (Andreadis *et al.* 1987, Porter & Mintz 1991). This would lend weight to the view that alternative splicing and the regulatory mechanisms it requires are maintained because of the selective advantage of the isoform diversity itself (Andreadis *et al.* 1987).

Thus the GST-T and T7-3 transcripts of *GSTM4* may represent either illegitimate transcription or examples of "experimental" alternative splicing in the testis. Incomplete cDNA clones of GSTM4 have been reported elsewhere but it is not known if these also represent possible alternatively spliced transcripts (Comstock *et al.* 1993). Evidence recently published suggests that a 1.5-1.6kb GSTM4 mRNA is expressed in appreciable amounts in heart, brain, liver, muscle and kidney tissue (Comstock *et al.* 1993), using a probe from the 5' untranslated region which is likely to be specific for GSTM4 rather than hybridize with other members of the Mu class. But the presence or absence of a shorter mRNA was not noted. Further studies are required to determine the full extent of tissue expression of *GSTM4* and if truncated forms of the mRNA or protein are detectable. GSTM4 has not been purified from any tissue and it may be that complete transcripts are only expressed under certain conditions. Alternatively the GSTM4 isoenzyme may not have been detected because of its low activity with CDNB (Chapter 4).

It is interesting to note that Taylor *et al* (1991) have recently speculated that the human Mu class genes may be subjected to a novel form of regulation by somatic gene conversion where intronic sequences containing elements that influence transcription may convert a related gene and alter its regulation. If this proposed mechanism does function within the Mu class genes, it is not immediately evident what role, if any, alternative splicing would play.

3.7 SUMMARY

Two cDNA clones encoding a new Mu class glutathione *S*-transferase have been isolated from a human testis cDNA library. Both clones are incomplete and appear to result from alternative splicing. One clone is missing the sequence encoding exon 4 and the other is missing exon 8. The complete sequence of the previously undescribed isoenzyme can be deduced from the two cDNA clones. This is the first report of alternative

splicing in a GST transcript and may represent either a novel form of regulation in this multigene family or illegitimate transcription and experimental alternative splicing which may be part of a selection process, eventually leading to novel protein isoforms or transcription in other cell types.

CHAPTER 4 CHARACTERISATION OF THE GSTM4 ISOENZYME AND COMPARISON WITH OTHER HUMAN MU CLASS ISOENZYMES

4.1 INTRODUCTION

Studies of the human Mu class isoenzymes in a number of tissues have suggested that the class is quite extensive (Mannervik 1985, Suzuki *et al.* 1987, Board *et al.* 1988, Campbell *et al.* 1990, Tsuchida *et al.* 1990, Suzuki, *et al.* 1991, Hussey *et al.* 1991, Singhal *et al.* 1991). So far, five functional Mu class genes have been identified by the isolation of cDNA clones (Seidegård *et al.* 1988, DeJong *et al.* 1988a, Campbell *et al.* 1990, Vorachek *et al.* 1991, Listowsky *et al.* 1993 and Chapter 3), it is not clear if these cDNAs represent all the functional Mu class genes.

Broad substrate specificity is a characteristic of detoxication enzymes. Although isoenzymes within the GST family generally show overlapping substrate specificities, individual isoenzymes can have highly specific substrate preferences. The Mu class isoenzymes are no exception and appear to exhibit both the versatility and specificity characteristics of the family as a whole (Chapter 1.4.3.1).

There are indications that GSTs may also play extensive roles beyond the conjugation reaction. There is considerable evidence for a binding function, predominant in the Alpha class but still considerable for the isoenzymes of the Mu class (Habig and Jakoby 1981). The extent and full significance of this apparent storage and regulatory role is yet to be resolved. Even less well defined is the role of GSTs in biosynthesis. Although both Alpha and Mu class isoenzymes from rat are capable, *in vitro,* of converting leukotriene A_4 to C_4, it is the rat Mu class GSTs 4-4 and 6-6 (isolated from brain) that have the highest activity of those tested (Mannervik *et al.* 1984b, Tsuchida *et al.* 1987, Abramovitz and Listowsky 1987). Given that a specific microsomal

enzyme leukotriene C_4 synthase has been isolated and purified recently (Penrose *et al.* 1992), the extent to which the cytosolic GSTs participate in this pathway *in vivo*, needs to be clarified.

Individual Mu class isoenzymes have been characterised with a range of substrates in several laboratories, however, it has not been possible to directly compare results before. Although the recombinant form of GSTM1-1 has been studied previously (Widersten *et al.* 1991), the recent publication of the cDNA sequences of several additional human Mu class isoenzymes now permits comparative studies of the isolated recombinant proteins to be undertaken. This may lead to a fuller understanding of the role individual isoenzymes play in detoxication and protection of the cell.

4.2 AIM

The cloning of a cDNA encoding a novel human Mu class GST is described in Chapter 3. The aim of the experiments reported in this chapter was to engineer the expression of GSTM4 in *E. coli*, therefore permitting the characterisation of the purified isoenzyme for the first time. A further intention of this study was to compare of the purified GSTM4 with previously described human Mu class GSTs.

4.3 MATERIALS

4.3.1 Bacterial strains, and plasmids

E. coli (TG1) was again used as the host strain for these experiments (Chapter 2.1.1). The vector pKK261 was used to express the proteins. This vector has been modified from pKK223-3 (Pharmacia) by removing a *Bam*HI-*Sal* I fragment from the backbone of the vector and was a gift of Dr. Rohan Baker, Molecular Genetics Group, JCSMR, A.N.U. RS broth (2% tryptone, 1·5% yeast

extract, 0·5 % NaCl) was used for E. coli cultures harbouring pKK expression plasmids. Details of the reagents, enzymes and buffers used are listed in Tables 2.1.and 2.2.

4.4 METHODS

Standard recombinant DNA and preparative methods used in this study are described in Chapter 2, those techniques unique to these experiments are given here in detail.

4.4.1 Construction of a full length GSTM4 cDNA transcript

To determine if the GSTM4 protein is functional, a plasmid containing a combined cDNA was constructed from the λGST-T and λT7·3 sequences and placed down stream of a tac promoter by the strategy shown in Figure 4.1. The plasmid subclones pGST-T and pT7·3 were digested with *Nco*I, which cuts once within the cDNA, and *Hind* III, which cuts only in the polylinker. The vector containing the 5′ portion of insert from pT7·3 and the *Nco*I to *Hind* III insert fragment (consisting of the 3′ sequence) from pGST-T were recovered by agarose gel electrophoresis. These fragments were ligated together and transformed into *E. coli* (TG1). The structure of the resulting clone (pGSTM4) was verified by restriction analysis with *Bam*HI and *Bgl* I. Polymerase chain reaction (**PCR**) amplification with Taq polymerase (Promega) was used in the construction of the final insert, essentially following the manufacturer's basic protocols. To facilitate cloning and expression an *Eco* RI site and a ribosome binding site were added to the GSTM4 cDNA with the 5′ forward oligonucleotide primer (GTT ExA). The 3′ reverse oligonucleotide primer (GT1 ExB) incorporated a *Sal* I restriction site (primers are detailed in Table 4.1). PCR was carried out for an initial 5 cycles with the denaturing step at 93°C for 1 minute, annealing at 40°C for 1 minute and extension at 70°C for 1 minute. This was followed by 25 cycles with the denaturing step at 93°C for 1 minute,

annealing and extension at 70°C for 2 minutes on an Corbett FTS-1 thermal cycler (Corbett Research, Sydney, Australia).

Figure 4.1 Schematic diagram showing the construction of pGSTM4. The *Nco* I to *Hind*III fragment was removed from pT7·3 and replaced by the equivalent fragment from pGST-T. The backbone plasmid is pUC18.

The PCR product was digested with *Eco*R1 and *Sal* I and isolated by agarose gel electrophoresis. The fragment was purified using Geneclean II, ligated with the expression vector pKK261 and transformed into *E. coli* (TG1). Screening of clones was carried out by the growth of IPTG-induced 5ml cultures, sonication and spectrophotometric GST assay with CDNB (section 4.4.4). A selection of clones showing activity were recloned into M13 and sequenced for verification. The clone used in the following studies for the expression of GSTM4-4 was termed pKKGSTM4.

Table 4.1. Oligonucleotide primers used for amplification and addition of restriction sites to human Mu class cDNAs.

cDNA	oligonucleotide	sequence
GSTM2	GT4 ExA	5′ TCGAATTCAGGAGACTGCTATCATGCCCATGACACTG 3′
	GT1 ExB	5′ AAGTCGACCCTACTTGTTGCCCCAGAC 3′
GSTM3	GTM3 ExA	5′ TCGAATTCAGGAGACTGCTATCATGTCGTGCGAGTG 3′
	GTM3 ExB	5′ AACAAGCTTGCAAGTCTGCCTCCTGC 3′
GSTM4	GTT ExA	5′ TCGAATTCAGGAGACTGCTATCATGTCCATGACACTG 3′
	GT1 ExB	5′ AAGTCGACCCTACTTGTTGCCCCAGAC 3′

The names of the forward primers end in A and the names of the reverse primers end in B.

4.4.2 Generation of other recombinant Mu class GSTs

To compare the properties of all the known human Mu class GSTs the following enzymes were expressed and purified; GSTM1b-1b (pKKGST1, generously provided by Dr. Gordon S. Rule, Department of Biochemistry, University of Virginia), GSTM2-2 (pKKGSTM2) and GSTM3-3 (pKKGSTM3). The latter two were generated by PCR amplification from the testis cDNA library. The primers were derived from the published sequences and are detailed in Table 4.1. GSTM2-2 and GSTM3-3 were amplified and cloned using similar procedures as those described for the production of pKKGSTM4 (section 4.4.1), except that the 3′ reverse primer for GSTM3 has a *Hind* III site instead of a *Sal* I site.

Clones expressing GSTM2-2 and GSTM3-3 were sequenced for confirmation. In comparison with the published nucleotide sequence (Vorachek *et al.* 1991) the GSTM2 sequence had a single T to C substitution at base 222 (numbered from the initiating ATG). This is a silent change and does not alter the amino acid sequence. This may be a natural polymorphism. It is unlikely that it is a PCR error as the same change was noted in a direct clone retrieved from the cDNA library (Chapter 3.5.2). In the GSTM3 sequence we identified a T to G substitution at base 440 when compared to the published sequence (Campbell *et al.*1990). This makes a Trp 147 Gly substitution which may also be a natural polymorphism. A PCR error can be discounted because the same substitution has been identified independently (M. Widersten and B. Mannervik personal communication). This substitution is unlikely to have a critical effect on function as this residue is Gly in GSTM1, GSTM2 and GSTM4.

4.4.3 Purification of GST enzymes from *E. coli*

Each *E. coli* (TG1) clone containing the required expression plasmid was grown in two 500ml cultures to an OD_{555} of 0·4 then induced with 0·2mM IPTG and grown overnight. Bacteria were collected by centrifugation, resuspended in an equal volume of buffer (50mM Tris-HCl pH7·2; 50mM EDTA; 15% glucose; 1mg/ml lysozyme) and incubated on ice for 1 hour. After addition of 170µM phenylmethanesulfonyl fluoride (PMSF) the cells were sonicated (Branson Sonifer 250) 3X20 seconds at setting 4 and the soluble fraction was collected after centrifugation at 25,000g for 20 minutes. The extraction procedure was repeated on the pellet and the soluble fractions pooled.

All enzymes were purified by affinity chromatography on pentyl-glutathione agarose according to a modification of the method described by Mannervik and Guthenberg (1981) using 50mM Tris-HCl pH7·2. Purification was monitored by activity assay using CDNB (section 4.4.4.). The GSTs were eluted from the column in 50mM Tris-HCl pH7·2 containing 5mM pentyl-glutathione.

The fractions exhibiting activity were dialysed against 50mM Tris-HCl pH7·2, 1mM EDTA, 0·2mM dithiothreitol. The protein was then concentrated in a Diaflo PM10 ultrafilter (Amicon Corp., Lexington MA., USA) and immediately assayed for activity with a range of substrates (Table 4.3). All operations in the purification procedure were performed at 4°C. The purified enzymes were examined by SDS/PAGE using the method of Laemmli (1977). Protein assays were carried out using the method of Lowry *et al.* (1951) as modified by Peterson (1977).

4.4.4 Enzyme assays

Specific activities were determined with a range of substrates (see Table 4.3). The spectrophotometric assays were carried out according to the procedures of Habig and Jakoby (1981) and Brophy *et al.* (1989). The assays depend on a direct change in the absorbance spectrum when the substrate is conjugated with GSH. The CDNB activity assay is the standard used in such studies for monitoring purification and for characterisation. It is described in more detail as a model for the other assays.

CDNB undergoes nucleophilic displacement of the chloro substituent by GSH, the spectral change accompanying its conversion to the thioether product can be monitored at 340nm (a yellow colour develops and can often be seen visually). A typical assay contains, in a final volume of 1ml, 100µl 100mM Na phosphate, 1mM EDTA, pH6.5; 100µl 20mM GSH; an aliquot of enzyme sample (usually 10µl), 100µl 20mM CDNB. After brief but thorough mixing in a cuvette, the change in absorbance at 340nm is recorded, for 2-4 minutes on a spectrophotometer. Glutathione peroxidase activity was determined by the spectrophotometric method described by Beutler (1975). A complete assay mixture minus the enzyme served as control in all experiments. All spectrophotometric assays were carried out at 30°C.

Activity with tritiated *trans*-stilbene oxide was determined by a procedure based on the method of Seidegård *et al.* (1984) and Seidegård

and Pero (1985). The activity was measured in 250mM Tris-HCl pH7·2 containing 5mM reduced glutathione and 4·75μM *trans*-stilbene oxide. Initially the reaction mixture was incubated at 37°C and terminated after 5, 10 or 15 minutes. Over the longer time span it was found that the reaction rate was non-linear therefore all subsequent reactions were determined over 5 minutes. All reactions were terminated by extraction of the unreacted substrate with two volumes of hexanol. Vortexing and centrifugation produced phase separation. It has been previously determined that 99·8% of the unreacted substrate is extracted by this method (Gill *et al.*1983). The amount of tritiated product was determined in a 20μl aliquot of the aqueous phase by liquid scintillation counting.

4.4.5 Amino acid sequencing

The four recombinant Mu class enzymes studied were subjected to N-terminal amino acid sequencing for at least six cycles on an Applied Biosystems automated sequencer model 477A. Each of the isoenzymes has a unique amino acid sequence within this section of the N-terminal. This was carried out by Dr. Dennis Shaw, Division of Biochemistry and Molecular Biology, JCSMR, A.N.U. My contribution was limited to supplying the protein samples and interpretation of results.

4.4.6 Immunological characterisation

Purified proteins were separated on 12% SDS/PAGE and electro-blotted onto nitrocellulose filters as described by Towbin *et al.* (1979). The filters were then probed with antiserum directed against either denatured GSTM1b-1b or native GSTM2-2. Cross-reacting antigen was detected using a rabbit primary antigen and goat anti-rabbit IgG second antibody coupled with alkaline phosphatase, as described by Board *et al.* (1988). The antiserum was the gift of Dr. Philip Board, Molecular Genetics Group, JCSMR, A.N.U.

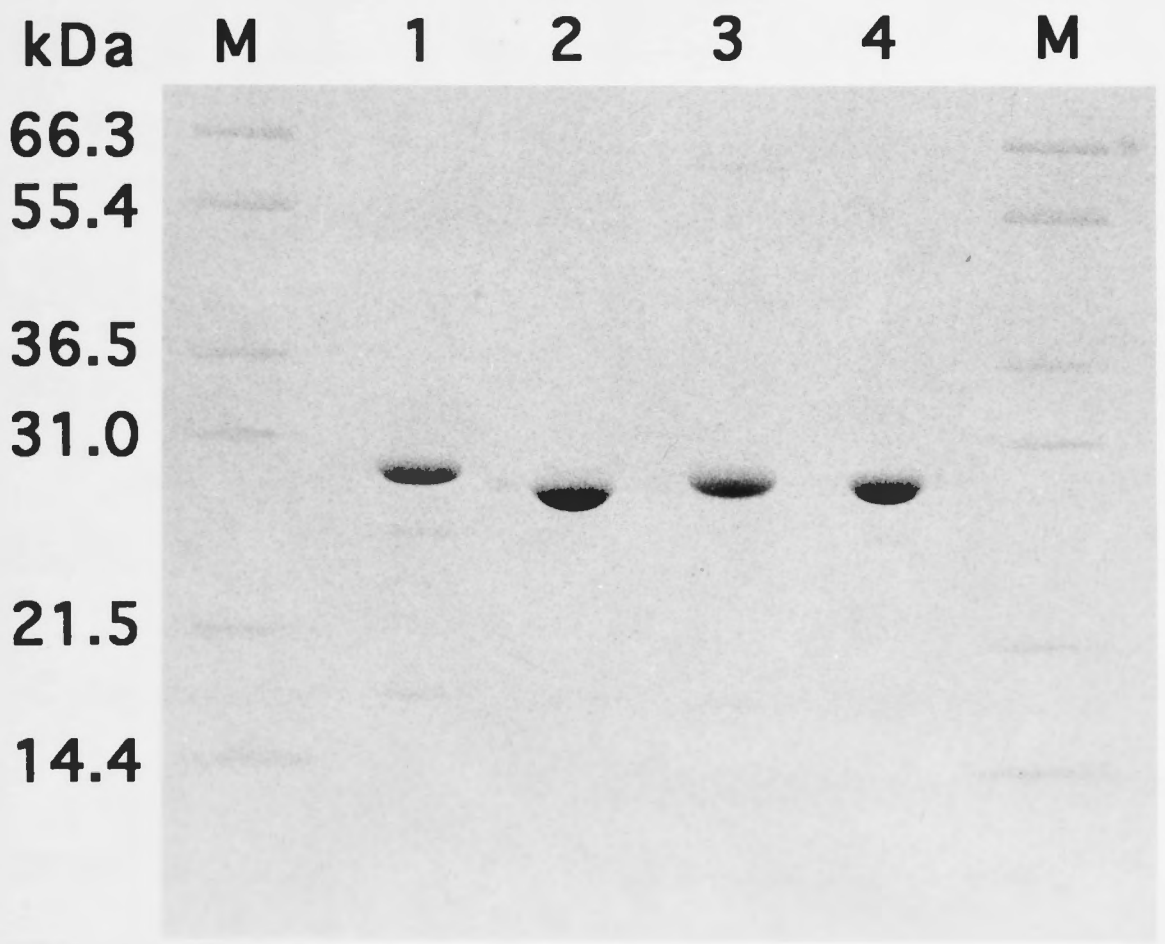


Figure 4.2 SDS/PAGE of recombinant human Mu class GSTs. Lane 1, GSTM1b; lane 2, GSTM2; lane 3, GSTM3; lane 4, GSTM4; lanes M, standard size markers.

4.4.7 Native size determination

Purified recombinant GSTs were passed through an FPLC superose-12 column (Pharmacia). Elution volumes were determined in a buffer consisting of 150mM NaCl, 10mM $NaPO_4$, pH7·2, in comparison with protein standards of known size (bovine serum albumin 66kDa, ovalbumin 45kDa and trypsinogen 24kDa).

4.4.8 Isoelectric focusing

To determine the isoelectric point, the recombinant proteins were subjected to isoelectric focusing in the range pH 4 - 6·5 on polyacrylamide gels with standard proteins of known pI on a Phastgel system (Pharmacia).

4.5 RESULTS

4.5.1 Characterisation of recombinant human Mu class GST enzymes

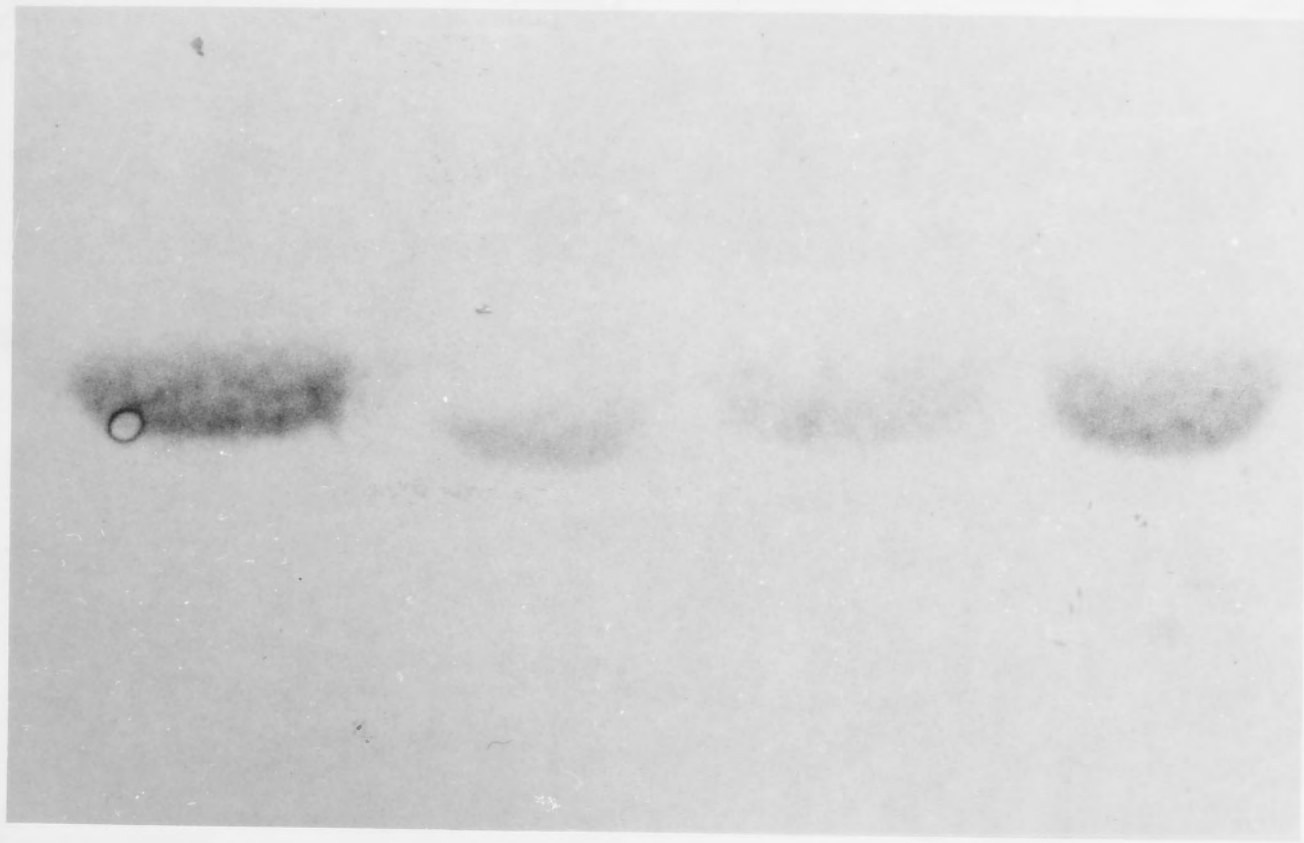
When pKKGSTM4 was grown in the presence of IPTG significant quantities of recombinant GSTM4-4 were expressed and preliminary determinations demonstrated it was functional with CDNB as a substrate. Similar results were obtained with the plasmids pKKGST1, pKKGSTM2 and pKKGSTM3. Large scale preparations yielded purified recombinant GSTM4-4 at the rate of 12mg/L. Recombinant GSTM1 was produced at the rate of 4mg/L by pKKGST1; GSTM2-2 was produced by pKKGSTM2 at the rate of 14mg/L and pKKGSTM3 produced GSTM3-3 at a rate of 6mg/L.

4.5.2 Subunit and protein size, and immunological relationships

Samples of the purified GSTs were analysed by SDS/PAGE and are shown in Figure 4.2. The M_rs of GSTM1, GSTM3 and GSTM4 are similar. In this experiment GSTM2 appears to be slightly smaller than the other subunits despite the fact that it has a deduced molecular weight similar to that of GSTM1 and GSTM4. Although GSTM3 has a deduced molecular weight

A

1 2 3 4



B

1 2 3 4

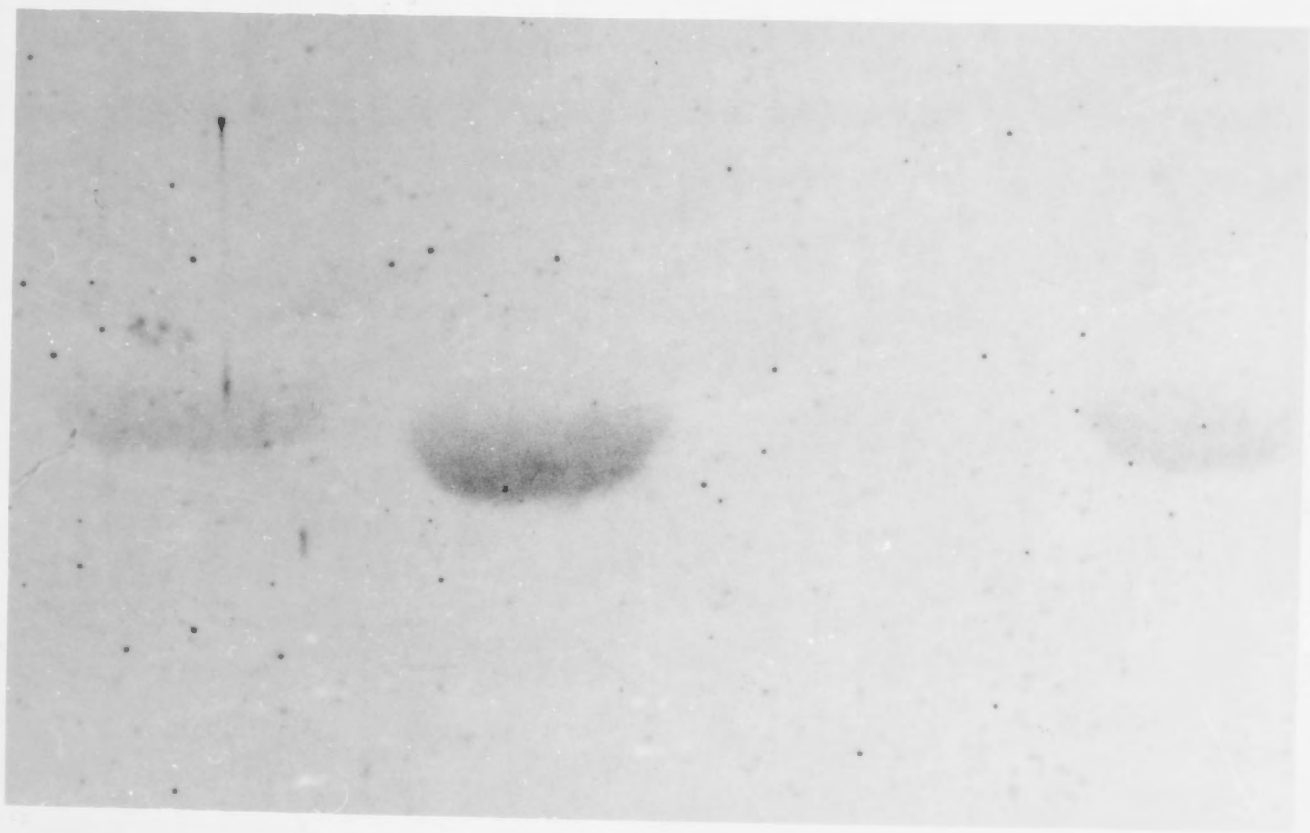


Figure 4.3 Western blots of SDS/PAGE gels probed with **A.** antiserum raised against denatured GSTM1-1. **B.** antiserum raised against native GSTM2-2. Lane 1, GSTM1-1, lane 2, GSTM2-2, lane 3, GSTM3-3, lane 4, GSTM4-4.

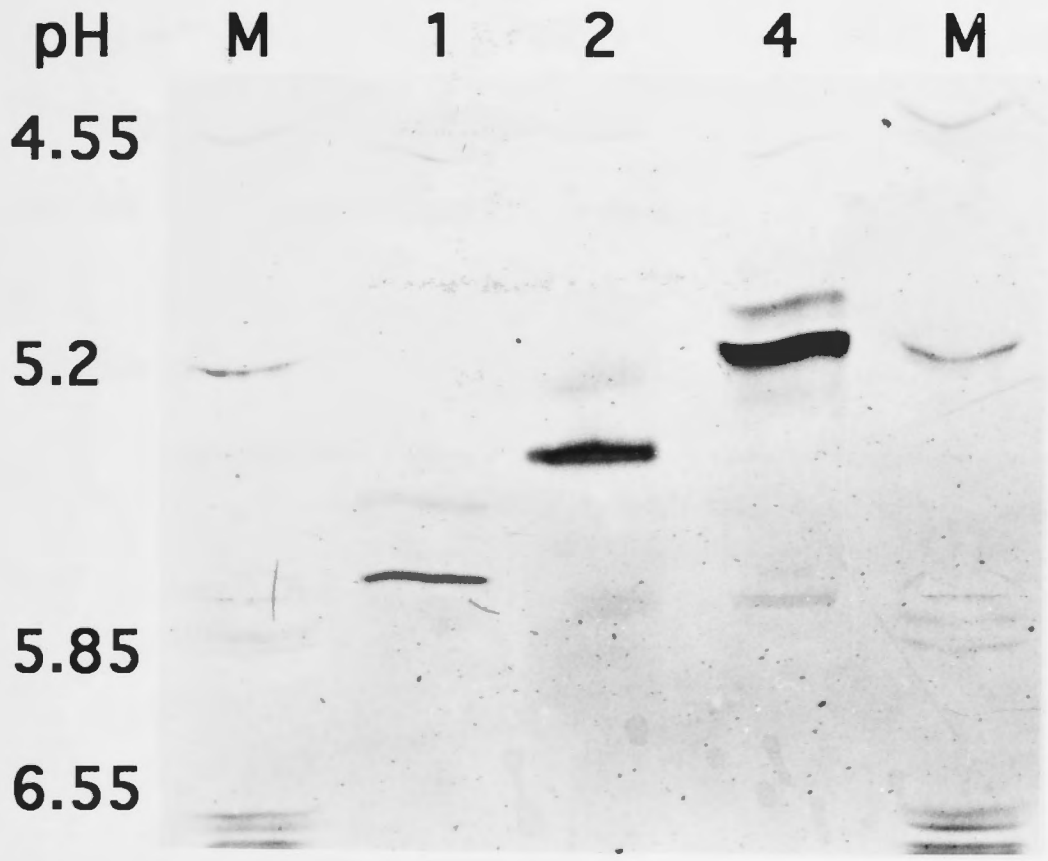


Figure 4.4 Isoelectric focusing of recombinant human Mu class enzymes. Lane 1, GSTM1b-1b; lane 2, GSTM2-2; lane 4, GSTM4-4; lanes M, standard proteins of known isoelectric point.

approximately 1kDa greater than that of the other subunits, this difference could not be reliably detected in a number of experiments.

To confirm that the recombinant GSTs were assembling in *E. coli* as dimers, the purified proteins were subjected to gel filtration on an FPLC superose-12 column in comparison with standard proteins of known size. The estimated sizes of the recombinant proteins are shown in Table 4.2 in comparison with their deduced subunit molecular weights. Although they are slightly smaller than may be predicted from their deduced sequence, they are clearly in the range expected for dimers.

Table 4.2 Size of native recombinant Mu class GSTs determined by gel filtration.

	deduced subunit kDa	native protein[a] kDa
GSTM1b-1b	25.7	40.6
GSTM2-2	25.7	41.7
GSTM3-3	26.7	49.0
GSTM4-4	25.6	41.9

[a] mean of two determinations.

Western blots of SDS/PAGE gels were probed with antiserum raised against either denatured GSTM1-1 or native GSTM2-2 (Figure 4.3). Antiserum directed against GSTM2-2 recognised GSTM1 and GSTM4, however, it failed to recognise GSTM3. In contrast, antiserum raised against denatured GSTM1-1 cross reacted with all Mu class subunits. Possibly denaturation of the antigen reveals some epitopes conserved in all four isoenzymes.

4.5.3 Isoelectric focusing

The isoelectric points of the recombinant GSTs were determined by Isoelectric focusing in the range pH 4 - 6·5 in polyacrylamide gel (Figure 4.4). The p*I* of GSTM1b-1b was pH5·7, the p*I* of GSTM2-2 was pH5·4 and the p*I* of GSTM4-4 was pH5·2. A comparison of GSTM4-4 with GSTM1b-1b reveals 6

[1] The p*I* quoted for each GST enzyme is for the strongest band. Clearly there is more than one band present in each lane,.this which may be due to subtle post-translational modifications, such as the removal or retention of the initiating methionine, as described in the amino acid sequencing (section 5.5.4). The smearing consistently present in the GSTM3-3 lanes (data not shown) may also be due to this or to initiation of translation from the second methionine (section 5.5.4).

amino acid substitutions that result in charge differences. Although these charge changes are balanced, and there is no net change, the positions of the substitutions in the mature folded protein clearly influence its isoelectric point. GSTM3-3 consistently migrated as a smear on isoelectric focusing gels and an accurate p*I* could not be estimated by this method.[1]

4.5.4 Amino terminal sequencing

The amino terminal sequences for all four recombinant proteins were identical with those predicted from their respective cDNAs. Approximately 15% of the GSTM3 preparation retained the initiating methionine, whereas that fraction was ≤ 5% in the cases of GSTM1, GSTM2 and GSTM4. The amino terminal sequencing also detected a minor (≅2%) fraction where the first three residues were deleted, suggesting that in a small fraction of transcripts the translation of protein is initiated from the second methionine in the sequence. The small percentages of proteins with either the retention of the first methionine or initiation of transcription from the second methionine may well be due to expression of the isoenzymes in *E coli* rather than mammalian cells. Alternatively, this low level of aberrant translation may occur, *in vivo*, and not have been previously noted. None of the previous studies of purified Mu class enzymes have reported retention of the initiating methionine or initiation from the second methionine.

4.5.5 Substrate specificities

The compounds used for enzyme rate assays are known to be substrates for GST enzymes from the different classes, though not all have been reported as efficient Mu class substrates (Table 4.3). Recombinant GSTM4-4 has a lower level of activity with the model substrate CDNB in comparison to either GSTM1b-1b or GSTM2-2 but the value is similar to that for GSTM3-3. All the Mu class enzymes showed a low but consistent activity with ethacrynic acid.

Table 4.3 Comparison of specific activity of human Mu class isoenzymes with a range of known GST substrates.

substrate	GSTM1		GSTM2		GSTM3		GSTM4	
1-Chloro 2,4-dinitro benzene	107.4	±8.1	186.4	±7.1	6.0	±0.5	1.25	±0.02
1,2-Dichloro 4-nitro benzene	0.7	±0.30	3.4	±0.5	0.11	±0.01	n.d.	
1,2-epoxy-3-(*p*-nitrophenoxy) propane	n.d.		7.0	±0.5	n.d.		n.d.	
trans -4-phenyl-3-buten-2-one	0.06	±0.01	0.07	±0.07	n.d.		n.d.	
p-nitrophenol acetate	n.d.		1.7	±1.13	0.20	±0.02	0.03	±0.01
ethacrynic acid	0.93	±0.23	3.22	±0.09	0.06	±0.01	0.04	±0.002
cumene hydroperoxide	n.d.		n.d.		n.d.		n.d.	
t-butyl hydroperoxide	n.d.		n.d.		n.d.		n.d.	
trans-stilbene oxide	10477*	±1059	0.29*	±0.58	0.41*	±0.81	2.95*	±2.54
trans, trans-hexa-2,4-dienal	0.15	±0.05	0.17	±0.02	0.03	±0.01	0.004	±0.001
trans, trans-hepta-2,4-dienal	0.34	±0.04	0.43	±0.03	0.07	±0.06	0.02	±0.001
trans, trans-deca-2,4-dienal	0.36	±0.005	0.52	±0.01	0.08	±0.03	0.03	±0.01
trans-hex-2-enal	0.58	±0.02	0.12	±0.02	0.35	±0.01	0.01	±0.001
trans-oct-2-enal	1.07	±0.20	0.24	±0.01	0.54	±0.03	0.09	±0.07
trans-non-2-enal	1.48	±0.05	0.31	±0.04	0.75	±0.18	0.10	±0.02

Values are expressed as µmol.mg-1.min-1. n.d., not detectable.
* Values for *trans*-stilbene oxide activity are expressed as nmol.mg-1.min-1
All values are the mean of triplicate determinations ± standard deviation.

GSTM4-4 has no discernible activity with either 1,2-dichloro 4-nitro benzene (DCNB) or 1,2-epoxy-3-(*p*-nitrophenoxy)propane (NPEP), although all the others have some activity with the former and at least GSTM2-2 has activity with the latter. In this study GSTM1b-1b had no detectable activity with the second substrate, p-nitrophenyl acetate (pNPA) and the activity exhibited by the remaining enzymes is low.

Mu class GST enzymes have been considered as efficient catalysts with some epoxide substrates (Seidegård and Pero 1985, Mannervik and Danielson 1988). However, in this comparison of Mu class recombinant enzymes GSTM1b-1b has at least a thousand-fold higher activity with *trans*-stilbene oxide than any of the other enzymes. The reaction rate of GSTM4-4 is marginally higher, though not significantly so, when compared to GSTM2-2 and GSTM3-3. In this study there was no detectable peroxidase activity in any of the recombinant enzymes, however, Campbell *et al.* (1990) reported a low level of activity against cumene hydroperoxide with GSTM3-3 purified from human testis.

Although there was no detectable activity with the lipid hydroperoxide model substrates, all the Mu class enzymes showed conjugating activity with both classes of reactive carbonyl compounds *trans*-alk-2-enals and *trans, trans*-alk-2,4-dienals. The activity of each enzyme increased with increasing carbonyl chain length in both series. GSTM1b-1b and GSTM2-2 are more reactive with the *trans, trans*-alk-2,4-dienals than either GSTM3-3 or GSTM4-4. Conversely, GSTM3-3 has a higher activity with the *trans*-alk-2-enals than does GSTM2-2.

4.6 DISCUSSION

Western blots of human lung samples show evidence of a ubiquitous protein which cross-reacted with GSTM1-1 (M) antibodies and is slightly smaller than the liver GSTM1-1 standard (Carmichael *et al.* 1988). In light of the results

described here, it is entirely possible that the antibody used by Carmichael *et al.* (1988) was cross-reacting with one of the other Mu class GSTs, probably with GSTM2 which also appears to be a slightly smaller molecule on the comparative western blot described above. If this is the case then GSTM2, previously thought to be muscle specific, has now also been detected in lung (Carmichael *et al.* 1988) and in testis (Chapter 3).

The different classes of GST enzymes have characteristic but overlapping substrate specificities. Previously, it has been difficult to directly compare the catalytic properties of a range of isoenzymes from within a class. By the use of heterologous expression in *E. coli*, it has been possible to prepare recombinant forms of four known human Mu class isoenzymes and compare their properties under the same conditions.

The marked specificity of GSTM1b-1b for *trans*- stilbene oxide is clear. In comparison, the other isoenzymes show no similar dramatic substrate preferences. This suggests that the primary substrate specificities of the other enzymes may not have been included in this study and have yet to be identified. Previous studies by Seidegård *et al.* (1984) have suggested that individuals with GSTM1 deficiency can be identified by assaying blood monocyte GST activity with *trans*-stilbene oxide as a substrate. Although previous studies by Seidegård *et al.* (1984) have shown that Alpha and Pi class isoenzymes do not utilise this substrate the present results demonstrate for the first time that other human Mu class isoenzymes do not utilise *trans*-stilbene oxide at a significant rate. It has yet to be determined if the recently described GSTM5 isoenzyme uses *trans* stilbene oxide as a substrate. This study is further confirmation of the value of *trans*-stilbene oxide for the detection of GSTM1 deficiency.

The specific activity of GSTM4-4 was low with all the substrates tested in this study. It s low activity with CDNB possibly explains why this enzyme has not been detected in previous studies.

Consideration of the marked substrate specificity for *trans*-stilbene oxide shown by GSTM1-1 when compared with other human Mu class GST isoenzymes and the high level of sequence similarity in the human Mu class, led to a comparison of the the amino acid sequences being undertaken in an attempt to learn more about the second substrate binding site. Ji *et al.* (1992) have solved the structure of rat subunit 3, a Mu class GST, providing evidence of both the residues concerned with GSH binding and regions of probable importance in the binding of xenobiotic substrates. The significance of the 11 residues implicated in GSH binding in the rat subunit 3 (Y7, L13, R43, W46, K50, N59, L60, P61, Q72, S73 and N102—numbers include initiating methionine) is confirmed by their conservation in other human Mu class sequences (Figure 3.4). Of those residues considered important in hydrophobic substrate binding, (Y7, V10, L13, M105, M109, I112, M113, Y116 and S210) residues I112, M113 and S210 are the only ones that are unique to GSTM1 among the human Mu class sequences. It is therefore possible that these may be of importance in *trans*-stilbene oxide binding. Interestingly, in a further study, using modular replacement mutagenesis, Zhang *et al.* (1990) implicated the C-terminal in determining the stereoselectivity of the second substrate. Thus there are some grounds for suggesting that S210 may be involved in the marked stereoselectivity of GSTM1-1 in favour of *trans*-stilbene oxide compared to the *cis*- form as reported by Seidegård *et al.* (1984). Further analysis of sequence comparisons between all the classes of GSTs are reported in Chapter 6 and within the human Mu class in Chapter 7.

4.7 SUMMARY

By combining components from both clones, pGST-T and pT7.3 a complete cDNA has been constructed and the encoded GSTM4 protein expressed in *E. Coli*. Similar constructs have been made of other human Mu class GST isoenzymes. The activity and characterisation of the purified proteins

thus produced has been compared. In general, the recombinant GSTM4-4 enzyme has relatively low activity with all the substrates tested and this may explain why it has not been previously detected.

CHAPTER 5 CHROMOSOMAL LOCATION OF THE MU CLASS GLUTATHIONE S-TRANSFERASES

5.1 INTRODUCTION

Five human Mu class GSTs have now been cloned and characterised; GSTM1, the predominant form expressed in the liver (Board 1981a, Suzuki *et al.* 1987, DeJong *et al.* 1988a, Seidegård *et al.* 1988); GSTM2, considered to be essentially muscle specific (Laisney *et al.* 1984, Suzuki *et al.* 1987, Board *et al.* 1988, Vorachek *et al.* 1991) and known to be expressed in testis (Chapter 3); GSTM3, observed in brain and testis (Campbell *et al.* 1990), GSTM4, the novel gene reported in Chapter 3 and GSTM5 recently reported by Listowsky *et al.* (1993). Evidence has also been presented for two additional Mu class isoenzymes, GST5 and GST6 (Suzuki *et al.* 1987, 1991). Tsuchida *et al.* (1990) characterised five Mu class subunits from human heart and aorta. N-terminal amino acid sequencing indicated that two of these were previously undescribed variants and may be products of distinct genes. It is not yet clear how these isoenzymes relate to the cDNAs that have been cloned.

GSTM1, GSTM2, GSTM4 and GSTM5 show marked similarity in both nucleotide and amino acid sequences implying a common origin for these three genes and there have been suggestions of recombination and gene conversion occurring between the members of this class (DeJong *et al.* 1991, Taylor *et al.* 1991, Vorachek *et al.* 1991, Board *et al.* 1993 and Chapte7 5).

Reports of the chromosomal locations of the human genes encoding enzymes from other GST classes have been outlined in Chapter 1.7.4 and are not the subject of contention. In contrast, the chromosomal locations of the known human Mu class GST genes have not been clearly resolved and there is contradictory evidence in the published literature. The difficulties arise because of the increasing number of closely related genes in this family which

may cross-hybridise, and the uncertainty as to how many Mu class genes there are in total. A further complication arises as it is clear that the GSTM1 locus has a polymorphic null allele and the frequency of GSTM1 deficiency exceeds 40% in most populations (Board 1981a and b, Board *et al.* 1990a). The null allele appears to be due to the absence of the gene (Seidegård *et al.*1988, Board *et al.* 1990a).

In situ hybridisation data reported by DeJong *et al.* (1988a) utilised a GSTM1 cDNA probe and indicated the presence of the Mu class GSTM1 locus at 1p31. However, the presence or absence of the *GSTM1* gene in the individuals studied was not clear, therefore the relative contribution of the other cross-hybridising Mu class genes was unknown. This doubt has been reinforced, given the high level of sequence similarity now apparent between these cDNAs.

In contrast, Islam *et al.* (1989) studied human/rodent somatic cell hybrids with a rat Mu class cDNA probe and reported a human Mu class locus on chromosome 3. Again in this study, the GSTM1 status of the cell lines involved was unknown. Only one human Mu class locus was identified, as the probe, when hybridised to both human/rat and human/mouse somatic cell hybrids, also identified an extensive complex of both rat and mouse Mu class genes. The human gene identified was contained in a 6.2kb *Bam*HI fragment, however, Seidegård *et al.* (1988) reported that an 11.5kb *Bam*HI fragment is lost in GSTM1-null individuals. Therefore it was inferred that the chromosome 3 locus which hybridised is not *GSTM1* but another Mu class gene.

In a further study DeJong *et al.* (1991), using Southern blots of *Eco*RI digested DNA from mouse/human somatic cell hybrids, suggested that the polymorphic GSTM1 locus could be assigned to chromosome 13, with additional loci on chromosomes 1 and 6. When genomic DNA digested with *Eco*RI is hybridised with a GSTM1 cDNA probe, four fragments can be detected in individuals expressing GSTM1 and an approximately 8kb fragment is missing in individuals known to be of the GSTM1-null phenotype (Seidegård *et al.* 1988,

Board *et al.*1990a). However, the GSTM1 status of the cell lines used was again unclear and a complicating factor. Therefore, in an analysis of the data presented by DeJong *et al.* (1991) it is not possible to distinguish between those cell lines lacking the signal for the ≅8kb fragment because of the null phenotype and those cell lines lacking the signal because the chromosome carrying the GSTM1 locus is one of the missing chromosomes. On this basis it is not possible to assign the GSTM1 locus to a single chromosome as the ≅8kb fragment signal could be associated with any of chromosomes 1, 3, 6, 13, 15, 17, 18 or 19.

The CEPH (Centre d'Etude du Polymorphisme Humain) consortium linkage map of chromosome 1 places the GSTM1 locus in association with two markers, D1S11 and D1S73, reliably ordered in their analysis to a ~50cM region, 1p13.1 to 1p21 (Dracopoli *et al.* 1991). In an attempt to integrate this linkage map and the physical map derived from cytogenetic assignments of chromosome 1, Collins *et al.* (1992) noted the discrepancy in the summary map produced. The only chromosomal localisation of a Mu class gene in the literature with a definite band assignment (1p31) is from the original report by DeJong *et al.* (1988). This assignment places *GSTM1* distal to PGM1 (a confirmed and reliably ordered linkage reference locus) on the cytogenetic map, however, their calculations place it proximal to PGM1 in the resulting genetic map.

Two recent studies were published while the present work was in progress. Zhong *et al.* (1992) used human intron specific sequences derived from the *GSTM1* gene to amplify a predicted 718bp fragment from human/rodent somatic cell hybrid cell lines, suggesting that the human Mu class is located on chromosome 1. Linkage analysis of eight CEPH families was also carried out and Zhong *et al.* (1992) reported a location of 1p13. Consideration of the the evidence indicates that the oligonucleotide primers used for PCR amplification may not have been specific for the *GSTM1* gene. The oligonucleotide primers were shown to be specific for human genomic

DNA when compared to rodent and hamster DNA, however, all the cell lines containing human chromosome 1 were positive despite population studies have indicating clearly approximately 40% of any population have the GSTM1-null phenotype. So it is unlikely that their GSTM1 probe is specific for the *GSTM1* gene. In the second study, Pearson *et al.* (1993) used a combination of locus-specific PCR amplification of each of the five Mu class genes from human/hampster somatic cell hybrid cell lines, isolation of a YAC clone containing all five genes and fluorescence *In situ* hybridisation to indicate that the human Mu clas is clustered with a location at or near 1p13.3.

5.2 AIM

In view of the apparent contradictions in the literature and the complexity of the human Mu class, this study was initiated to examine the extent of hybridisation between the four characterised human Mu class genes, GSTM1, GSTM2, GSTM3 and GSTM4 and, by the use of *in situ* hybridisation, to localise the Mu class GST gene family, in subjects with and without the *GSTM1* gene.

5.3 MATERIALS

Reagents used in the following experiments and materials required for Southern hybridisations were supplied as stated in Chapter 2. The tritiated dinucleotides ((^{3}H)dNTPs) and the kit used for radioactive labelling by nick translation were supplied by Amersham.

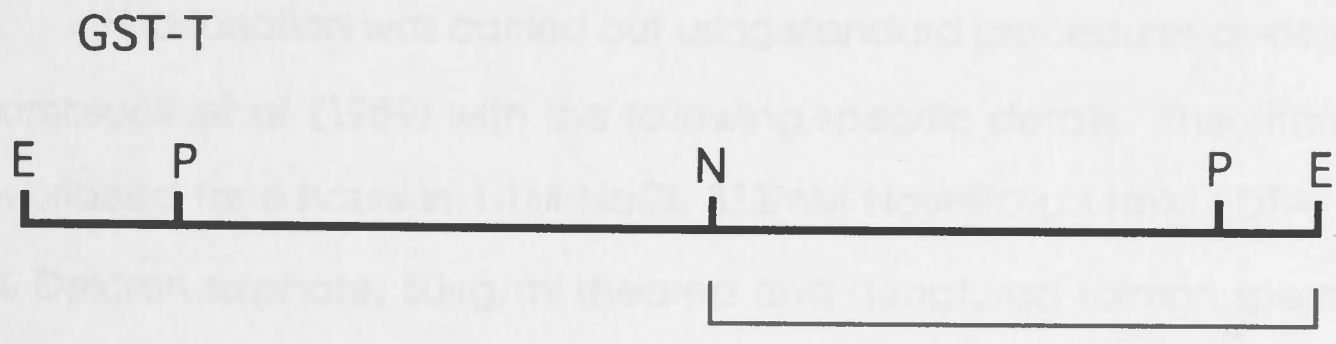
5.4 METHODS

The preparative methods used in this study have been described in Chapter 2.2 and are not repeated. Methods unique to this section are given in detail.

5.4.1 Hybridisation probe

The probe was excised from the GSTM4 encoding plasmid pGST-T by digestion with *Nco*I and *Eco*RI as shown in Figure 5.1. After electrophoretic separation the 666bp DNA fragment was purified using Geneclean II and labelled appropriately as described below. The probe contains 218bp of coding sequence and 448bp of 3′ noncoding sequence.

Figure 5.1 Schematic showing the derivation of the hybridization probe from the GSTM4 encoding plasmid pGST-T. The 3′.end of the cDNA was excised from the plasmid with *Eco*RI and *Nco*I.



E =*Eco* RI, N =*Nco*I, P =*Pst* I.

5.4.2 Capillary transfer and hybridisation of nucleic acids (Southern blot procedure)

cDNAs of four previously described human Mu class genes were prepared from *E. coli* (TG1) cultures containing the plasmids, pKKGST1, pKKGSTM2, pKKGSTM3, and pKKGSTM4 which are described in detail in Chapter 3.4.2. 500ng of each plasmid was linearised by digestion with *Eco*RI and after electrophoretic fractionation on an agarose gel, the DNA was transferred to a nylon membrane (Genescreen plus) using the Southern Blot technique (Southern 1975) as modified by Reed and Mann (1985). Prior to transfer the DNA was depurinated by soaking the gel in 0·25M HCl, then rinsed in distilled H_20. Transfer to the membrane was completed in 0·4M NaOH overnight.

Genomic DNA was extracted from samples by the methods of Grunebaum *et al.* (1984). Following digestion of the DNA with a range of

enzymes (*Eco*RI, *Bcl*I, *Msp*I and *Taq* I) transfer to a nylon membrane was carried out as previously described.

The filters were then hybridised with a radioactive probe labelled with (α^{32}P)dATP by the random primer method of Feinberg and Vogelstein (1983a and b) using an Amersham kit. For hybridisation with both the cDNA and genomic filters, 200ng of DNA per filter was labelled using (α^{32}P)dATP (>3000μCi/mmol, ≅10μCi/μl); the extension was carried out at 37°C for 3 minutes and the termination reaction for 5 minutes. The labelled DNA was ethanol precipitated along with 1μg of sheared salmon sperm carrier DNA and resuspended in 50μl of H_2O.

Hybridisation was carried out using standard procedures as described by Sambrook *et al.* (1989) with the following specific details. The filters were prehybridised for 6 hours in 1·1M NaCl, 333mM Na_2HPO_4, 11mM EDTA, pH6·2; 18·5% Dextran sulphate; 50μg/ml sheared and denatured salmon sperm DNA at 65°C. Hybridisations were carried out in 15ml of the above solution containing the labelled probe DNA at 65°C for 16 hours. The filter was then washed in 2XSSC (150mM NaCl; 15mM Na citrate) at room temperature, followed by a wash in 2XSSC containing 0·1% SDS at 65°C for 15 minutes. Filters were exposed at –70°C with an intensifying screen.

5.4.3 *In situ* hybridisation

In situ hybridisation involves hybridisation of intact chromosomes with a radioactively labelled DNA probe. The technique includes removal of non-specifically bound probe and incorporates autoradiographic exposure and development. Finally chromosomes are G-banded to provide reliable identification and to subdivide the chromosomes into small regions for more accuracy in physical location.

In order to locate the Mu class GST gene loci, human chromosomes were hybridised with the probe derived from the GSTM4 clone pGST-T described (Figure 5.1) . This was tritium-labelled by the nick translation method

of Kelly *et al.* (1970) using an Amersham kit and following the manufacturer's protocol. Three tritium labelled deoxynucleotides were used, (^{3}H)dATP, (^{3}H)dCTP and (^{3}H)dTTP and specific activities of 1·1 - 2·5 X 10^8CPM/µg were achieved. A selection of slides was acetylated by the method of Pardue (1985) in an attempt to prevent nonspecific hybridisations due to charge effects, however, no difference in the background level of hybridisation was observed. The probe was used at a concentration of 200ng/ml and the chromosome slide preparations were denatured, probed and stringency rinsed according to Donlon (1986). Slides were dipped in Ilford L-4 emulsion and exposed for 13-45 days. Preparation of G-banded chromosomes followed the method of Buckle and Craig (1986) and were prepared from the peripheral blood of four GSTM1-plus and four GSTM1-null individuals. However, only a single individual of each group was found to have both sufficient cell divisions present and to have chromosomes G-banded efficiently to be included in the study.

In situ hybridisation on a further three GSTM1-plus and three GSTM1-null individuals was carried out by Dr. G. C. Webb (Genetics Department Queen Elizabeth Hospital, Adelaide, South Australia) in whose laboratory all the *in situ* hybridisation experiments were carried out.

Silver grains on, or within one-half-chromatid width of the chromosomes were plotted onto a 550-band ideogram of all the human chromosomes (Harden and Klinger 1985). The initial scoring revealed peaks of grains over the proximal region of 1p (Figure 5.4), so chromosome 1 was investigated in greater detail using only prophasic chromosomes which exhibit extended bands and allow more accurate detailed mapping. The results were recorded on an ideogram of chromosome 1 increased 1·5 times in size (Figure 5.5). The diagrams of silver grain distribution were prepared on CHROMOMAP version 3·0 (unpublished program, Dr. P.G Board).

Sample chromosomal spreads were obtained from eight normal individuals in total, of both sexes including four GSTM1-plus and four GSTM1-null

individuals. The presence or absence of the *GSTM1* gene was ascertained in each individual by gene specific polymerase chain reaction amplification, essentially as described by Comstock *et al.* (1990) and the details are shown in Table 5.1.

Table 5.1 Sex and GSTM1 status of the subjects scored by *in situ* hybridisation for the localisation of the Mu class gene family.

subject	sex	GSTM1 status
AC	M	—
HAN	F	—
MN	M	—
PF	M	—
GP	M	+
GW	M	+
LK	F	+
TM	F	+

5.5 RESULTS

5.5.1 Southern blot analysis

The nucleotide sequence similarity between GSTM4, and GSTM1, GSTM2 and GSTM3 suggests there would be significant cross-hybridisation, even at relatively high stringencies. To confirm that this was indeed the case, the GSTM4 probe was hybridised to southern blots of the four cDNAs (Figure 5.2) and to genomic DNA of two individuals known to be GSTM1-plus or GSTM1-null (Figure 5.3). The hybridisation of the GSTM4 probe with human genomic DNA confirms that this probe is capable of detecting the GSTM1 gene. GSTM1-null individuals in Figure 5.3 are clearly discernable due to the absence of the 8kb *Eco*RI fragment, as is hybridisation with other bands representing a number of other Mu class GST genes.

The GSTM4 probe hybridised strongly to the cDNAs of GSTM1, GSTM2 and GSTM4, the signal was weaker, though clearly detectable, with the GSTM3

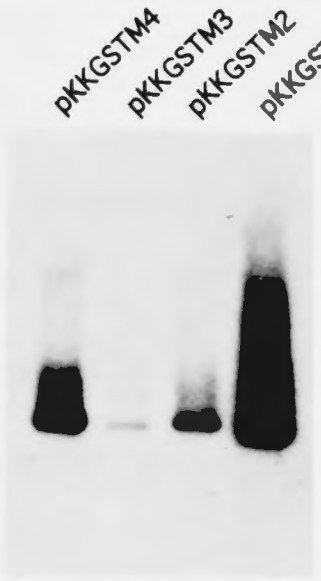


Figure 5.2 Southern blot of human Mu class GST cDNAs probed with a 3′ hybridisation probe derived from the GSTM4 encoding plasmid pGST-T.

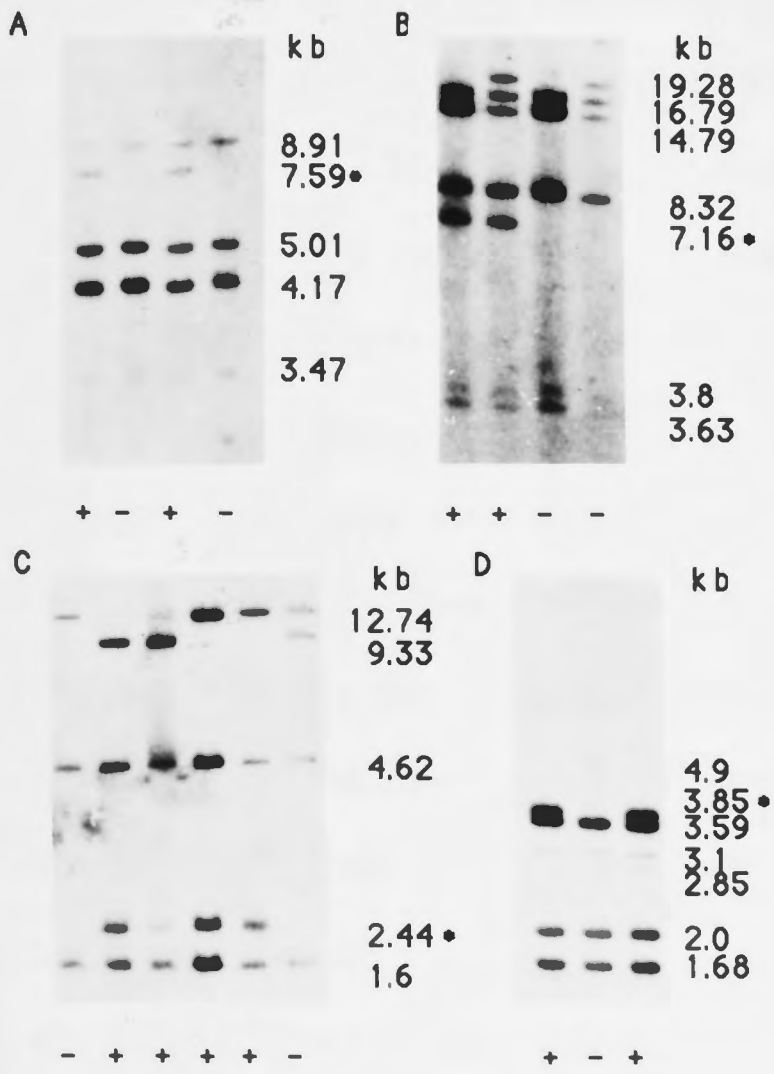


Figure 5.3 Southern blot of human genomic DNA digested with various restriction endonucleases (A: *Eco*RI, B: *Bcl*I, C: *Msp*I, D: *Taq*I). These were hybridised with the same 3′ probe derived from the GSTM4 encoding plasmid pGST-T as used for the cDNAs in Figure 5.2. The + and — symbols indicate the presence or absence of GSTM1 expression.. The approximate sizes of the hybridising fragments are indicated in kb. The bands marked (*) are deleted in individuals with the GSTM1-null phenotype. Other RFLPs are detectable with *Bcl*I and *Msp*I but are not correlated with the deficiency.

cDNA. The coding region of GSTM3 has 69% sequence similarity with GSTM4, compared to GSTM1 and GSTM2 which share greater than 88% similarity with GSTM4 in the coding region (using distances corrected according to the 2-parameter method of Kimura (1980), see Chapter 6.3.4 and Table 6.6. In addition, the probe used contained a significant proportion of 3′ noncoding which has a lower level of sequence similarity. Therefore identical levels of hybridisation would not be expected with this cDNA probe.

5.5.2 *In situ* hybridisation

The GSTM4 probe was generally predisposed to a relatively high frequency of background, even at the shortest exposures (Figure 5.4). Even so, only two significant concentrations of grains can be observed. A very low peak of grains over band 6q21 was repeatably observed in all eight subjects studied representing individuals of both GSTM1-plus and GSTM1-null genotypes (Figures 5.4a and b). The major peak, in all individuals, of both GSTM1 types, was an accumulation of grains over the region 1cen.–1p21, which in well-scored cases, contained 23-25% of all grains (Figure 5.5). The tallest peak in the target region was always over the sub-band 1p13·3 (Figure 5.6). The results shown are from males, however, males and females were represented in both GSTM1-plus and GSTM1-null samples in this study and no differences in the hybridisation pattern were observed.

Detailed scoring of chromosome 1 using prophasic chromosomes from approximately 100 cells each of GSTM1-plus (Figure 5.5a) and GSTM1-null (Figure 5.5b) individuals respectively, showed 75% and 77% of grains over the 1cen.–1p21 target region. The two tallest peaks of grains were over sub-band 1p13·3 and contained 64-65% of the grains over the target region.

From the above results it is concluded that the GST Mu class family of genes is probably in the region 1cen.–1p21 with possible point localisation in sub-band 1p13·3.

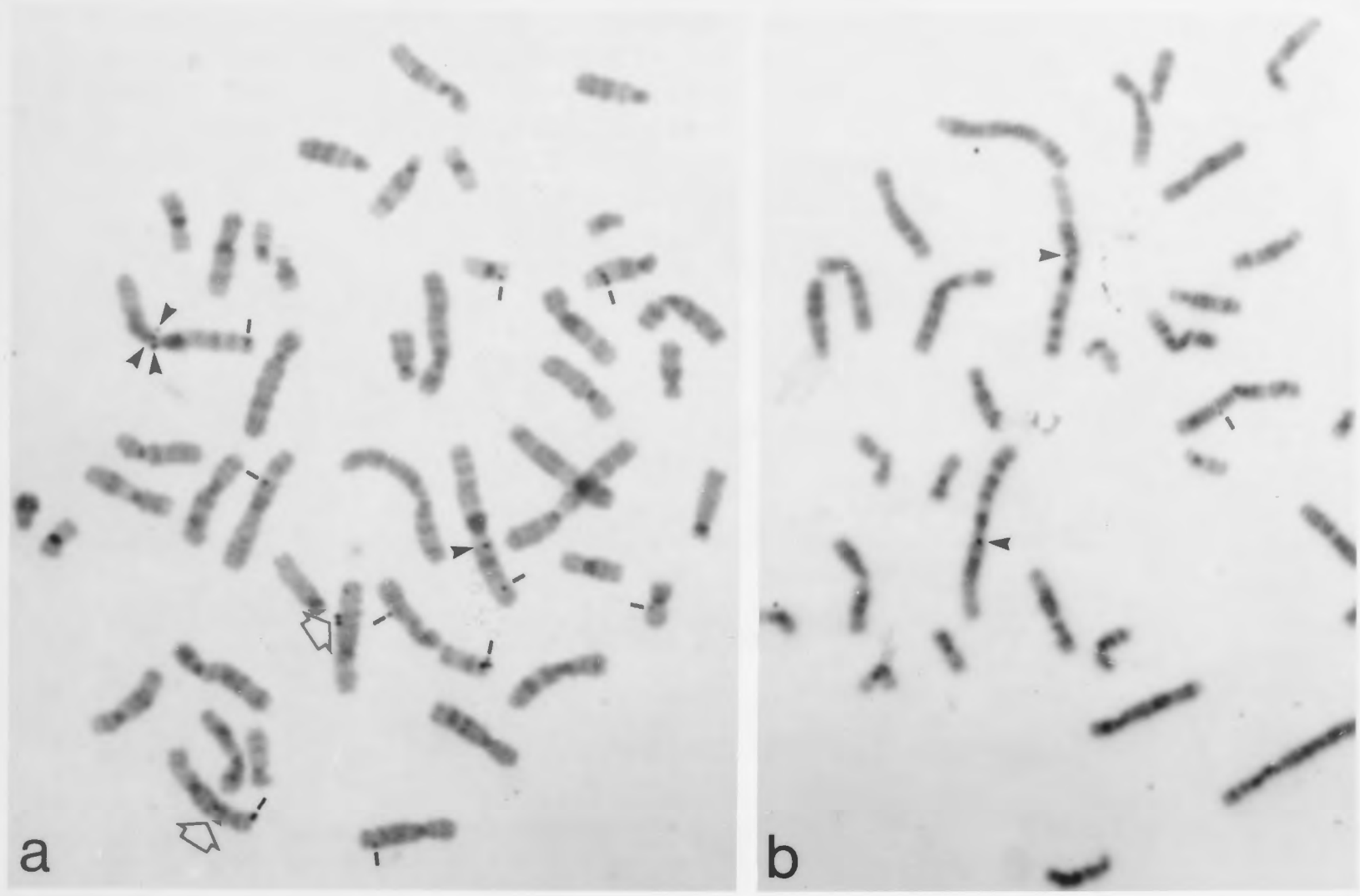


Figure 5.4 G-banded chromosome spreads probed with a H^3-labelled 3′ probe derived from the GSTM4 encoding plasmid pGST-T.

(a) Metaphase from a *GSTM1*-null individual and used to compile Figures 5.5a and 5.6a. Grains over the target region are indicated by solid arrows. The open arrow indicates grains associated with the small peak on chromosome 6. Background grains, indicated by dashes, are characteristically frequent.

(b) Chromosomes in late prophase from the *GSTM1*-plus individual used to compile Figures 5.5b and 5.6b. The arrows indicate grains over 1p13·3. Background grains, unusually low in this case, are indicated by dashes.

Figure 5.5 Plot of grains over all the chromosomes in approximately 100 cells probed with H3-labelled GSTM4 3′ probe described in Figure 5.1. The low peak in the background grains over 6q21 is seen to be repeatable. In both the GSTM1-null and GSTM1-plus individuals the peak over the target region, 1cen.-1p21 contains 23-25% of all grains and is regarded as the probable location of the GST Mu class family of genes. The tallest column of grains is over sub-band 1p13·3, which is regarded as the possible point locality of the GST Mu class family. (a) GSTM1-null individual. (b) GSTM1-plus individual.

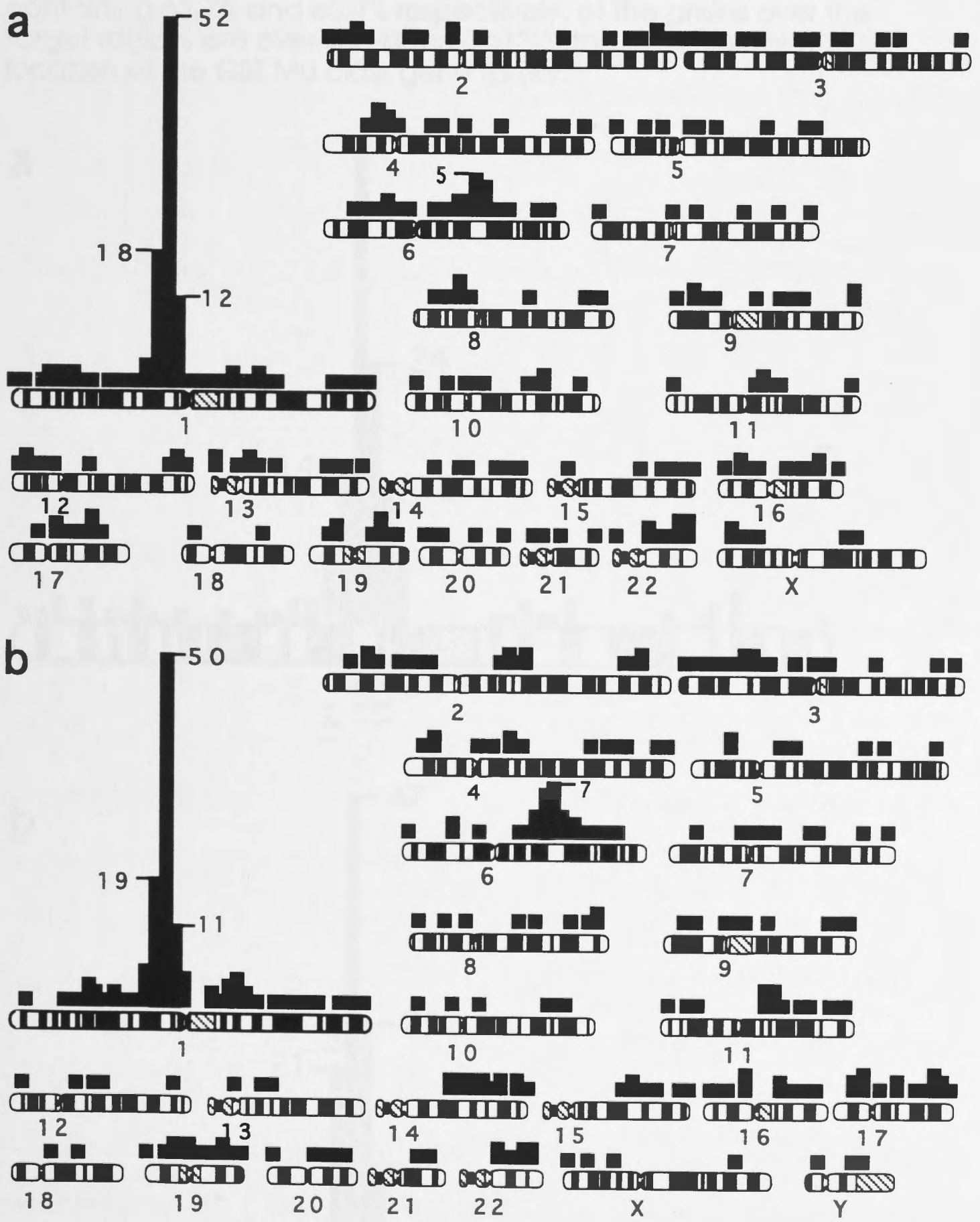
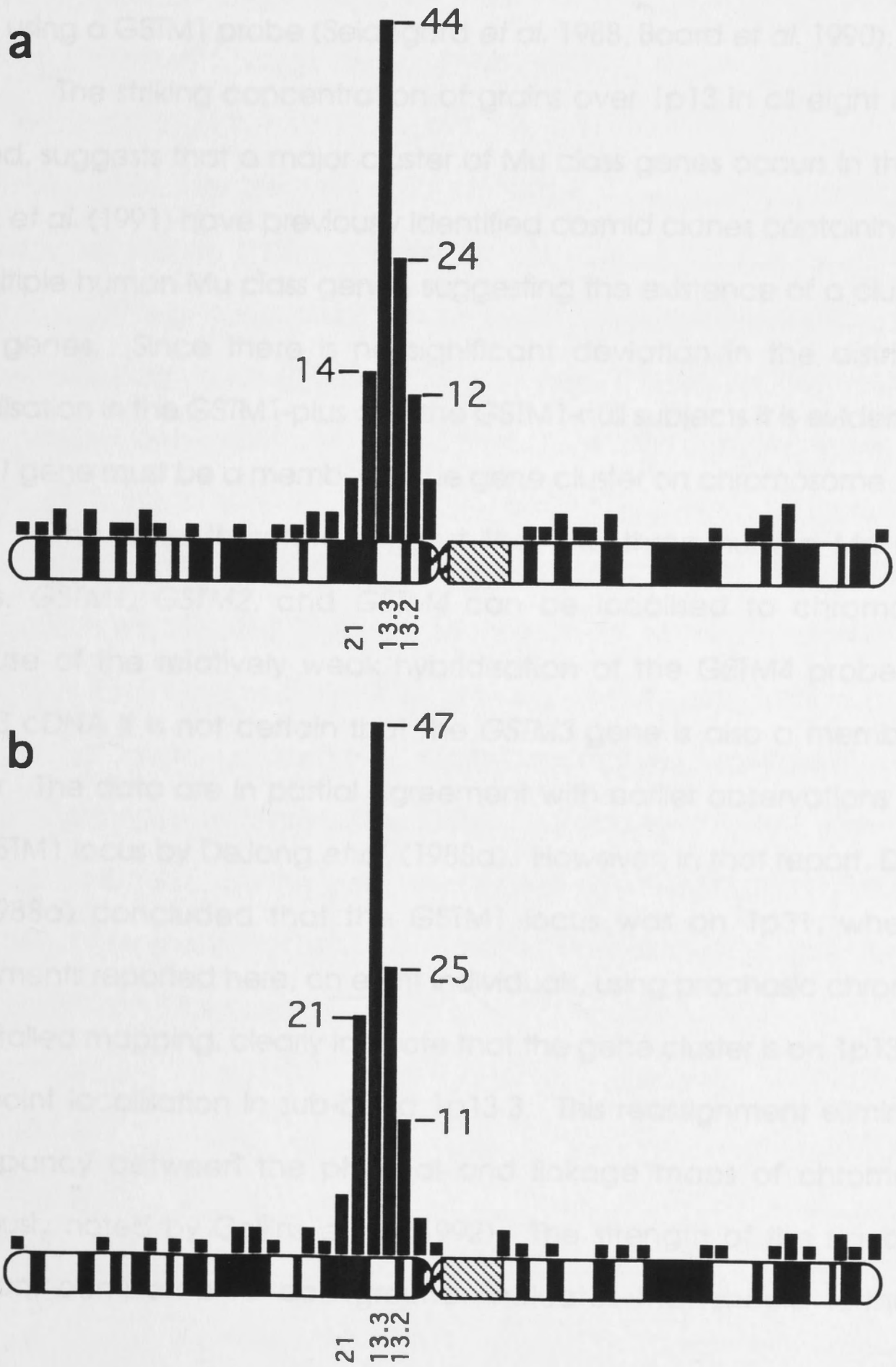


Figure 5.6 Plots of grains over prophasic chromosomes from a GSTM1-plus individual (a) and a GSTM1-null individual (b), probed with H_3-labelled 3′ probe derived from the GSTM4 encoding plasmid pGT-T 76·8% and 74·8% respectively, of the grains are over the target region. The two tallest columns, containing 63.7% and 65.4% respectively, of the grains over the target region, are over sub-band 1p13·3, the possible point location of the GST Mu class gene family.



5.6 DISCUSSION

The Southern blot analysis shows that the *in situ* hybridisation probe used cross-hybridises strongly with the coding region of GSTM1 and GSTM2 cDNAs and weakly with a GSTM3 cDNA, indicating that it should be capable of localising the corresponding genes. It has been demonstrated that the GSTM4 probe (pGST-T) can recognise the ≅8kb *Eco*R1 fragment containing the *GSTM1* gene and also generates the same hybridisation pattern as that observed when using a GSTM1 probe (Seidegård *et al.* 1988, Board *et al.* 1990).

The striking concentration of grains over 1p13 in all eight individuals studied, suggests that a major cluster of Mu class genes occurs in that region. Taylor *et al.* (1991) have previously identified cosmid clones containing portions of multiple human Mu class genes, suggesting the existence of a cluster of Mu class genes. Since there is no significant deviation in the distribution of hybridisation in the GSTM1-plus and the GSTM1-null subjects it is evident that the *GSTM1* gene must be a member of the gene cluster on chromosome 1.

The data therefore suggest that the three human Mu class GST genes, *GSTM1, GSTM2,* and *GSTM4* can be localised to chromosome 1. Because of the relatively weak hybridisation of the GSTM4 probe with the GSTM3 cDNA it is not certain that the *GSTM3* gene is also a member of the cluster. The data are in partial agreement with earlier observations localising the GSTM1 locus by DeJong *et al.* (1988a). However, in that report, DeJong *et al.* (1988a) concluded that the GSTM1 locus was on 1p31, whereas the experiments reported here, on eight individuals, using prophasic chromosomes for detailed mapping, clearly indicate that the gene cluster is on 1p13, possibly with point localisation in sub-band 1p13·3. This reassignment eliminates the discrepancy between the physical and linkage maps of chromosome 1 previously noted by Collins *et al.* (1992). The strength of the signal at that location, compared to background, indicates that should further cross-

hybridising Mu class GST genes be characterised it would be probable that they will also be included in the cluster. The reassignment reported here is also in agreement with the localisation reported recently by Zhong *et al.* (1992) which suggested the human Mu class *GSTM1* localised to chromosome 1 p13.

In a previous study, using somatic cell hybrids, DeJong *et al.* (1991) suggested the presence of the *GSTM1* gene on chromosome 13. The present studies of individuals with and without this gene did not identify any specific hybridisation on that chromosome. Nor did the *in situ* hybridisation data detect the locus reported on chromosome 3 by Islam *et al.* (1989). However, weak hybridisation at 6q21 was repeatably detected. The report of DeJong *et al.* (1991) also indicated the presence of a Mu class gene on chromosome 6 and it will be of interest to determine if this locus is a functional gene, a diverged pseudogene or is perhaps more distantly related.

Unambiguous localisation by *in situ* hybridisation of each of the Mu class genes would require gene specific probes. The high degree of similarity in nucleotide sequences of the cDNAs that have been characterised suggests the use of intronic sequence may be necessary. However, this approach may also have some difficulties as the partial genomic clones described by Taylor *et al.* (1991) also show 87% sequence similarity in introns 2, 3 and 4, between *GSTM4* and the gene *mu3* (GSTM2), although it not clear that this continues in other introns. Intronic sequence information is not yet available for *GSTM1* but considering that, in intron 3 at least, the sequence similarity between the rat gene (Lai *et al.* 1988) on the one hand and *mu2* (GSTM4) and *mu3*(GSTM2) on the other, is of the order of 61%, there are indications of a considerable degree of conservation in this class extending beyond coding exons.

5.7 SUMMARY

The chromosomal localisation of the human Mu class glutathione *S*-transferase genes has been complicated by two factors, the total number of

genes is unknown and there is a polymorphism which results from the presence or absence of the *GSTM1* gene. Four human Mu class GST isoenzymes have been characterised by others, GSTM1, GSTM2, GSTM3 and GSTM5 and the cloning and characterisation of a fifth, GSTM4, another member of this class is reported in Chapter 3. A probe derived from GSTM4 hybridises to genomic DNA, generating the same hybridisation pattern as that observed when using a GSTM1 probe. The same GSTM4 probe cross-hybridises with three of the other human Mu class GSTs cDNAs. *In situ* hybridisation with the GSTM4 probe localised a major region of hybridisation on chromosome band 1p13. These data indicate that the human Mu class gene family is largely clustered and not dispersed on different chromosomes. The identical *In situ* hybridisation patterns in individuals with or without the *GSTM1* gene suggests that this locus is a component of the Mu class GST gene cluster.

CHAPTER 6 AN EVOLUTIONARY PERSPECTIVE OF THE GLUTATHIONE *S*-TRANSFERASES

6.1 INTRODUCTION

6.1.1 Molecular evolution

Nucleotide and amino acid sequences not only provide an ever increasing wealth of information concerning present biological function but also represent a very extensive source of knowledge about the past, often paralleling the fossil record and providing remnants of patterns from very early forms of life. Until twenty years ago most studies in systematics utilised morphological characters as evidence for relationships, however, the contribution of molecular approaches to phylogenetic research has increased dramatically. An advantage of molecular phylogenetics is that it is possible to trace the early evolution of life. It is not possible to estimate genetic distances among widely divergent species, such as plants, animals and bacteria, by any morphometric measure, but this is possible using either nucleotide or amino acid sequence data.

A discussion of evolution in the broad sense is well beyond the scope of this thesis. There are, however, some pertinent points of a more general nature that need to be made before limiting this discussion. Firstly, the conceptual basis to the methods used in estimating evolutionary relationships is still in a relatively early stage of development. Simplifications and assumptions are needed to allow the limited computational ability available, to cope with the immense complexity of a biological system.

Phylogenetic inferences are based on the premise of inheritance of ancestral characteristics and on an evolutionary history which can be defined by changes in these. Inferring phylogenetic relationships requires an estimation of evolutionary history from necessarily incomplete information. In the context of molecular systematics we do not have direct genetic

information from the past (although in a limited way this is changing), we have access only to contemporary species and molecules.

Secondly, because it may be possible to postulate evolutionary scenarios to produce almost any chosen phylogeny from the observed data, some basis for selecting one or more preferred trees from among the set of possible phylogenies must be used. This is accomplished in two ways; by defining a specific sequence of steps (algorithm) for constructing the best tree and by defining a criterion for comparing alternative phylogenies and deciding which is better, or that they are equally good (Felsenstein 1982, 1988).

Thirdly, it should be stressed that in the topology of phylogenetic trees based on molecular sequence data there is an absolute dependence on the alignment of the sequences, and any difficulties in alignment will cause ambiguity in the resulting trees, lowering the level of confidence. As genes accumulate mutations there is an ever increasing likelihood that some of the changes will be at the same sequence positions. Subsequent substitutions cannot decrease similarity but can increase it (by reversion, parallelism, ect.). Thus similarity does not decline at a uniform rate but rapidly at first, then more slowly.

One of the assumptions required is that patterns of similarity of characters reflect homology, i.e. the similarity results from common ancestry (Hillis and Mortiz 1990). Analysis of molecular data can be made more complex by the need to assume positional homology. This means that the nucleotides/amino acids under analysis at a given position should all trace their ancestry to a single position that occurred in a common ancestor. Insertion and deletion events must nearly always be postulated to conform to this assumption.

Finally, it must be understood that the proposed evolution of a single enzyme group, however large, cannot be used in isolation to infer the general path of evolution. The restraints and/or selective forces for change on the members of a single enzyme group may not be consistent with the

forces acting on an organism as a whole. However, both the use of single gene families (eg. rRNA) and the combination of data from several genes and families of genes has led to major insights into significant questions in evolution.

6.1.2 Multigene families

Multigene families are defined on the basis of sequence homology and related overlapping functions (Hood *et al.* 1975). Gene families cross species lines and there may be more sequence similarity between the homologous gene in different species (orthologues) than between different but related members of a gene family in the same species (paralogues). Supergene families are defined as broader groups of genes containing one or more domains of common origin (Dayhoff 1978, Doolittle 1981). However, care should be taken for in some cases the extent of sequence similarity between proteins can be misleading with regard to catalytic function (Umbenhauer *et al.* 1987) and certainly the reverse is true, where similar catalytic function can be performed by proteins with little apparent sequence similarity.

Gene duplication and the subsequent process of divergence of genetic loci have become accepted as significant mechanisms for the evolution of genes encoding novel functions (MacIntyre 1976, Ohno 1970, Dayhoff 1978). Integral to this proposition is the view that a duplication event may free one of the loci from the constraints of natural selection (Ohno 1970, Kimura 1983). The supernumerary is thus free to accumulate mutations that would have previously been deleterious to the original function of the gene, eventually allowing it to emerge as a new gene. Meanwhile the original function is being continued by the other copy of the gene. In eukaryotes, the existence of multigene families as clusters of related genes in close proximity on a chromosome, provides some corroborative evidence of gene duplication and subsequent divergence.

Multi-function isoenzymes may represent an early step in this pathway (Ohno 1970) or, alternatively, it may be necessary for duplicate genes to pass through a non-functional stage to allow the rapid accumulation of mutations (Koch 1972). Many eukaryote genomes do have a pool of non-functional duplicated genes, such as the globin pseudogenes (Poudfoot and Maniatis 1980). Pseudogenes that could be included in this category are also known in the human Alpha class GSTs (Suzuki *et al.* 1993) and a rat Pi class gene (Okuda *et al.* 1987).

Once duplication has occurred, unequal crossover events may also contribute to increases in the number of genes. The resultant genes can be expected to further diverge by mutation and may be the subjects of selection processes. Additionally, gene conversion has been recognised as a significant process that may sometimes lead to divergence but often tends to "homogenise" genes in a multigene locus.

Molecular phylogeny is based on the comparison of sequences. Sequence differences that result from substitutions are the product of a complicated process of mutation, selection and genetic drift. Each of these factors will have a different relative importance in different multigene families. However, it has been widely accepted that most of the divergence observable among related genes is largely neutral to selection (Kimura 1983). Random mutation is the mechanism for accumulation of these neutral substitutions and selection is thought to be minimal, because such mutations have little or no effect on function. So, for example, mutations may be maintained at a different frequency in noncoding regions compared to coding regions and also in the largely redundant third position of a codon compared to the other two, positions (Miyata *et al.* 1980, Kreitman 1983, Kimura 1983).

6.1.3 The glutathione *S*-transferases

GSH was once thought to be essential for life because of its widespread taxonomic distribution, despite the specific function being

unknown. Lack of evidence for the presence of GSH in many prokaryotes, including some strict aerobes, has modified that view (Fahey *et al.* 1978). GSH is now considered to have a general function in the protection of cells against oxygen toxicity (Mannervik 1985, Mannervik and Danielson 1988). There are several systems of enzymes that facilitate this protection by GSH, such as GSH reductases, peroxidases and thioltransferases. The GSTs are at present the most extensively studied family of enzymes involved in GSH metabolism.

An understanding of the origin and evolution of what is now considered the GST supergene family is likely to develop quickly in the next few years since sequence data are rapidly appearing. The complete coding regions of 72 sequences are known from GSTs in various species and are included in the present study. In addition, GST activity has been characterised in many more species (Clark 1989) and even in the better characterised mammalian species there may be still more genes to be cloned (Suzuki *et al.* 1987, Tsuchida *et al.* 1990). In some studies the level of activity reported may be somewhat misleading because such surveys have invariably been carried out using CDNB as the substrate and it has become clear that there are a significant number of GSTs that exhibit no, or very low, activity with this substrate (Chapter 1.4.3).

The general characteristics and functions of this enzyme family are described in Chapter 1. However, there are a few salient points worth reiterating. GSTs exist in higher organisms as families of enzymes each of which has a broad but overlapping substrate specificity, they have been divided into cytosolic and microsomal forms. The cytosolic forms occur as both homodimers and heterodimers of individual subunits. The cytosolic enzymes have been assigned to four separate classes Alpha, Mu, Pi and Theta, on the basis of biochemical, immunological and sequence characteristics. Within the mammalian species studied, structural similarity among genes in each of the classes has been observed (Chapter 1.5). Additionally, definition of mammalian GST classes includes amino acid

sequence similarities of at least 50% and as much as 95%. However, between classes the similarity is thought to be only of the order of 30% (Mannervik and Danielson 1988, Mannervik *et al.* 1992). When comparing sequences from more diverse species this level of similarity can further decrease.

The earliest published GST sequences from non-mammalian species were from parasitic Schistosomes (Smith *et al.* 1986, Henkle *et al.* 1989). Comparison with mammalian sequences indicated that the parasite sequences from the 26kDa proteins could be considered as Mu class whereas the affiliation of the 28kDa GST sequence were more Alpha-like but less certain.

There is only a single form of microsomal GST known from each of rats, and humans. In a pairwise comparison these exhibit 83% similarity. However, the level of similarity, when compared with the cytosolic GSTs, was seen as sufficiently low to have prompted speculation of either very early divergence or even independent evolution (Mannervik and Danielson 1988, Morgenstern and DePierre (1989). Currently there is no evidence to suggest that the microsomal and cytosolic GSTs share a common ancestor.

It has been suggested that the level of sequence conservation observed between the mammalian members of the newest class of GSTs, Theta, and such diverse species as maize, *Drosophila* and *Methylbacterium* implied the mammalian Theta class may be representative of the ancient progenitor GST gene (Meyer *et al.* 1991, Pemble and Taylor 1992). Very recently, Buetler and Eaton (1993) have also included most of the non-mammalian GST sequences (insect, plant, bacteria and yeast) in the Theta class as a "catch all" of what was admitted to be a group with heterogeneous sequence structure.

On the basis of their recent phylogenetic study, Buetler and Eaton (1993) have proposed a further class, Sigma, containing primarily the cephalopod eye lens crystallin sequences. However, this class has yet to gain general recognition. A discussion of the class classification system and data used to define the relationships is also to be found in Chapter 1.3.

6.1.3.1 Taxonomic distribution

In the face of an ever broadening range of toxic compounds assaulting the cell, detoxication enzymes would find a strategy of simple proliferation insufficient, on its own, to fulfil their broad protective role. Adaptability to new substrates and flexibility in regulation in response to changing environments are also required. In this light it is not surprising that detailed studies of GST activity in many organisms have identified an intricate pattern of expression in development and in differing cell types. So, for example, in addition to the work already discussed on mammalian GSTs (Chapter 1.6.3), studies of such diverse species as the amphibian *Bufo bufo* (Del Boccio *et al.* 1987b) and the rubber tree *Hevea brasiliensis* (Balabaskaran and Muniandy 1984), have shown developmental and tissue specific patterns of expression by a range of GST isoenzymes. Arthropods such as crustaceans (Keeran and Lee 1987) and several insect species appear to have relatively few isoenzymes. However, in insects such as the housefly (*Musca domestica*) and the cockroach (*Periplaneta americana*), which are both scavengers with a varied diet and environment, multiple GSTs have been observed (Clark *et al.* 1977). GSTs in *Drosophila spp.* (vinegar-fly) and *Musca domestica* (housefly) have been linked to insecticide resistance and have been studied in some detail (Shamaan*et al.* 1993, Beall *et al.* 1992, Cochrane *et al.* 1991, Wool *et al.* 1982). As a result, nucleotide sequences for representative corresponding isoenzymes have now been cloned (Cochrane *et al.* 1991, Toung *et al.* 1990). Evidence has been presented which suggests that in *Drosophila,* GSTs are not closely related to each other and it has been suggested that a dispersed gene family analogous to that established in mammals may exist in insects (Toung *et al.* 1990).

In some plants, for instance the rubber tree (Balabaskaran and Muniandy 1984), the pea (*Pisum savitum*; Diesperger and Sanderman 1979), maize (*Zea mays*; Lamoureux and Frear 1979, Moore *et al.* 1986), and tobacco (*Nicotiniana tabacum*; Takahashi and Nagata 1992) in particular,

increased expression of GSTs has been associated with the administration of some herbicides (Mozer *et al.* 1983, Moore *et al.* 1986). Evidence of multiple soluble isoenzymes was obtained from both maize and pea and a microsomal enzyme from the pea has been reported (Diesperger and Sanderman 1979). cDNA transcripts and genes encoding some of the maize enzymes have been cloned (Grove *et al.* 1988, Moore *et al.* 1986). GSTs have also been found in cephalopods, in this case as species specific eye lens crystallins (Tomarev and Zinoviera 1988, Harris *et al.* 1991, Tomarev *et al.* 1991, 1992).

Within the non-vertebrate eukaryotes GST enzymes with activity towards CDNB have been purified from several euglenoid species and found to also have activity with cumene hydroperoxide (Lau *et al.* 1980, Overbaugh and Fall 1982). The native molecular weight was determined to be 37kDa. Similarly a trypanosome species GST enzyme was purified as a heterodimer consisting of 20 and 17kDa subunits (Yawetz and Agosin 1981). Unusually, the GST purified from *Tetrahymena* is reported to be a 37kDa monomer (Overbaugh *et al.* 1988). GST activity has also been found in a dinoflagellate and a green alga (Lau *et al.* 1980). Using CDNB as a substrate, GST activity was found in *Candida* and *Neurospora* species of yeast and a GST-like cDNA has been cloned from *Saccharomyces cerevisiae* (Lau *et al.* 1980, Coschigano and Magasanik 1991). In addition, two immunologically distinct subunits of 22 and 23.5 kDa were purified from the yeast, *Issatchenkis orientalis,* and one has now been cloned (Tamaki *et al.* 1991).

The extent of GST expression in prokaryotes is less well defined. The GSH-dependent dichloromethane halogenase of *Methylobacterium* sp. is considered a GST enzyme (La Roche and Leisinger 1990). Although this enzyme exists as a hexamer, an unusual arrangement in GSTs, the subunits are 37.4 kDa, a similar size to the dimeric GST enzymes purified from some of the non-vertebrate eukaryotes, whereas mammalian GSTs are dimers consisting of 23-28 kDa subunits. Other known prokaryote sequences

considered to have similarity with GSTs, one from *Flavobacterium* and one from *E. coli*, have not been fully characterised in terms of their GST activity.

6.1.3.2 Mu class GSTs

The Mu class GSTs are a focus of this thesis. In this class the number of known enzymes and their sequences has grown dramatically since the initiation of this work, particularly within humans (Chapter 1.7.2). Several Mu class sequences have been available for rats (Lai *et al.* 1986, Lai *et al.* 1988, Abramovitz and Listowsky 1987). This has led to speculation on the relationships among these genes. Lai *et al.* (1986, 1988) observed that the nucleotide differences within the rat Mu class GSTs are not random but concentrated into a few specific regions, whereas the level of sequence similarity in some other areas of the genes, including some introns, is extremely high. A similar pattern of conservation between gene family members has been noted in a class of cytochrome P450 genes (Atchison and Adesnik 1986). In both cases gene conversion, a non-reciprocal recombination process, has been invoked.

Evidence, from *in situ* hybridization studies of the clustering of human Mu class genes at chromosome band 1p13, has been presented (Chapter 5). This supports the suggestion of clustering from the sequencing of a cosmid insert which contained two partial human genes (Taylor *et al.* 1991). Recent confirmation of the proximity of the Mu class genes on chromosome 1 has been provided by Pearson *et al.* (1993) from analysis of a single Yac clone containing five genes. Evidence of a high level of conservation in both intron/exon structure, splice sites and in general sequence similarity among several human Mu class cDNAs and extending to the rat GST subunit 4 gene (Taylor *et al.* 1991). This evidence was used to support suggestions that gene conversion among the human Mu class may have been a causative factor in the high sequence similarity between the human genes (Taylor *et al.* 1991). The regions thought to be involved are surrounded by a Chi-related tetramer which has been associated with gene rearrangements in other studies

(Kenter and Birshtein 1981) and direct repeats which have also been implicated in gene conversions (Stachelek and Liskay 1988). Buetler and Eaton (1993) further suggested that it is the 5′ and 3′ portions of GSTM1 and GSTM2 cDNAs, that have been involved in conversion, leaving a middle section with a higher level of divergence.

From an evolutionary perspective, GSTs constitute a very interesting family of proteins, with their combination of multiple functions, the complexity of their sub-cellar location, tissue and developmental specificities, all of which are coincident with a very broad taxonomic range. The maintenance of GST catalytic functions through evolutionary history supports the view that the role of GST enzymes is fundamental.

6.2 AIMS

In view of the recently available evidence from crystallographic studies that have shown three dimensional structural similarities among representative enzymes from each of the Alpha, Mu and Pi classes (Cowan *et al.* 1989, Reinemer *et al.* 1991, 1992, Ji *et al.* 1992, Sinning *et al.* 1993 and Chapter 1.5.2) and of the increasing number of GST sequences available, this study was initiated to gain a greater understanding of the evolutionary relationships of the GSTs .

The intention was to construct a multiple alignment, incorporating information from the known crystallographic structure, thus providing a comparison of the currently available GST sequences, including GSTM4 (Chapter 3). Subsequent production of a phylogenetic tree of this enzyme family and an evaluation of its reliability is then possible. A further intention of this study was to analyse the intensively studied Alpha, Mu and Pi class GSTs so as to provide an estimation of divergence times among these enzymes.

6.3 MATERIALS AND METHODS

6.3.1 Origin of sequences

Sequences used in this study were extracted from GENBANK and EMBL data bases. Table 6.1 lists the accession numbers, loci and the sequence names used in this study. The sequence for GSTM4 derives from work described in Chapter 3. The nomenclature has been slightly modified for ease and consistency. An attempt has been made to name the majority of sequences according to the system of Mannervik *et al.* (1992). Rat sequences are an exception, as nomenclature systems had been previously established. These are designated by subunit number, as this system is potentially the most compatible. Mouse sequences have been named according to the orthologous rat subunit where this is possible and has been characterised, because that system is already in common use. All other sequences have a species designation and are numbered in order of the publication of the sequence for that species. Thus a wheat (*triticum aestivum*) sequence (Accession N^{o} X56004, Locus TAGSTAGST) has been designated TAGST2.

Coding regions were identified and translated for amino acid sequence comparison.

Table 6.1 Shows GENBANK accession numbers, locus names; species origin of the sequences and the names used in this study.

GENBANK accession	GENBANK locus	tree designation	SPECIES
M98559	FVBDEHALO	FBDH	*Flavobacterium sp.*
X05088	ECSSPB	ECSSPB	*Escherichidia coli* (stringent starvation protein)
M32346	MTBDCMAA	MBDMD	*Methylbacterium sp.*
X13689	CEGST1	CEGSTP1	*Caenorhabditis elegans* (free living nematode)
X14233, S51044	DMGST	DMelGST1	*Drosophila melanogaster*
M97702	DROGLUSTD	DMelGST3	*Drosophila melanogaster*
M95198	DROGTT	DMelGST2	*Drosophila melanogaster*
M77682	FHEGSTD	FHGST1	*Fasciola hepatica* (parasitic nematode)
M83249	MDOGST1A	MDGST1	*Musca domestica* (house fly)
X61302	MDGST1	MDGST2	*Musca domestica*
L23126	LUCGLTR	LCGST1	*Lucillia cuprina* (sheep blow fly)
M36937	OMMMLPA	OSGST2	*Ommastrephes sloani* (squid)
M36938	OMMMLPB	OSGST1	*Ommastrephes sloani*
M74326	OMMCRYSC	OSGST3	*Ommastrephes sloani*
M65184	OCTCRYS1	ODCRY1	*Ommastrephes sloani*
M65185	OCTCRYS2	ODCRY2	*Ommastrephes sloani*
M65186	OCTCRYS3	ODCRY3	*Ommastrephes sloani*
M65187	OCTCRYS4	ODCRY4	*Ommastrephes sloani*

Table 6.1 continued

GENBANK accession	GENBANK locus	tree designation	SPECIES
X65543	OVOCTS1	OVGST1	*Octopus vulgaris*
X65544	OVOCTS2	OVGST2	*Octopus vulgaris*
X75502, X75820	ASGTA1	ASGST1	*Ascaris suum* (parasitic nematode)
M14654	SCMAG	SJGST1	*Schistosoma japonicum*
M73624	SCMGLUSTRA	SMGST1	*Schistosoma mansoni*
M98271	SCMGSTM	SMGST2	*Schistosoma mansoni*
M87799	SCMGSTX	SHGST1	*Schistosoma haemotobium*
M87800	SCMGSTY	SBGST1	*Schistosoma bovis*
L12057	ATHGST1X	ATGST3	*Arabidopsis thaliana*
X68304	ATGLUTRA	ATGST2	*Arabidopsis thaliana*
L07589	ATHGLUGRFS	ATGST1	*Arabidopsis thaliana*
X58390, M64268	DCCARSR8	DCGST	*Dianthus caryophyllus* (carnation)
X06754	MZEGSTI	ZMGST1	*Zea mayes* (maize)
X06755	ZMGST3	ZMGST3	*Zea mayes*
D10524, D90500	TOBPARB	NTGST1	*Nicotinia tabacum* (tobacco)
M84968	SIPGTSTF	SCGST1	*Silene cucubalus*
X56012	TAGSTA1	TAGST1	*Triticum aestivum* (wheat)
X56004	TAGSTAGST	TAGST2	*Triticum aestivum*
M35268	YSCURE2	SacCURE2	*Saccharomyces cerevisiae* (yeast)
X57957	IOGSTY2	IOGST2	*Issatchenkia orientalis* (yeast)
(PIR) S13780	(PIR) S13780	SSGSTP1	*Sus scrofa* (pig)
X61233	BTGST	BTGSTP1	*Bos Taurus* (cattle)
X08020	HSGST4	HSGSTM1a	*Homo sapiens* (human)
X06547	HSGSTPI	HSGSTP1	*Homo sapiens*
M63509	HUMGLUTRA	HSGSTM2	*Homo sapiens*
J03746	HUMGST	HSGSTMIC	*Homo sapiens*
M99421	HUMGSTAA	HSGSTM4	*Homo sapiens*
M99422	HUMGSTAB	HSGSTM4	*Homo sapiens*
M21758	HUMGSTB	HSGSTA1	*Homo sapiens*
		HSGSTM5	*Homo sapiens*
M16594	HUMGSTC	HSGSTA2	*Homo sapiens*
J03817	HUMGSTD	HSGSTM1b	*Homo sapiens*
J05459	HUMGSTMUA	HSGSTM3	*Homo sapiens*
M59772	HAMGST	MAGSTM1	*Mesocricetus auratus* (hamster)
X57489	CLY1	CLGST1	*Cricetulus logicaudatus* (chinese hamster)
X65021, S40516	MMGLUT	MMGST2	*M.musculus* (mouse)
X53451	MMGSTII	MMGST7	*M.musculus* (mouse)
J04632	MUSGLUTA	MMGST3	*M.musculus*
J04696	MUSGLUTB	MMGST4	*M.musculus*
J03958	MUSGSTB	MMGST1	*Mus musculus*
L06047	MUSGTF	MMGST8	*Mus musculus*
S59019	S59019	OCGSTA1	*Oryctolagus cuniculus* (rabbit)
S59023	S59023	OCGSTA2	*Oryctolagus cuniculus*
X67654	RNGLTF	RNGST5	*Rattus norvegicus* (RAT)
J03752	RATGST	RNGSTMIC	*Rattus norvegicus*
K00136	RATGST1YA	RNGST1a	*Rattus norvegicus*
K01931	RATGST2YA	RNGST1b	*Rattus norvegicus*
M28241	RATGSTAA	RNGST3	*Rattus norvegicus*
M14364, J02690	RATGSTPPS	RNGSTPPS	*Rattus norvegicus*
J02744	RATGSTYB	RNGST4	*Rattus norvegicus*
K01932, M10960	RATGSTYC	RNGST2	*Rattus norvegicus*
D10026	RATGSTYRS	RNGST12	*Rattus norvegicus*
X02904	RNGSTP	RNGST7	*Rattus norvegicus*
X62660	RRGTS8	RRGST8	*Rattus rattus*
M38219	CHKCL3	GGGST3	*Gallus gallus* (chicken)
X58248	GGGSTCL2	GGGST2	*Gallus gallus* (chicken)
X63761	PPGLUST	PPGST1	*Pluronectes platessa* (plaice)

6.3.2 Alignment of all GST sequences

Initially similarity between the sequences was obtained using the hierarchical clustering algorithm of Higgins and Sharp (1989) by means of their program CLUSTALV (Higgins 1993). This is based on a modified version of the method of Feng and Doolittle (1987) using alignments of progressively larger and larger groups of sequences according to the branching order provided by an initial phylogenetic tree, or dendrogram, in an iterative method based on the fast, approximate method of Wilbur and Lipman (1983) and the cluster analysis method of Sneath and Sokal (1973). Optimal alignments at all stages are achieved by using the algorithm of Myers and Miller (1988). In amino acid sequence comparisons a Dayhoff (1978) PAM matrix is used.

However, it was found that the CLUSTALV generated amino acid alignment of all sequences failed in some cases to align some of those residues known from the crystallographic data to be important for GSH binding in mammalian species. The solution of the crystallographic structure of several GSTs has provided insights into the structural motifs of the three-dimensional (3D) protein which a CLUSTALV alignment also cannot take into consideration. A minor problem of the computer generated alignments was their inability to adequately deal with sequences of differing lengths. Consequently, it was necessary to substantially edit the amino acid alignment manually and the GENOMIC DATA ENVIRONMENT (GDE) package (version 2.2, Smith 1993) was used. The full alignment of all sequences used is to be found in Appendix 1.

6.3.3 Alignment of individual GST classes

Alignments of the cDNA sequences from the three well characterised mammalian classes, Alpha, Mu and Pi, were achieved using CLUSTALV and the default parameters, the amino acid sequence alignment is shown in Appendix 2, the nucleotide in Appendix 3 For these, little manual editing was required.

6.3.4 Estimation of genetic distance

Evolutionary distance between a pair of sequences is measured by the number of nucleotide or amino acid substitutions between them. The simplest method is where distance is merely the proportion of sites at which two sequences are different. Distances are calculated after multiple alignment rather than alignment of separate pairs of sequences, as the latter may increase the error considerably (Higgins 1991). Simple pairwise distances between amino acid sequences were obtained by CLUSTALV based on the manually edited alignment and using a PAM 250 matrix, these are shown in Table 6.2.

Distances were also calculated for a truncated version of the alignment, removing both the 3′ and 5′ extensions but the estimated distances were similar. This is because any position with a gap for either sequence in a pairwise comparison is excluded and so these extensions were largely excluded from the calculations anyway.

There are several methods available for correcting the estimation of genetic distance to overcome the problems of multiple substitutions at a single site. For protein sequences the distance matrices were calculated under three different models of amino acid replacement whereas for the nucleotide sequences two models of replacement were used.

The three models of amino acid substitution are; one based on PAM amino acid relationship matrices developed by Dayhoff (1978); one developed by Kimura (1980); and one developed by Felsenstein (1993) based on the division of amino acids into groups. Each of these relies on a different weighting system for different types of substitutions. The Kimura model was not applicable to the comparison of all GST sequences because if any two sequences are more than 83% divergent the distance becomes infinite (Kimura 1980). It was possible to use this correction for the Alpha, Mu and Pi class comparisons. Trees based on the PAM matrix, Felsenstein's model and on simple percentage distance were very similar in topology (data not shown), the only differences being slight length

1 All numerical methods for inferring phylogenies result in a single, point, estimation of relationships, using some form of probabilistic model as a basis. These two methods were chosen for two reasons, firstly because of their independence, being based on differing algorithms and secondly, because of their widespread use in the field allowing a more general appraisal of the study.

differences in some of the branches. Kimura's 2-parameter method of correction was used in the comparison of the nucleotide sequences within each of the Alpha, Mu and Pi classes.

6.3.5 Inferring phylogenetic relationships

An estimate of evolutionary relationships of all currently available, complete, Mu class GST amino acid sequences was determined by distance, and maximum parsimony methods. Two packages were used to generate the phylogenetic trees, CLUSTALV (Higgins and Sharp 1989) and PROTPARS in PHYLIP (version 3.4, Felsinstein 1992).[1]

To estimate the phyletic relationships among the GSTs from nucleotide or amino acid distances, matrices were computed and unrooted phylogenetic trees were constructed by the neighbor-joining (NJ) method (Saitou and Nei 1987).

The NJ method of tree construction is conceptually and computationally simple, thus requiring far less computer time or power and increasing the number of sequences possible for inclusion in a tree. NJ trees were constructed in CLUSTALV and in PHYLIP. It is possible to produce trees by two different routes in PHYLIP, one provides a single tree similar to that produced in CLUSTALV. The other route involves generating bootstrap (see section 6.3.6 below) replicates of the alignment (generally one hundred) in SEQBOOT, followed by calculation of a distance matrix for the replicates, construction of each tree and, finally, production of a consensus tree in CONSENSE.

Phylogenies using an unrooted parsimony method were also generated using the programs in the PHYLIP package. Parsimony methods use the sequences directly by discrete character comparison. A MP tree derived from the alignment of all the sequences was constructed for comparison with the NJ tree.

For the class comparisons trees were constructed both directly through the maximum parsimony program PROTPARS, and by both routes in

PHYLIP. For the cDNA sequence comparisons of the Alpha, Mu and Pi classes two other tree construction methods were used. The first, DNAPENNY, finds all the most parsimonious trees implied by a data set using a branch-and-bound algorithm for the search which requires prohibitively high computational time in most instances over 10-11 sequences (Hendy and Penny 1982, Felsenstein 1993). Thus trees constructed for the Alpha (12 sequences) and Pi (6 sequences) classes were the most parsimonious found by that program, however, the tree constructed for the Mu class (15 sequences) was that found in 2.5 days of computation but was not necessarily the most parsimonious possible.

The third method used to infer evolutionary relationships within GST classes was maximum likelihood, which again is computationally prohibitive in cases, such as the GSTs, where there is a large number of sequences. This method incorporates a complex algorithm in which it is assumed that each site in a sequence evolves independently, at a transition/transversion rate that can be specified and different lineages may evolve at different rates (Felsenstein 1993). The topologies of the trees constructed from these two methods were similar to those found using the NJ method (data not shown).

The schematics of the phylogenetic trees were produced using TREETOOL (version 1.0, Maciukenas 1993) incorporated in the GDE package (version 2.2, Smith 1993).

6.3.6 Estimation of confidence levels

Phylogenetic trees were tested for reliability by the bootstrap method, a technique that can be used for placing confidence limits on statistics estimated without knowledge of the underlying distribution (Felsenstein 1985). Although it is not possible to prove that one tree is superior to another when the true phylogeny is unknown, it is possible to provide a measure of confidence in any given tree by subjecting it to a bootstrap analysis. The method takes random samples of data, with replacement, from the complete data set and a new data set is generated. Trees inferred

from these replicates can be compared to the original tree. The number of replicates that generate alternative trees provides a measure of the variation and can be used to calculate confidence intervals. These give an indication of how much support there is for the original tree. Bootstrap scores for the NJ tree of all sequences (uncorrected) are shown in Table 6.4. Phylogenetic trees of each of the Alpha, Mu and Pi classes, corrected for multiple substitution (section 6.3.4) were all subjected to boostrapping and are incorporated on the respective trees Figures 6.2 and 6.3 .

6.3.7 Estimation of evolutionary time

It is possible to use genetic distance data to estimate the span of evolutionary time between two species but all methods depend on the accuracy of calculated distances. There are several methods for estimating distances depending on the pattern of substitution (section 6.3.4). Corrections for multiple substitutions at a single site vary primarily on the assumptions related to the ease with which each type of substitution can occur. Genetic distances are generally calculated assuming a constant rate of substitutions at each site, for example Kimura's 2-parameter method, described above. Although it assumes the rate at each site is the same, it differentiates between transition and transversion substitution changes, weighting the latter as less likely. However, analyses have suggested that the rate of substitution varies from site to site, approximately according to a gamma distribution pattern (Tamura and Nei 1993). Thus, for the estimation of divergence times, distances were estimated using both a gamma distribution for the rate of substitutions at each site and a constant rate using Kimura's method (1980).

Evolutionary time can be estimated either as an absolute, in years, or in terms of the relative rates of mutation. Absolute methods rely on calibration from the fossil record, whereas relative rates do not. It is necessary to choose a suitable date for the calibration of absolute estimates. For the purpose at hand the bird/mammal split was considered adequate, as both Alpha and Pi classes contain sequences from these groups. In the fossil

1 The conclusions based on this premise could be in error if the bird and mammalian sequences are not orthologous or the fossil dating is found to be miscalculated. There is, however, a reasonable likelihood that the bird and mammalian GST sequences are orthologous, given that the calculated distances of both chicken sequences fall within the range of the mammalian differences. The miscalculation of the fossil data is always a possibility and may affect these calculations.

record dating of the divergence of mammalian orders is not entirely clear and the record is poor, however, the separation of the mammal-like reptiles from those leading to the bird lineage is less controversial, and considered to be at about 300 Ma (Mega Anni, millions of years) ago, during the Carboniferous era (Pennsylvanian era, in North America, Carrol 1988). [1]

A simple calculation is possible to estimate the time of divergence by using the genetic distance between a chicken sequence and a mammalian sequence and the formula,

$$D = X/G \times A$$

where A = millions of years since bird/mammal split,
G = genetic distance between chicken and mammals,
X = genetic distance between subject sequences
and D = Time since divergence of subject species

6.3.8 Relative rates

By calculating time from genetic distance in the above way, an assumption of similar rates of mutation in all lineages is made. Formal relative rate tests were performed by the method of Wu and Li (1985) to test the validity of this assumption in the case of the Alpha and Mu class GSTs. Relative rates of divergence between two sequences can be tested with the aid of a third, outgroup sequence, simply by calculating the total distance between the outgroup and each of the two sequences, less the distance estimated between the subject species. Independent verification of the accuracy of the estimations may be achieved using the fossil record. The coding region nucleotide sequence alignment used for the relative rate tests was the same as that used for the phylogenetic comparisons within the classes (section 6.5.3). The non-degenerate (K(0)) and four-fold degenerate (K(4)) substitution rates for representative species comparisons were calculated using the methods of Li *et al.* (1985) and Li (1993).

Table 6.2 Distance matrix of all pairwise comparisons of GSTs. Column 1 and row 73 orders sequences for ease of comparison. Matrix generated by CLUSTALV

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
1 MMGST7	0
2 RNGST7	8	0
3 HSGSTP1	15	14	0
4 BTGSTP1	14	14	14	0
5 SSGSTP1	15	15	17	13	0
6 CEGSTP1	58	60	60	60	59	0
7 MAGSTM1	69	70	71	70	67	68	0
8 HSGSTM4	71	70	71	71	68	68	20	0
9 HSGSTM5	69	69	70	69	66	68	19	16	0
10 HSGSTM1b	71	68	71	70	66	66	18	13	12	0
11 HSGSTM2	68	67	70	67	64	69	22	17	18	16	0
12 HSGSTM3	68	67	67	68	66	66	31	29	28	29	32	0
13 RNGST4	68	67	69	68	65	69	19	18	17	17	15	32	0
14 RNGST3	69	69	69	69	67	68	21	26	22	21	24	33	22	0
15 CLGST1	70	69	71	69	68	70	20	25	22	20	22	34	21	11	0
16 GGGSTM2	70	70	71	71	69	71	33	33	33	35	36	35	33	39	38	0
17 MMGST3	69	69	69	69	66	70	20	28	23	22	25	34	22	7	11	38	0
18 MMGST4	71	69	71	70	67	67	11	18	18	16	21	29	18	19	19	36	20	0
19 SJGST1	71	70	71	73	70	71	56	54	54	56	54	56	56	55	56	56	55	56	0
20 SMGST1	71	69	71	72	70	71	60	60	59	58	59	58	59	58	58	58	58	60	21	0
21 FHGST1	70	70	71	71	70	69	52	51	51	50	53	53	53	51	51	53	52	52	43	44	0
22 OCGST1	70	68	69	68	67	75	79	78	78	78	77	76	79	78	77	78	77	77	73	72	72	0
23 OCGST2	71	69	71	71	68	71	78	77	79	76	76	76	79	78	77	78	78	76	72	71	72	21	0
24 HSGSTA1	72	69	71	70	68	73	79	78	79	77	76	76	79	79	77	79	79	77	73	71	73	17	19	0
25 HSGSTA2	71	70	71	70	68	73	79	77	78	76	76	75	79	80	78	78	80	77	74	72	73	20	20	5	0
26 MMGST2	70	68	70	70	67	72	76	76	76	75	77	74	78	78	77	78	79	75	73	73	73	25	24	23	24	0
27 MMGST1	71	68	70	70	68	73	75	75	76	75	74	74	76	76	75	77	77	75	71	71	70	29	28	23	25	29	0
28 RNGST2	70	68	70	69	66	71	77	77	76	75	76	74	78	78	77	77	78	76	72	71	73	23	24	24	26	14	29	0
29 MMGST8	67	65	68	69	66	72	79	78	79	77	77	77	77	78	78	78	78	78	72	72	71	40	37	37	38	39	39	39	0
30 RRGST8	66	65	67	68	66	72	80	79	80	78	79	77	78	78	78	79	78	79	73	72	72	43	38	38	39	39	40	39	9	0
31 RNGST1a	72	70	71	70	69	74	75	76	76	75	74	73	76	76	75	77	76	75	72	72	72	30	29	24	26	30	5	31	41	40	0
32 RNGST1b	71	69	70	70	68	74	75	75	76	75	73	73	76	76	75	76	75	74	72	72	72	29	29	23	26	30	6	32	40	40	4	0
33 GGGSTA3	71	70	72	72	70	74	78	79	78	77	77	76	78	78	77	79	78	77	72	72	74	35	35	35	35	36	38	36	30	31	39	39	0
34 SHGST1	76	77	78	76	78	74	74	74	74	73	74	75	74	74	73	74	74	73	74	75	74	73	72	71	72	73	73	73	73	74	74	74	73	0
35 SBGST1	76	76	78	76	77	75	74	74	73	73	74	74	74	74	73	74	74	73	74	74	74	73	73	72	73	73	74	73	73	74	74	74	73	0	0
36 SMGST2	75	75	77	75	77	73	73	72	73	72	73	73	73	74	72	72	74	71	73	73	73	75	73	72	73	73	74	73	73	74	75	74	75	10	8	0
37 ASGST1	70	71	72	71	68	65	72	72	70	72	73	72	70	73	73	73	74	70	76	78	73	73	71	72	72	69	71	68	69	69	72	71	69	70	70	69	0
38 DMelGST2	75	76	78	77	76	72	78	77	80	78	77	77	78	75	75	77	75	76	82	82	77	80	79	78	80	81	78	79	78	79	79	78	79	74	74	73	67	0
39 MDGST2	72	72	74	74	73	69	75	75	77	75	75	75	75	73	73	74	73	72	81	79	75	78	77	76	78	77	76	76	76	76	77	76	76	72	71	70	64	14	0
40 OVCRY1	74	74	74	72	73	77	76	75	76	77	75	74	74	75	75	77	75	75	77	79	79	76	76	76	75	76	76	76	76	77	75	75	80	79	79	78	71	78	77	0
41 OVCRY2	71	71	73	70	71	73	75	74	75	76	74	73	73	74	76	76	74	74	76	77	77	72	73	72	71	72	72	72	72	74	71	72	73	73	73	74	69	77	73	31	0
42 ODCRY1	71	71	72	70	71	74	75	74	74	75	73	74	72	74	75	76	74	73	76	77	77	74	74	73	73	74	72	74	71	73	71	71	73	72	72	73	69	76	73	34	7	0
43 ODCRY2	72	72	74	71	72	73	74	73	73	74	72	73	71	74	74	75	73	73	77	77	78	70	71	70	70	72	72	71	71	72	71	71	71	71	70	72	68	75	72	33	20	22	0
44 ODCRY3	71	71	73	70	71	73	76	75	75	76	74	75	73	75	75	76	74	75	76	77	77	72	74	71	71	73	72	73	72	73	71	71	73	73	73	74	67	75	72	26	19	22	20	0
45 ODCRY4	70	70	72	71	71	70	75	74	74	75	74	73	73	76	76	76	75	74	78	79	78	71	71	70	70	71	72	71	71	72	71	71	73	73	73	74	68	73	68	33	20	25	22	21	0
46 OSCRY1	73	74	76	75	75	74	79	78	78	79	77	77	77	79	78	80	78	78	77	77	78	71	73	73	73	72	72	73	71	72	72	72	73	74	74	75	73	80	77	51	42	44	41	42	41	0
47 OSCRY2	70	70	73	71	70	71	75	74	75	75	74	76	74	76	74	76	75	74	78	80	78	74	76	74	74	75	73	76	72	73	73	72	74	73	73	75	69	75	71	57	45	43	44	45	41	46	0
48 OSCRY3	73	72	75	75	72	74	77	76	77	77	76	76	75	78	77	77	77	76	78	78	78	74	75	74	74	74	72	75	74	75	73	72	75	76	76	76	69	75	71	51	40	40	43	44	41	49	47	0
49 FBDH	80	79	79	80	79	81	87	87	87	86	86	86	84	86	86	86	86	86	85	86	86	82	81	81	82	83	84	85	83	83	83	83	84	89	89	88	82	86	85	85	82	82	84	83	83	83	83	86	0
50 PPGST1	78	77	78	78	77	80	83	85	83	83	84	83	83	83	83	86	83	82	84	83	82	84	85	84	84	84	83	84	82	82	83	84	84	83	83	82	78	83	84	83	83	84	82	83	81	84	81	85	79	0
51 IOGST1	83	83	84	84	85	82	81	81	80	81	81	83	81	81	81	83	81	81	82	83	78	81	80	80	79	80	81	81	79	79	79	79	79	83	83	83	76	84	82	84	81	81	79	79	81	80	78	82	82	81	0
52 SacCURE2	84	84	84	83	83	84	84	83	84	85	83	86	84	86	85	86	84	84	85	86	88	83	84	82	82	83	84	84	83	83	85	84	83	84	84	83	84	88	87	88	83	83	85	84	83	83	85	85	81	76	81	0
53 ZMGST3	81	81	80	81	81	79	79	83	82	81	83	84	83	81	80	84	82	81	80	80	81	82	81	79	79	81	80	82	81	81	80	81	80	83	83	82	77	83	84	84	84	84	85	87	86	85	82	87	76	74	77	74	0
54 ZMGST1	76	76	75	75	76	76	76	77	76	74	77	78	76	76	74	82	77	75	79	80	80	78	80	74	75	78	77	78	77	77	76	77	77	78	77	76	78	84	84	80	79	79	79	79	80	81	79	81	78	73	77	73	49	0
55 TAGST1	77	78	77	76	77	81	76	79	77	76	78	78	76	79	77	81	79	78	80	81	79	82	80	79	79	80	78	82	80	80	79	79	80	79	79	78	78	84	84	78	77	77	78	79	78	80	80	80	81	76	78	74	56	47	0
56 TAGST2	77	79	78	77	77	80	77	79	77	76	79	78	76	79	77	83	80	79	81	82	80	82	80	79	79	80	79	82	81	80	80	80	80	80	80	80	80	84	84	79	77	77	78	80	79	81	79	81	81	77	79	75	59	49	4	0
57 DCGST1	81	81	81	80	81	82	82	84	84	82	84	83	81	81	81	85	82	80	82	82	83	76	77	76	76	78	78	76	78	78	77	78	77	82	82	79	77	83	82	83	80	81	82	82	82	82	82	78	77	72	80	79	68	71	77	77	0
58 ATGST2	78	76	79	77	78	77	79	82	80	79	80	82	79	79	79	84	79	79	80	80	81	80	76	78	78	78	77	77	76	76	77	77	79	81	81	79	78	84	83	81	78	79	79	81	81	82	81	82	79	78	78	78	59	67	69	70	70	0
59 ATGST1	81	81	82	82	82	83	80	84	83	82	84	83	84	83	82	85	83	82	78	77	81	80	80	76	76	80	77	81	78	78	77	78	75	81	81	80	83	84	86	84	83	83	84	84	84	85	82	81	84	73	80	76	53	60	59	61	72	63	0
60 ATGST3	82	82	82	82	82	80	77	81	80	79	81	81	80	80	79	82	80	79	77	77	82	81	81	77	77	81	77	81	79	81	77	77	76	80	79	78	83	85	87	82	80	80	80	82	82	83	80	82	85	75	82	76	55	59	60	61	72	65	26	0
61 NTGST1	80	80	81	80	80	85	83	85	85	84	85	86	85	84	84	84	85	84	79	79	85	80	79	77	77	80	78	81	81	81	78	78	77	81	81	80	81	88	89	85	85	85	86	87	86	85	85	85	81	74	81	75	52	60	56	59	75	63	42	42	0
62 SCGST1	83	82	81	83	83	83	81	84	82	83	84	83	83	81	82	81	82	81	79	79	82	79	79	76	76	81	75	82	80	81	75	76	78	82	82	82	81	86	83	81	82	82	83	84	84	84	81	82	84	76	82	79	54	58	59	60	77	66	46	43	41	0
63 HSGSTMIC	81	81	81	81	80	82	83	82	83	84	83	81	84	82	78	84	82	82	81	80	80	80	83	80	79	81	78	81	80	79	77	78	84	83	84	86	83	85	86	81	73	73	74	74	78	73	79	79	83	82	86	86	80	81	83	84	79	85	88	87	84	84	0
64 RNGSTMIC	85	85	85	85	84	84	87	85	87	88	87	84	88	85	82	87	84	86	81	81	83	84	85	84	83	85	82	85	83	82	81	82	86	86	87	89	86	84	87	81	76	76	77	76	79	76	81	81	83	81	87	85	83	82	86	87	80	89	88	87	84	85	17	0
65 DMelGST1	84	84	84	85	83	84	85	86	85	85	85	86	85	84	85	87	84	85	82	82	84	79	77	77	77	76	78	77	79	78	77	76	76	83	83	83	79	81	82	85	82	83	82	83	83	83	82	84	81	77	76	77	73	77	75	77	73	73	77	78	79	80	76	79	0
66 DMelGST3	85	84	82	84	83	84	85	84	84	85	86	83	84	85	85	86	84	85	85	83	85	80	80	79	78	79	79	79	79	78	79	79	78	83	83	83	78	82	82	85	80	81	81	81	81	81	83	83	80	79	79	73	75	78	75	77	72	74	78	76	80	82	73	80	33	0
67 ECSSPG	81	81	80	82	82	85	83	82	82	83	82	84	82	82	85	86	82	82	87	87	84	82	81	81	81	81	82	80	84	84	82	82	82	87	87	86	83	89	87	87	83	83	83	84	83	82	84	86	83	82	83	84	75	78	82	82	76	80	80	76	80	79	81	82	73	73	0
68 MDGST1	82	82	82	82	81	84	83	84	84	84	83	85	83	82	83	85	82	83	78	79	82	79	77	77	77	77	77	78	79	77	77	76	77	81	81	81	79	82	83	86	82	83	83	84	84	85	83	83	78	77	76	74	73	75	73	75	70	71	75	76	76	78	75	79	15	34	75	0
69 RNGST5	81	80	79	79	79	80	81	79	80	80	81	78	79	81	80	79	81	80	80	81	79	79	78	78	79	76	78	76	76	76	78	78	77	81	82	80	80	82	82	82	79	79	79	80	80	79	78	79	80	77	81	81	73	76	76	76	76	75	75	74	76	80	76	78	74	73	79	73	0
70 RNGST12	81	80	80	80	81	81	79	80	80	80	83	78	78	78	78	79	78	78	80	82	78	82	82	81	82	80	80	81	78	78	81	81	81	84	84	83	79	85	83	80	78	78	80	81	81	81	79	80	76	79	82	79	73	80	77	78	73	74	79	78	81	80	75	78	76	74	83	74	44	0
71 LCGST1	84	84	84	83	82	84	83	85	85	85	85	86	83	83	84	86	83	84	79	80	83	80	79	79	79	79	79	80	81	79	79	78	79	83	83	83	80	84	85	85	83	84	84	85	85	86	84	84	80	78	76	75	75	76	74	76	71	71	78	77	78	79	74	77	18	37	75	9	74	75	0
72 MBDMD	77	78	77	78	76	76	80	82	82	82	83	83	82	79	80	84	80	82	79	79	79	80	81	81	82	79	81	79	79	79	81	82	79	82	81	81	75	81	83	82	80	80	82	82	81	83	82	83	84	80	79	78	74	74	80	79	76	78	81	78	80	79	79	81	74	74	80	72	73	74	74	0

6.4 COMPARISON OF THE GST SEQUENCES

6.4.1 Results

6.4.1.1 Alignment

Alignment of the deduced amino acid sequences for the 72 GSTs collated for this study incorporates information from the recently determined 3D structure of three mammalian enzymes, each representing one of the established classes (SSGSTP1, Reimemer *et al.* 1991; HSGSTA1, Sinning *et al.* 1993 and RNGST3, Ji *et al.* 1992). The alignment is shown in Appendix 1. Pairwise percentage differences between the sequences are shown in matrix form in Table 6.2. These pairwise comparisons are derived from the multiple alignment and are not necessarily identical to scores that may be obtained when any two sequences alone are compared in isolation. All sequences have names beginning with a species designation, the most common are HS for human, RN for rat and MM for mouse.

In the NJ phylogenetic tree the distance between each pair of sequences estimates the total branch length between two species. Thus distances are related to the degree of divergence between the sequences. For the comparison of the full complement of amino acid sequences, distance matrices were calculated using three different models of amino acid replacement, (as described in section 6.3.4). Differences among the replacement models in amino acid sequence comparisons produced little change in the topology of the trees (data not shown). The main differences were in the lengths of some branches. The simple distance model tree is discussed here.

The only residue conserved throughout this current alignment is Pro56 (all residues numbered as for HSGSTA1 and include the initiating methionine), a residue not considered to be part of either the GSH-binding site or the second substrate binding site. It is, however, thought to be of conformational importance in Alpha class HSGSTA1 (Sinning *et al.* 1993).

1) It should be noted that the branch lengths depicted in the MP tree are of a constant length, this is independent of the number of changes needed to move between adjacent nodes on the tree. Thus branch lengths in the MP tree do not represent genetic distances.

Two areas of generally higher similarity have also been identified. The first region is nearer the N-terminus and is comprised of the amino acid residues 56-80 (including the "snail/trail" sequence) of the first domain. Distance estimates for the entire length of the sequences (global distances) are in the range 60% to 82%, the mean being near 78%. For this N-terminal region the estimated distances in the majority of the sequence comparisons are encompassed within the range 54% to 76%, the mean nearer 63% (Table 6.3a). Sequences with higher levels of divergence in this region are generally those with high levels overall such as FBDH from *Flavobacterium* (distances of up to 89% when compared to some of the Alpha class). Evidence from crystallographic studies suggests several of the residues in this region are included in the GSH-binding site and others in maintenance of the 3D conformation of the protein (Reinemer *et al.* 1991, Ji *et al.* 1992 and Sinning *et al.* 1993).

The second region of increased similarity is nearer the C-terminus and is comprised of the residues 147-181. This region is not as well conserved as the N-terminal region, with distances generally reduced from the global estimates (60% to 82%) to between 56% and 78%, with the mean near 68% (Table 6.3b). Sequences that exhibit higher levels of divergence in this region are again those with high levels overall, such as FBDH. IOGST1, the yeast (*Issatchenkis orientalis*), shows little similarity with any sequence in this region, the lowest distance being shared with both Alpha class GSTs and DCGST1, at 69%. This C-terminal region contains no residues thought to be involved in either GSH or in second substrate binding, although, again, some are implicated in functions involved in structural integrity such as the interactions between the two domains of a GST subunit and in dimer interactions (Sinning *et al.* 1993).

6.4.1.2 Topology of the phylogenetic trees

Figure 6.1a shows a schematic of the maximum parsimony (MP) tree. Figure 6.1b shows the neighbor-joining (NJ) tree, constructed from distance estimates. Both are derived from the alignment of all available full length GST amino acid sequences. 1)

Table 6.3a Distance matrix of all pairwise comparisons of GSTs in the N-terminal region of increased similarity. Column 1 and row 73 orders sequences for ease of comparison. Matrix generated by CLUSTALV.

1	MMGST7	0
2	RNGST7	0	0
3	HSGSTP1	9	9	0
4	BTGSTP1	0	0	9	0
5	SSGSTP1	0	0	9	0	0
6	CEGSTP1	36	36	45	36	36	0
7	MAGSTM1	36	36	45	36	36	45	0
8	HSGSTM4	45	45	55	45	45	55	9	0
9	HSGSTM5	36	36	45	36	36	45	0	9	0
10	HSGSTM1b	45	45	55	45	45	55	9	0	9	0
11	HSGSTM2	36	36	45	36	36	45	0	9	0	9	0
12	HSGSTM3	36	36	45	36	36	45	0	9	0	9	0	0
13	RNGST4	27	27	36	27	27	36	9	18	9	18	9	9	0
14	RNGST3	36	36	45	36	36	36	18	27	18	27	18	18	9	0
15	CLGST1	36	36	45	36	36	45	0	9	0	9	0	0	9	18	0
16	GGGSTM2	27	27	36	27	27	55	9	18	9	18	9	9	18	27	9	0
17	MMGST3	27	27	36	27	27	36	9	18	9	18	9	9	0	9	9	18	0
18	MMGST4	27	27	36	27	27	36	9	18	9	18	9	9	0	9	9	18	0	0
19	SJGST1	45	45	55	45	45	45	27	36	27	36	27	27	36	36	27	18	36	36	0
20	SMGST1	45	45	55	45	45	45	27	36	27	36	27	27	36	36	27	18	36	36	0	0
21	FHGST1	45	45	55	45	45	55	27	36	27	36	27	27	36	27	27	18	36	36	18	18	0
22	OCGSTA1	55	55	64	55	55	64	55	55	55	55	55	55	55	64	55	45	55	55	55	55	55	0
23	OCGSTA2	64	64	73	64	64	64	55	55	55	55	55	55	64	64	55	45	64	64	45	45	45	18	0
24	HSGSTA1	55	55	64	55	55	64	45	45	45	45	45	45	55	64	45	36	55	55	45	45	45	9	9	0
25	HSGSTA2	55	55	64	55	55	64	45	45	45	45	45	45	55	64	45	36	55	55	45	45	45	9	9	0	0
26	MMGST2	55	55	64	55	55	64	45	45	45	45	45	45	55	64	45	36	55	55	45	45	45	18	18	9	9	0
27	MMGST1	55	55	64	55	55	64	45	45	45	45	45	45	55	64	45	36	55	55	45	45	45	9	9	0	0	9	0
28	RNGST2	55	55	64	55	55	64	45	45	45	45	45	45	55	64	45	36	55	55	45	45	45	9	9	0	0	9	0	0
29	MMGST8	45	45	55	45	45	64	45	45	45	45	45	45	36	45	45	36	36	36	45	45	45	27	36	27	27	36	27	27	0
30	RRGST8	45	45	55	45	45	64	45	45	45	45	45	45	36	45	45	36	36	36	45	45	45	27	36	27	27	36	27	27	0	0
31	RNGST1a	55	55	64	55	55	73	45	45	45	45	45	45	55	64	45	36	55	55	45	45	45	18	18	9	9	18	9	9	27	27	0
32	RNGST1b	55	55	64	55	55	73	45	45	45	45	45	45	55	64	45	36	55	55	45	45	45	18	18	9	9	18	9	9	27	27	0	0
33	GGGSTA3	55	55	64	55	55	64	45	45	45	45	45	45	55	64	45	36	55	55	45	45	45	9	9	0	0	9	0	0	27	27	9	9	0
34	SHGST1	64	64	73	64	64	55	55	64	55	64	55	55	55	55	55	55	55	55	55	55	55	64	64	64	64	64	64	64	73	73	73	73	64	0
35	SBGST1	64	64	73	64	64	64	55	64	55	64	55	55	55	55	55	55	55	55	55	55	55	73	73	73	73	73	73	73	73	73	73	73	73	9	0
36	SMGST2	64	64	73	64	64	64	55	64	55	64	55	55	55	55	55	55	55	55	55	55	55	73	73	73	73	73	73	73	73	73	73	73	73	9	0	0
37	ASGST1	36	36	45	36	36	45	45	55	45	55	45	45	36	36	45	36	36	36	36	36	36	55	55	55	55	55	55	55	45	45	45	45	55	55	55	55	0
38	DMelGST2	64	64	64	64	64	64	73	82	73	82	73	73	64	64	73	73	64	64	73	73	73	91	91	91	91	91	91	91	82	82	91	91	91	73	73	73	64	0
39	MDGST2	64	64	64	64	64	64	73	82	73	82	73	73	64	64	73	73	64	64	73	73	73	91	91	91	91	91	91	91	82	82	91	91	91	73	73	73	64	0	0
40	OVCRY1	67	67	78	67	67	67	67	78	67	78	67	67	56	56	67	67	56	56	67	67	67	78	78	78	78	78	78	78	67	67	78	78	78	44	44	44	56	56	56	0
41	OVCRY2	45	45	55	45	45	36	36	45	36	45	36	36	27	27	36	45	27	27	36	36	45	64	64	64	64	64	64	64	55	55	64	64	64	45	45	45	36	55	55	33	0
42	ODCRY1	45	45	55	45	45	36	36	45	36	45	36	36	27	27	36	45	27	27	36	36	45	64	64	64	64	64	64	64	55	55	64	64	64	45	45	45	36	55	55	33	0	0
43	ODCRY2	55	55	64	55	55	45	45	55	45	55	45	45	36	36	45	55	36	36	45	45	55	55	55	55	55	55	55	55	45	45	55	55	55	55	55	55	45	64	64	33	9	9	0
44	ODCRY3	55	55	64	55	55	45	45	55	45	55	45	45	36	36	45	55	36	36	45	45	55	73	73	73	73	73	73	73	64	64	73	73	73	55	55	55	45	45	45	22	9	9	18	0
45	ODCRY4	55	55	64	55	55	45	45	55	45	55	45	45	36	36	45	55	36	36	45	45	55	73	73	73	73	73	73	73	64	64	73	73	73	55	55	55	45	45	45	22	9	9	18	0	0
46	OSCRY1	64	64	73	64	64	64	64	73	64	73	64	64	55	55	64	64	55	55	55	55	64	64	64	64	64	64	64	64	55	55	64	64	64	36	36	36	55	73	73	22	27	27	18	36	36	0
47	OSCRY2	45	45	55	45	45	45	45	55	45	55	45	45	36	36	45	45	36	36	36	36	45	64	64	64	64	64	64	64	55	55	64	64	64	55	55	55	36	55	55	44	18	18	27	27	27	36	0
48	OSCRY3	55	55	64	55	55	45	45	55	45	55	45	45	36	36	45	55	36	36	45	45	55	55	55	55	55	55	55	55	45	45	55	55	55	55	55	55	45	64	64	33	9	9	0	18	18	18	27	0
49	FBDH	67	67	67	67	67	67	67	67	67	67	67	67	56	56	67	67	56	56	67	67	67	89	89	89	89	89	89	89	67	67	89	89	89	78	78	78	67	67	67	71	67	67	78	78	78	78	56	78	0
50	PPGST1	64	64	73	64	64	73	73	73	73	73	73	73	64	64	73	64	64	64	64	64	64	73	73	73	73	73	73	73	64	64	73	73	73	64	64	64	55	82	82	44	64	64	73	64	64	64	64	73	67	0
51	HSGSTMIC	64	64	73	64	64	73	64	64	64	64	64	64	64	73	64	64	64	64	73	73	73	73	82	73	73	73	73	73	73	73	73	73	73	73	73	73	73	82	82	78	64	64	73	73	73	73	64	73	78	82	0
52	RNGSTMIC	82	82	91	82	82	73	82	82	82	82	82	82	82	82	82	82	82	82	73	73	82	91	91	91	91	91	91	91	91	91	91	91	91	82	82	82	82	82	82	78	73	73	82	73	73	82	73	82	89	82	25	0
53	DMelGST1	55	55	64	55	55	64	64	64	64	64	64	64	55	55	64	55	55	55	55	55	55	55	55	55	55	64	55	55	45	45	55	55	55	55	55	55	45	82	82	56	55	55	64	64	64	55	55	64	67	45	73	82	0
54	DMelGST3	55	55	64	55	55	55	45	45	45	45	45	45	36	45	45	55	36	36	64	64	64	55	64	55	55	64	55	55	45	45	55	55	55	55	55	55	55	82	82	56	45	45	55	55	55	55	55	55	67	55	64	82	27	0
55	ECSSPB	64	64	64	64	64	73	73	73	73	73	73	73	64	55	73	64	64	64	64	64	55	64	64	64	64	73	64	64	55	55	64	64	64	64	64	64	55	82	82	67	64	64	73	73	73	64	64	73	67	55	82	91	27	36	0
56	MDGST1	55	55	64	55	55	64	64	64	64	64	64	64	55	45	64	55	55	55	55	55	45	55	55	55	55	64	55	55	45	45	55	55	55	55	55	55	45	82	82	56	55	55	64	64	64	55	55	64	67	45	73	82	9	27	18	0
57	RNGST5	50	50	60	50	50	60	50	50	50	50	50	50	40	50	50	50	40	40	60	60	50	60	70	60	60	60	60	60	50	50	60	60	60	50	50	50	50	80	80	56	50	50	60	60	60	50	50	60	63	50	60	80	40	30	50	40	0
58	RNGST12	50	50	60	50	50	60	40	40	40	40	40	40	30	40	40	40	30	30	50	50	50	60	70	60	60	60	60	60	40	40	60	60	60	50	50	50	50	80	80	56	50	50	60	60	60	50	50	60	50	50	50	70	40	20	50	40	30	0
59	LCGST1	55	55	64	55	55	64	64	64	64	64	64	64	55	45	64	55	55	55	55	55	45	55	55	55	55	64	55	55	45	45	55	55	55	55	55	55	45	82	82	56	55	55	64	64	64	55	55	64	67	45	73	82	9	27	18	0	40	40	0
60	MBDMD	64	64	73	64	64	64	55	64	55	64	55	55	55	55	55	55	55	55	55	55	45	64	73	73	73	73	73	73	73	73	73	73	73	36	36	36	55	64	64	44	45	45	55	55	55	45	45	55	67	64	73	82	45	45	55	45	40	50	45	0
61	IOGST1	73	73	82	73	73	82	55	55	55	55	55	55	64	64	55	45	64	64	45	45	45	64	55	55	55	55	55	55	55	55	55	55	55	64	64	64	64	91	91	67	73	73	73	82	82	64	73	73	67	64	82	91	55	64	55	55	60	50	55	64	0
62	SacCURE2	45	45	55	45	45	36	55	55	55	55	55	55	45	55	55	64	45	45	73	73	73	73	82	73	73	73	73	73	64	64	73	73	73	64	64	64	64	82	82	67	55	55	64	64	64	64	64	64	67	64	64	82	45	27	55	45	30	40	45	55	73	0
63	ZMGST3	55	55	64	55	55	64	45	55	45	55	45	45	55	55	45	36	55	55	36	36	36	55	45	45	45	55	45	45	55	55	45	45	45	45	45	45	45	82	82	56	55	55	64	64	64	55	55	64	78	55	73	82	36	45	45	36	50	50	36	45	45	64	0
64	ZMGST1	64	64	73	64	64	73	64	64	64	64	64	64	64	64	64	55	64	64	55	55	55	55	55	55	55	64	55	55	55	55	55	55	55	55	55	55	55	91	91	67	64	64	73	73	73	64	64	73	78	45	73	82	36	45	45	36	50	50	36	55	55	64	27	0
65	TAGST1	64	64	73	64	64	73	55	55	55	55	55	55	64	64	55	45	64	64	45	45	45	55	45	45	45	55	45	45	55	55	45	45	45	45	45	45	55	82	82	56	55	55	64	64	64	55	64	64	78	55	73	82	36	45	45	36	50	50	36	45	45	64	18	18	0
66	TAGST2	73	73	82	73	73	73	55	55	55	55	55	55	64	64	55	55	64	64	55	55	55	64	55	55	55	64	55	55	64	64	55	55	55	55	55	55	64	82	82	67	55	55	64	64	64	64	64	64	78	73	73	82	55	55	64	55	60	60	55	55	64	73	45	45	27	0
67	DCGST1	64	64	73	64	64	45	55	55	55	55	55	55	45	45	55	64	45	45	55	55	64	73	73	73	73	73	73	73	64	64	64	64	73	55	55	55	45	82	82	67	45	45	55	55	55	64	55	55	67	55	64	64	55	45	64	55	50	50	55	64	73	55	64	64	64	64	0
68	ATGST2	45	45	55	45	45	45	45	55	45	55	45	45	36	36	45	55	36	36	55	55	55	64	64	64	64	73	64	64	55	55	64	64	64	45	45	45	45	73	73	44	36	36	45	45	45	45	36	45	67	55	73	82	36	27	45	36	40	40	36	45	64	45	36	45	45	55	45	0
69	ATGST1	64	64	73	64	64	73	55	55	55	55	55	55	64	64	55	45	64	64	45	45	45	55	45	45	45	55	45	45	55	55	45	45	45	55	55	55	55	91	91	67	64	64	73	73	73	64	64	73	78	55	73	82	36	45	45	36	50	50	36	55	45	64	18	27	18	45	64	45	0
70	ATGST3	73	73	82	73	73	64	45	45	45	45	45	45	55	55	45	55	55	55	55	55	55	64	55	55	55	64	55	55	64	64	55	55	55	55	55	55	64	91	91	67	55	55	64	64	64	64	64	64	78	64	73	82	45	36	55	45	50	50	45	55	55	55	27	36	27	45	55	36	9	0
71	NTGST1	64	64	73	64	64	73	55	55	55	55	55	55	64	64	55	45	64	64	45	45	45	55	45	45	45	55	45	45	55	55	45	45	45	55	55	55	55	91	91	67	64	64	73	73	73	64	64	73	78	55	73	82	36	45	45	36	50	50	36	55	45	64	18	27	18	45	64	45	0	9	0
72	SCGST1	55	55	64	55	55	64	64	64	64	64	64	64	55	55	64	55	55	55	55	55	55	55	55	55	55	64	55	55	45	45	55	55	55	55	55	55	45	82	82	56	55	55	64	64	64	55	55	64	67	45	73	82	27	36	36	27	40	40	27	55	55	55	27	18	18	45	55	36	18	27	18	0
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72

Table 6.3b Distance matrix of all pairwise comparisons of GSTs in the C-terminal region of increased similarity. Column 1 and row 73 orders sequences for ease of comparison. Matrix generated by CLUSTALV.

1	MMGST7	0
2	RNGST7	6	0
3	HSGSTP1	0	6	0
4	BTGSTP1	0	6	0	0
5	SSGSTP1	13	13	13	13	0
6	CEGSTP1	56	63	56	56	56	0
7	MAGSTM1	69	75	69	69	75	69	0
8	HSGSTM4	56	63	56	56	63	75	19	0
9	HSGSTM5	63	69	63	63	69	75	13	6	0
10	HSGSTM1b	69	63	69	69	69	81	19	19	13	0
11	HSGSTM2	63	69	63	63	69	75	25	13	13	25	0
12	HSGSTM3	75	75	75	75	75	81	25	25	19	19	31	0
13	RNGSTYB3	63	69	63	63	69	75	19	13	6	19	6	25	0
14	RNGST3	75	81	75	75	81	69	13	19	13	25	25	25	19	0
15	CLGST1	81	88	81	81	88	81	31	31	25	38	38	38	31	19	0
16	GGGSTM2	63	69	63	63	69	75	25	6	13	25	19	19	19	19	31	0
17	MMGST3	75	81	75	75	81	69	13	19	13	25	25	25	19	0	19	19	0
18	MMGST4	81	75	81	81	81	75	19	31	25	13	38	25	31	13	31	31	13	0
19	SJGST1	69	75	69	69	75	69	56	56	56	63	56	63	56	50	56	56	50	56	0
20	SMGST1	75	69	75	75	75	75	63	63	63	56	63	63	63	56	63	63	56	50	13	0
21	FHGST1	81	81	81	81	75	69	69	69	69	63	69	69	69	63	75	69	63	56	44	44	0
22	OCGST1	63	56	63	63	63	56	81	75	81	75	88	75	88	81	88	69	81	75	69	63	63	0
23	OCGST2	63	56	63	63	63	50	75	69	75	69	81	69	81	75	81	63	75	69	69	63	63	6	0
24	HSGSTA1	63	56	63	63	63	50	81	75	81	75	81	75	81	81	88	69	81	75	69	63	63	19	13	0
25	HSGSTA2	63	56	63	63	63	50	81	75	81	75	81	75	81	81	88	69	81	75	69	63	63	19	13	0	0
26	MMGST2	63	56	63	63	63	56	88	81	88	81	88	81	88	88	94	75	88	81	69	63	63	13	19	13	13	0
27	MMGST1	69	63	69	69	69	69	75	69	75	69	75	69	75	75	88	63	75	69	63	56	63	38	44	31	31	31	0
28	RNGST2	63	56	63	63	63	63	88	81	88	81	88	81	88	88	94	75	88	81	69	63	69	25	31	25	25	19	31	0
29	MMGST8	44	38	44	44	44	63	88	81	88	81	88	81	88	88	94	75	88	81	75	69	69	31	31	31	31	31	44	44	0
30	RRGST8	44	38	44	44	44	63	88	81	88	81	88	81	88	88	94	75	88	81	75	69	69	31	31	31	31	31	44	44	0	0
31	RNGST1a	69	63	69	69	69	69	75	69	75	69	75	69	75	75	88	63	75	69	63	56	63	31	38	25	25	25	6	38	38	38	0
32	RNGST1b	69	63	69	69	69	63	69	63	69	63	69	63	69	69	81	56	69	63	63	56	63	38	31	19	19	31	13	44	38	38	6	0
33	GGGSTA3	50	44	50	50	50	63	88	81	88	81	88	81	88	88	94	75	88	81	75	69	69	25	31	25	25	25	31	38	13	13	25	31	0
34	SHGST1	63	69	63	63	69	56	69	63	69	75	69	75	69	63	63	63	63	69	63	69	81	69	63	63	63	69	69	69	69	69	69	63	69	0
35	SBGST1	63	69	63	63	69	56	69	63	69	75	69	75	69	63	63	63	63	69	63	69	81	69	63	63	63	69	69	69	69	69	69	63	69	0	0
36	SMGST2	63	69	63	63	69	56	69	63	69	75	69	75	69	63	63	63	63	69	63	69	81	69	63	63	63	69	69	69	69	69	69	63	69	0	0	0
37	ASGST1	69	69	69	69	69	69	69	69	75	75	75	75	75	69	75	69	69	69	63	63	75	69	69	69	69	69	69	69	69	69	69	69	69	50	50	50	0
38	DMelGST2	80	80	80	80	80	73	73	73	73	73	67	67	73	73	80	67	73	73	80	80	87	73	67	67	67	73	67	60	67	67	73	67	67	53	53	53	67	0
39	MDGST2	87	80	87	87	87	80	80	80	80	73	73	73	80	80	87	73	80	73	87	80	93	73	67	67	67	73	67	60	67	67	73	67	67	60	60	60	73	0	0
40	OVCRY1	69	75	69	69	75	69	63	63	63	69	63	69	63	63	69	63	63	69	56	63	69	75	75	75	75	75	75	75	75	75	75	75	81	69	69	69	69	73	80	0
41	OVCRY2	63	69	63	63	69	75	63	63	63	69	63	69	63	63	69	63	63	69	63	69	75	81	81	81	81	81	81	81	69	69	81	81	75	63	63	63	69	73	80	13	0
42	ODCRY1	63	69	63	63	69	75	69	69	69	75	69	75	69	69	75	69	69	75	63	69	75	81	81	81	81	81	81	81	69	69	81	81	75	63	63	63	69	73	80	19	6	0
43	ODCRY2	63	69	63	63	69	75	63	63	63	69	63	69	63	63	69	63	63	69	63	69	75	81	81	81	81	81	81	81	69	69	81	81	75	69	69	69	75	73	80	6	6	13	0
44	ODCRY3	63	69	63	63	69	75	63	63	63	69	63	69	63	63	69	63	63	69	63	69	75	81	81	81	81	81	81	81	69	69	81	81	75	69	69	69	75	73	80	6	6	13	0	0
45	ODCRY4	56	63	56	56	63	69	63	63	63	69	63	75	63	69	75	69	69	75	69	75	81	81	81	81	81	81	81	81	69	69	81	81	75	69	69	69	75	73	80	19	19	25	13	13	0
46	OSCRY1	69	75	69	69	75	81	69	69	69	75	69	75	69	69	63	69	69	75	69	75	75	88	88	88	88	88	88	88	75	75	88	88	81	75	75	75	88	87	93	25	25	31	19	19	31	0
47	OSCRY2	69	75	69	69	75	75	63	63	63	69	63	69	63	69	63	69	69	75	75	81	81	88	88	88	88	88	88	88	81	81	88	88	88	69	69	69	88	87	93	56	63	69	56	56	50	44	0
48	OSCRY3	88	88	88	88	88	81	69	69	69	69	69	63	69	69	69	69	69	69	63	63	69	81	81	81	81	81	75	81	88	88	75	75	88	75	75	75	69	80	87	25	25	31	31	31	44	38	69	0
49	FBDH	63	63	63	63	63	75	81	88	81	81	88	81	81	81	88	88	81	81	81	81	69	69	69	69	69	69	88	75	63	63	81	81	63	81	81	81	75	80	87	75	69	69	69	69	69	75	81	81	0
50	PPGST1	69	69	69	69	63	63	81	88	81	75	88	81	81	81	81	88	81	75	88	88	75	75	75	75	75	75	81	69	69	69	88	88	69	75	75	75	69	73	80	81	75	75	75	75	69	75	81	88	63	0
51	IOGST1	81	75	81	81	81	75	81	75	75	75	75	88	75	81	88	81	81	81	88	81	81	69	69	69	69	69	81	69	75	75	75	75	75	88	88	88	81	87	80	75	81	88	81	81	69	88	75	81	81	81	0
52	SacCURE2	69	75	69	69	75	69	63	56	63	69	63	63	63	63	69	50	63	69	69	75	81	56	50	50	50	50	63	56	63	63	63	56	63	50	50	50	63	53	60	69	63	63	69	69	69	81	75	75	75	75	88	0
53	ZMGST3	69	75	69	69	75	69	69	75	69	75	75	75	69	69	69	75	69	75	63	69	75	69	69	69	69	69	69	69	75	75	69	69	75	63	63	63	69	67	73	63	69	69	69	69	69	75	75	75	75	63	81	63	0
54	ZMGST1	63	69	63	63	69	63	75	81	75	81	81	75	75	75	75	75	75	81	69	75	69	50	50	50	50	50	69	50	63	63	69	69	63	63	63	63	69	73	80	69	69	69	75	75	75	75	75	75	69	63	81	50	38	0
55	TAGST1	56	63	56	56	63	69	63	63	56	63	63	69	56	69	75	69	69	75	63	69	75	69	69	69	69	69	69	69	75	75	69	69	75	56	56	56	56	73	80	63	63	63	69	69	63	81	75	69	75	63	75	50	38	38	0
56	TAGST2	50	56	50	50	56	63	69	69	63	69	69	75	63	75	81	75	75	81	69	75	69	63	63	63	63	63	75	63	69	69	75	75	69	63	63	63	63	80	87	69	69	69	75	75	69	81	75	75	69	56	75	56	44	31	6	0
57	DCGST1	69	75	69	69	75	63	81	81	81	88	81	88	81	75	75	81	75	81	63	69	69	69	69	69	69	69	75	69	75	75	75	75	75	44	44	44	63	67	73	69	69	69	75	75	75	75	75	75	81	69	69	63	50	56	56	56	0
58	ATGST2	75	69	75	75	75	75	75	81	75	69	81	75	75	75	75	81	75	69	69	63	75	63	63	63	63	63	56	63	63	63	56	56	63	63	63	63	63	73	73	69	69	69	75	75	75	81	81	69	75	56	69	56	38	44	38	44	50	0
59	ATGST1	88	88	88	88	88	81	75	81	75	75	81	69	75	75	75	81	75	75	69	69	75	75	75	69	69	75	63	75	81	81	63	63	75	63	63	63	69	80	87	75	75	75	81	81	81	81	81	63	81	63	81	63	44	50	38	44	63	31	0
60	ATGST3	88	94	88	88	94	81	69	75	69	75	75	69	69	69	75	75	69	75	63	75	75	81	81	75	75	81	63	81	88	88	63	63	81	63	63	63	75	87	87	75	75	75	81	81	81	81	75	75	88	69	88	63	50	56	44	50	63	44	31	0
61	NTGST1	81	88	81	81	88	75	69	69	69	75	69	75	69	69	75	69	69	75	63	69	69	75	75	69	69	75	56	75	81	81	56	56	75	56	56	56	63	80	87	69	69	69	75	75	75	75	75	69	88	63	88	56	44	50	31	38	56	38	25	25	0
62	SCGST1	81	88	81	81	88	81	75	75	75	81	75	81	75	75	81	75	75	81	69	75	75	81	81	75	75	81	56	81	81	81	56	56	75	69	69	69	69	87	87	75	75	75	81	81	81	81	81	75	94	69	81	63	50	56	44	50	63	38	38	31	13	0
63	HSGSTMIC	82	82	82	82	73	82	82	82	82	82	82	82	82	73	64	82	73	73	73	73	82	91	91	91	91	91	82	82	82	82	82	82	82	64	64	64	73	73	80	82	73	73	73	73	73	73	82	82	82	0.6	82	64	82	82	91	73	73	73	82	64	75	0	0
64	RNGSTMIC	82	82	82	82	73	82	82	82	82	82	82	82	82	73	64	82	73	73	73	73	82	91	91	91	91	91	82	82	82	82	82	82	82	64	64	64	73	73	80	82	73	73	73	73	73	73	82	82	82	0.6	82	64	82	82	91	73	73	73	82	64	75	0	0	82
65	DMelGST1	69	75	69	69	75	63	69	75	69	75	75	69	69	69	75	69	69	75	63	69	81	50	50	50	50	44	63	56	63	63	56	56	63	50	50	50	63	60	67	63	69	69	69	69	69	81	75	75	69	69	75	50	56	50	50	56	56	63	69	69	63	69	82	82	0
66	DMelGST27	63	69	63	63	63	56	69	75	69	75	75	69	69	69	75	69	69	75	63	63	81	56	56	56	56	50	56	63	56	56	50	50	56	56	56	56	69	67	73	63	69	69	69	69	69	81	75	75	63	69	75	50	56	56	56	63	63	56	69	69	69	69	73	73	25	0
67	ECSSPB	73	80	73	73	80	67	73	73	73	80	67	80	73	73	80	73	73	80	73	80	73	73	73	73	73	73	73	67	80	80	73	73	80	67	67	67	87	79	79	73	80	87	80	80	80	73	67	80	87	73	80	87	60	67	80	73	60	73	73	53	60	60	82	82	67	67	0
68	MDGST1	63	69	63	63	69	63	69	75	69	75	75	69	69	69	75	69	69	75	63	69	81	63	63	63	63	56	56	63	63	63	56	56	63	44	44	44	56	60	67	63	63	63	69	69	69	81	75	69	75	69	81	50	56	50	44	50	50	50	63	63	56	56	82	82	13	25	67	0
69	RNGST5	56	56	56	56	56	69	81	69	75	81	75	88	75	81	81	75	81	88	75	81	75	63	63	63	63	63	69	56	56	56	75	75	56	56	56	56	69	73	80	81	75	75	75	75	69	75	75	88	69	56	75	69	69	63	63	56	75	69	69	69	69	81	73	73	81	81	80	81	0
70	RNGST12	56	56	56	56	63	75	75	81	75	75	81	75	75	69	69	81	69	69	69	69	69	75	75	75	75	75	69	81	56	56	69	69	63	56	56	56	63	73	80	63	50	50	56	56	56	63	69	69	56	63	88	69	63	63	63	69	63	50	63	69	69	69	55	55	69	63	87	56	69	0
71	LCGST1	63	69	63	63	63	56	69	75	69	75	75	69	69	69	75	69	69	75	63	69	81	63	63	63	63	56	56	63	63	63	56	56	63	50	50	50	56	67	73	63	63	63	69	69	69	81	75	69	75	63	81	50	56	50	44	50	56	50	63	63	56	56	73	73	19	19	67	6	81	56	0
72	MBDMD	69	75	69	69	69	44	75	75	75	88	75	88	75	63	75	75	63	75	69	75	69	69	69	69	69	63	81	69	75	75	81	81	75	63	63	63	63	80	87	69	69	69	75	75	69	81	81	75	75	63	69	63	63	56	63	56	50	69	75	75	69	75	73	73	63	56	67	56	69	69	50	0
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72

It can be clearly seen in the phylogenetic trees (Figure 6.1a and b) that there are several major groupings of GSTs exhibiting higher levels of sequence similarity within each group. In each of these groups sequences from different species can show more similarity to each other than sequences from the same species but from a different group. Three of these groups have been well characterised, primarily by methods other than sequence comparison, and consisted initially of enzymes only from mammalian species. These are the Alpha, Mu and Pi classes designated formally by Mannervik *et al.* (1985). The Alpha, Mu and Pi classes are highlighted in blue, green and red, respectively, in Figure 6.1a and b. In addition, mammalian sequences that have been characterised as the newer, Theta class (Hiratsuka *et al.* 1990, Ogura *et al.* 1991, Meyer *et al.* 1991, Hussey and Hayes 1992) are clearly separated from any other mammalian sequences. The Theta class is highlighted in Gold in Figure 6.1. Bootstrap confidence levels (shown in Table 6.4) for the clustering of the each of the Pi and Alpha classes are 100%. The Mammalian sequences of the Mu class cluster at this level also but, in this class, the addition of the non-vertebrate sequences lowers confidence marginally, to 99%. Only two rat sequences currently are generally thought to constitute Theta class but these also cluster at the 100% confidence level. Undoubtedly more Theta class sequences will soon become available. Table 6.5 shows percentage differences between the mammalian classes for representative sequences. The characteristics of these classes are discussed further in section 6.5.

Buetler and Eaton (1993) proposed a new class, Sigma (highlighted in aqua in Figure 6.1), for the cephalopod (octopus and squid species) eye lens crystallin sequences and the 28kDa protein sequences from *Schistosome* species (represented here by SMGST2, SBGST2 and SHGST2; SJGST2 was not included in the alignment as it lacked the N-terminus). The similarities among the crystallin sequences are generally within the range designated for a class, with differences between 7% and 57% and all cluster together at the 100% confidence level.

Figure 6.1a Phylogenetic tree of all GST sequences constructed by maximum parsimony. Alpha class is coloured blue, Mu is green, Pi is red, Sigma is purple and Theta is gold.

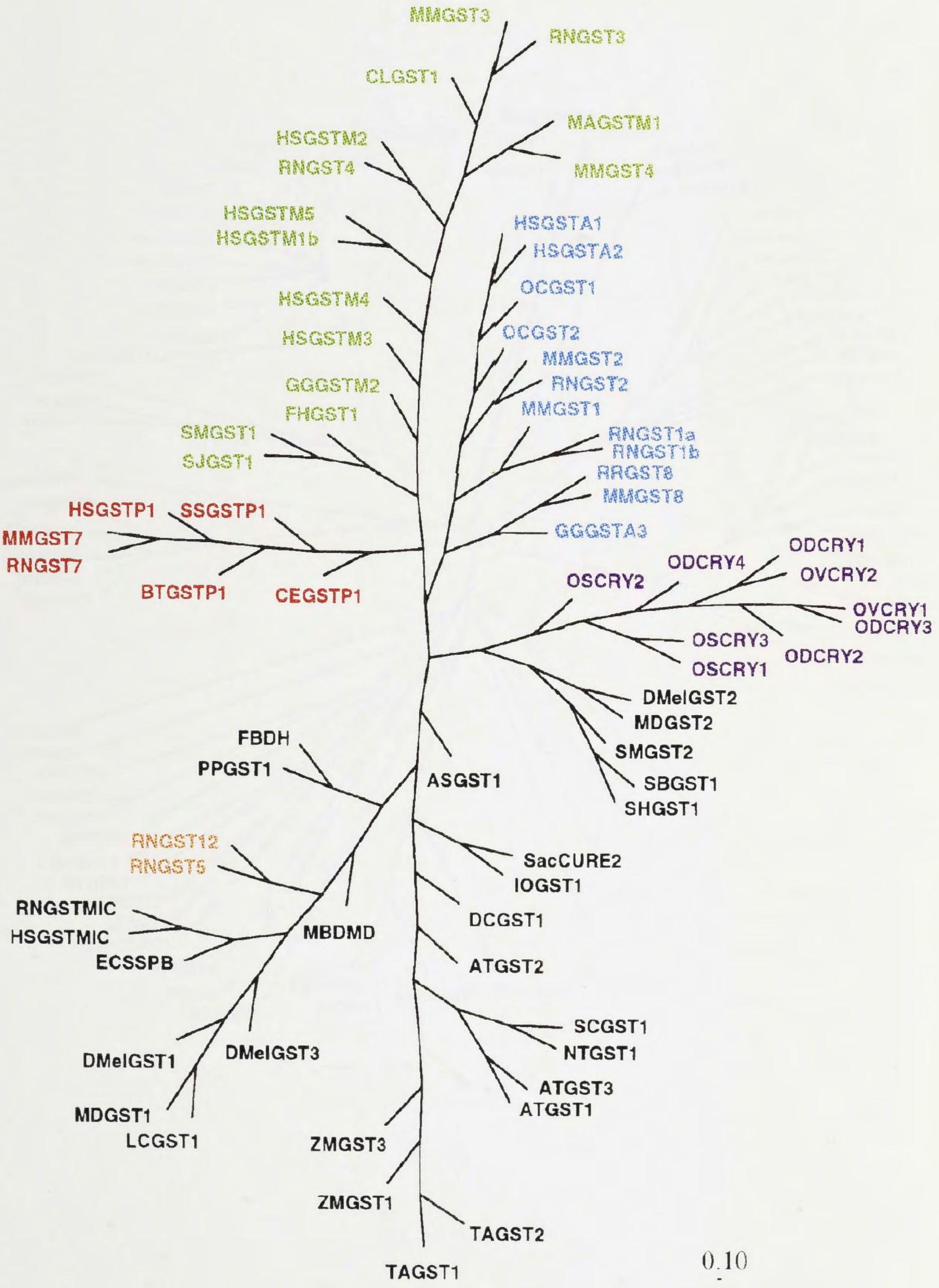
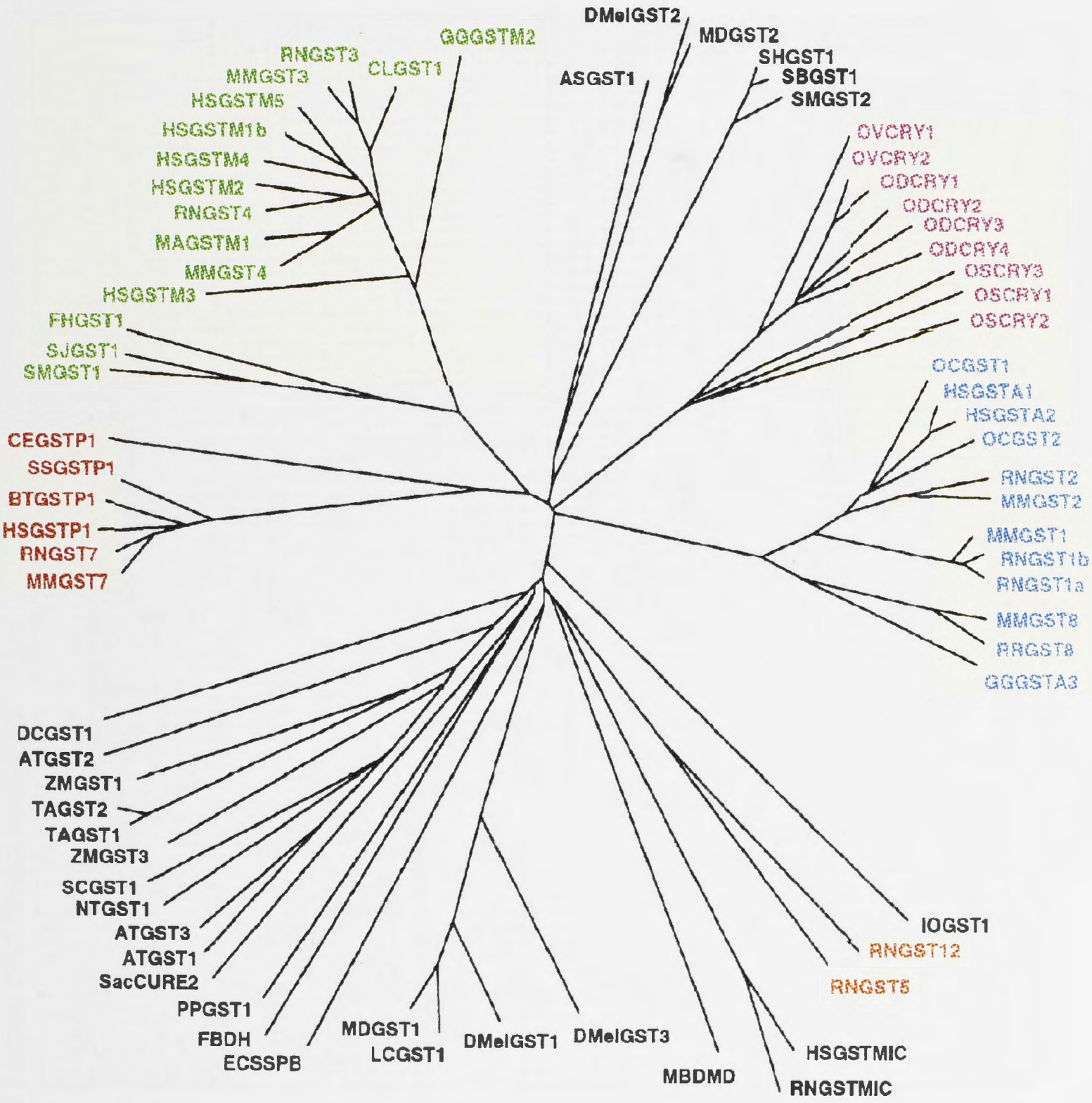
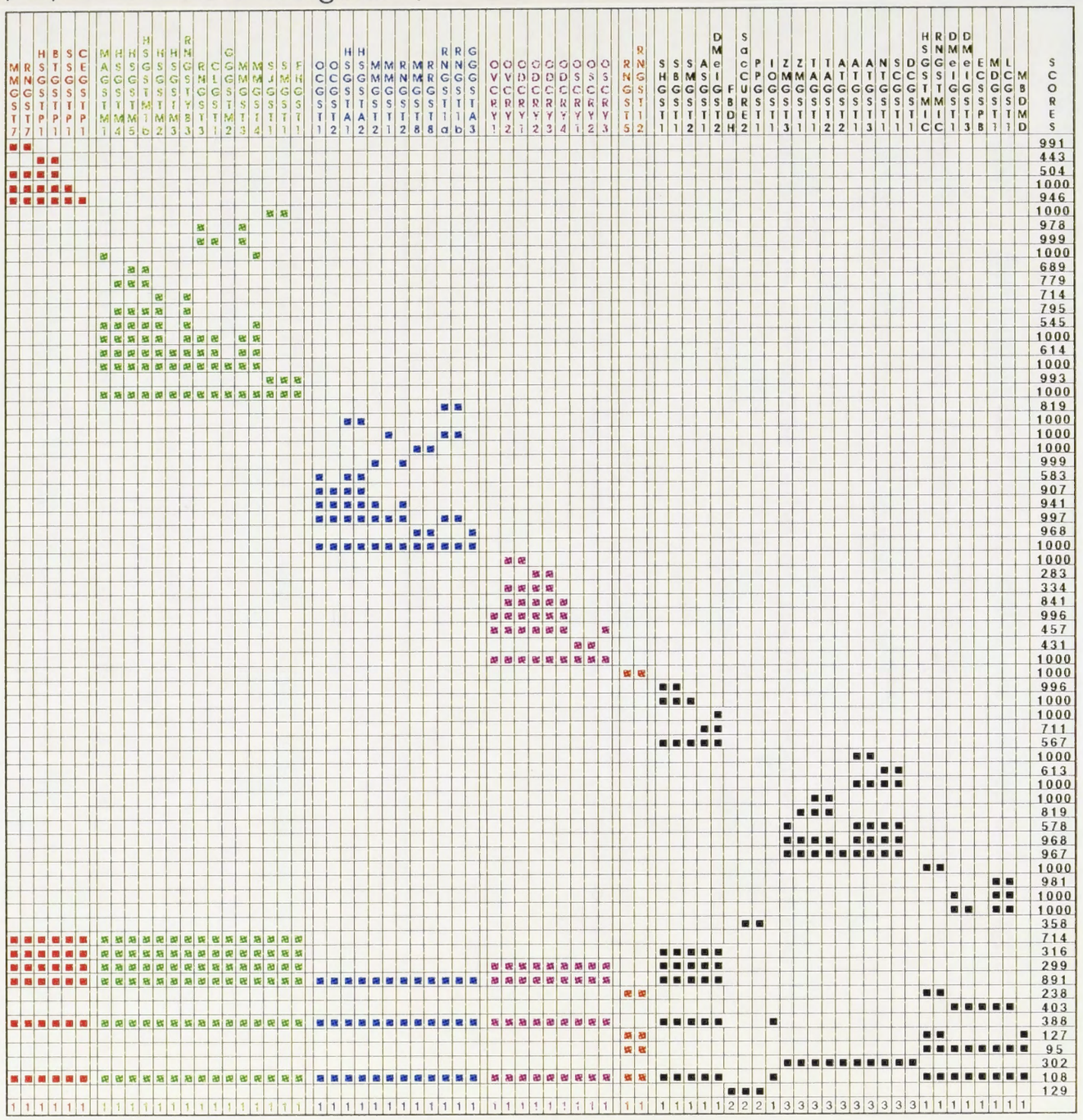


Figure 6.1b Phylogenetic tree of all GST sequences constructed by neighbor-joining. Alpha class is coloured blue, Mu is green, Pi is red, Sigma is purple and Theta is gold.



0.10

Table 6.4 Schematic of the construction of a NJ tree derived from comparisons of all GST sequences. The sequences joining at each node are represented by solid squares. Bootstrap values for each node in the tree are shown in the far right column. Alpha class sequences are blue, Mu are green, Pi are red, Sigma are purple and Theta are gold.



However, on the basis of the present alignment, comparison of distances between the 28kDa schistosome sequences and the crystallin sequences (all ~73% distance) are very close to those between the 28kDa sequences and the Alpha class sequences (again all ~73%). Additionally, since the proposal of the Sigma class, an *Ascaris* (parasitic nematode, ASGST1) sequence has been reported and single sequences from *D. melanogaster* (DMelGST2) and *Musca domestica* (MDGST2), all of which also cluster, albeit distantly, with the 28kDa proteins. The bootstrap confidence levels for the clustering of the schistosome/dipteran group are 57% and the confidence level for the joining of this branch to the crystallin branch is low, at only 30%.

TABLE 6.5 Percentage difference between representative rat amino acid sequences (uncorrected) from each of the Alpha, Mu, Pi and Theta classes illustrating the level of divergence within a species but between the classes.

		RNGSTA1	RNGST3	RNGST7	RNGST12
Alpha	RNGSTA1	0			
Mu	RNGST3	78	0		
Pi	RNGST7	70	68	0	
Theta	RNGST12	84	85	83	0

The topology of the NJ tree includes a clear internal branch with a fan-like array of sequences from each end of this branch (Figure 6.1b). When using two different collations of representative sequences from lineages in each of these arrays, bootstrap confidence levels for the internal branch responsible for this dichotomy are 100%. The topology of the MP tree is slightly different but an internal branch is maintained. However, in the MP tree a parasite sequence (*Ascaris*, ASGST1) joins directly onto the internal branch whereas this sequence is part of the crystallin lineage in the NJ tree. In both trees the three established classes, Alpha, Mu and Pi, along with the eye lens crystallins cluster at one end of the major internal branch, here referred to as sector A, whereas, Theta, the prokaryotes and the plant sequences are grouped at the other, sector B.

Sector A topology is similar in both trees and generally well supported by bootstrap confidence levels of the NJ tree. In contrast, there are some differences in the branching order of the two trees in sector B, for example, in the MP tree both yeast sequences, SacCURE2 (*Saccharomyces cerevisiae*) and IOGST1 (*Issatchenkis orientalis*), cluster together, whereas in the NJ tree they do not. In neither tree are they closely associated with any other sequences. Both show themselves to be equally distant from both plants (Minimum distance 73%) and animals (~73% from the dipteran group). SacCURE2 joins with the fish sequence, PPGST1 (*Pleuronectes platessa,* plaice) at the 36% confidence level and IOGST1 joins directly to the internal branch of tree at only 11%. In the MP tree both yeast sequences join the plant lineage.

Although sequences from sector B species cluster, they also show greater divergence than is evident within the Alpha, Mu and Pi classes. Multiple sequences from several plant species cluster together with differences of between 4% within a species to 72% between species (excluding comparison with the carnation sequence, DCGST1). There are three plant species with multiple GST sequences published, *Arabidopsis thalisinia, Zea mays (maize),* and *Triticum aestivum* (wheat). Representatives of each of these have differences in the same intra-class range as the mammalian classes. Some, however, also show greater similarity to sequences from other species than to those from the same species. For example, a maize (ZMGST3) sequence joins wheat sequences (TAGST1, TAGST2) at the 82% confidence level, whereas an *Arabidopsis* sequence, (ATGST2), diverges from the plant branch very early compared to either of the other two *Arabidopsis* sequences available (ATGST1, ATGST3). Estimated distances between the carnation (DCGST1), and other plant sequences are generally higher (~81%). A minimum distance of 70% with both an *Arabidopsis* (ATGST2) and a house fly sequence (MDGST1) has been calculated. Bootstrap confidence levels for the clustering of the other plant sequences

1 The microsomal GST sequences were included despite the lack of previous evidence for a shared common ancestor with the cytosolic enzymes to explore the possibility of such a relationship. Inclusion of the microsomal sequences does not affect results for the other sequences (data not shown).

are at 97% but the carnation sequence, DCGST1, is placed on the plant lineage at only the 30% level of confidence.

Several dipteran sequences have been reported that fall into two very distinct groups, one is placed in sector A, whereas the other is placed in sector B and is described here. The differences among the sequences in the sector B dipteran group are less than 36% and the confidence level for the clustering of this group is 100%. However, also joining this branch, but with a considerably higher level of divergence, is the stringent starvation protein of *E. coli*, (ECSSPB) with distances of ~74% from the dipteran sequences and a confidence level of 40%.

Other prokaryote sequences, represented by the *Methylbacterium* species di-methylhalogenase (MBDH) and FBDH, from *Flavobacterium*, have almost uniformly high distances from all other GSTs (~83%). Both also show less divergence from the Theta class (74% and 76%, respectively) but whereas MBDH joins with the Theta group in the NJ tree at only 13% confidence level, FBDH joins the fish/*Saccharomyces* branch, but also at only a 13% confidence level.

The single fish sequence included in this study, PPGST1 (*Pleuronectes platessa*, Plaice) has a high level of difference with all other GSTs (~82%) except an *Arabidopsis* sequence, ATGST2 (72%), and a dipteran, MDGST1 (76%). Buetler and Eaton (1993) refer to a report of the immunological cross-reactivity of a sequence from *Pleuronectes* to the Alpha class rat GST subunit 8. However, the single sequence available from the fish has ~83% divergence with the Alpha class. Quite possibly there are further GSTs to be isolated from this species.[1]

The distance between the two mammalian microsomal GST sequences currently available is estimated to be 16%. There was difficulty in aligning the microsomal sequences with the cytosolic GSTs, partially due to differences in length, 156 amino acids, compared to just over 200 for the cytosolic GST sequences. The alignment was constructed incorporating evidence from DeJong *et al.* (1988b) and Andersson *et al.* (1988) that N-

terminal residues 11-35 are a membrane spanning region and thus unlikely to appear in the cytosolic GSTs. Divergence from the cytosolic GST sequences is of the order of 84% and human microsomal GST shows slightly less divergence, overall, than does the rat form. The microsomal GSTs join the Theta class sequences in the NJ tree but this has a confidence level of only 24%.

6.4.3 Discussion

Although enzyme activity surveys across phyla have clearly demonstrated the presence of multiple GST enzymes in many species, the predominant interests of researchers in mammalian GSTs have strongly biased available sequence data. Thus the variety currently represented in the phylogenetic trees in this study are likely to significantly underestimate the complexity of the GST multigene family, at least in non-mammalian species.

Among the mammalian sequences, between class distances are quite large and very few amino acids can be found conserved throughout all mammalian sequences. This can be partially explained by the functions attributed to the enzymes, which, due to the broad range of hydrophobic substrates conjugated might be expected to have had some influence in the level of diversity among the sequences. However, the binding of GSH is a constant feature of all these enzymes and thus the binding site for this compound may be under some constraint. This analysis has found only one residue to be conserved throughout but there were two regions of increased similarity among the sequences. The conserved residue is Pro56 and is situated in the loop region between α2-helix and β3-strand. In the Alpha class, GSH lies antiparallel to this loop and forms hydrogen bonds with residue Val55. In addition, a further two residues within this loop, Met51 and Phe52 are involved in dimer interactions (Sinning *et al.* 1993). Proline is a unique residue in that considerable stability is provided by its ring structure and the lack of a proton to form other bonds. Proline can also exist as either a *cis*, as in HSGSTA1, or *trans* conformation (Sinning *et al.* 1993). It is possible that a proline

at the junction between this functionally important loop and the β–strand is necessary to maintain orientation of specific residues with respect to GSH and also in subunit interactions within the dimer.

Several residues that are known to be involved in the GSH-binding domain are widely, but not universally, conserved. The residue which is considered to "activate" GSH by donating a proton from its OH side group is Tyr9. In a limited number of sequences this residue is not conserved, being replaced in two plant sequences with His, rat microsomal GST with a Gln, the fish sequence, PPGST1, with a Trp and in five other sequences by Phe. None of these residues is capable of such a proton donation but it is possible that the Ser or Thr residues invariably situated nearby may be used for the proton donation. In the Alpha class HSGSTA1, site-directed mutagenesis of the Tyr9 residue reduced activity to only 10% but did not entirely abolish it. In this situation Arg15 was found to have been capable of partially fulfilling the role of proton donor (Björnestedt *et al.* (1994).

Two of the residues within the more conserved N-terminal region, Gln67 and Thr68, are known to be part of the GSH-binding site in mammalian enzymes. The "snail-trail" sequence region which is well conserved in this region includes these residues. Residues from this region are also involved in hydrogen bonding between the β–sheet and α–helices of domain one, while others, in addition to Pro56, are involved dimer interactions(Reinemer *et al.* 1991, 1992, Ji *et al.* 1992, Sinning *et al.* 1993). The second, C-terminal, region which exhibits a higher level of conservation contains residues that are also involved in the dimer interactions. All but one GST enzyme is reported to be dimeric and interactions between the monomers are important in stabilisation of the 3D structure. In each dimer two active sites are positioned at each end of the V-shaped crevice formed at the interface of the subunits.

Recently determined 3D structures in three GSTs have revealed a high degree of structural similarity (Reinemer *et al.* 1991, 1992, Ji *et al.* 1992, Sinning *et al.* 1993) that is not immediately obvious in a comparison of sequences across classes. There are now several studies indicating that 3D

structural conservation is not always evident in sequence comparisons (Lesk and Clothia 1980, Ollis *et al.* 1992, Cygler *et al.* 1993). In divergent sequences it is often the core rather than the surface residues which are more highly conserved (Lesk and Clothia 1980). In an extensive study of the relationship between sequence conservation and 3D structure in esterases and related proteins, Cygler*et al.* (1993) has provided evidence that conserved residues may include those involved in disulphide bridges and salt bridges that play a role in stabilising the 3D structure of the protein and in orientation of the substrate and its binding site. The most conserved of all in the esterase study (Cygler*et al.* 1993) were those residues sited at the edge of secondary structural elements, at the junctions of turns or loops with the edges of β-strands or α-helices. The residues most conserved among the GSTs are also those that are considered to be involved in structural stabilisation such as subunit domain, and also dimer, interactions. The most conserved of all residues in the current alignment (Pro56 in HSGSTA1) is situated on the junction of a loop and a β-strand (Sinning *et al.* 1993). Thus conservation of 3D structural framework may be more constrained than particular residues, except in specific instances, such as tight turns or difficult conformations.

Although there are some differences between the alignments derived from CLUSTALV, based on the sequences alone (data not shown), and from the manual alignment, incorporating evidence from the 3D structures of the mammalian enzymes , the resulting relationships among and within sector A groups show no apparent differences and there are only minor differences in the bootstrap confidence levels. However, there are considerable differences in some of the relationships within sector B. The plant sequences themselves (except for the carnation sequence, DCGST1) and those within the Theta and dipteran group are very stable in all methods of phylogenetic estimation but the relationships among these groups and the sequences from the yeasts, prokaryotes, the fish and the microsomal are not so robust. Neither the CLUSTALV nor the manual alignments allow distance

1 There are several distinct problems encountered when inferring very ancient evolutionary events from sequence data. For example, the number of possible residue substitutions at each site over the time frame postulated here may either obscure previous close relationships or, due to convergence, it is possible that more recent similarities may be suggested (see sections 6.3.4 and 6.3.7). Secondly each construction of a phylogenetic tree must, of necessity, be only one interpretation of the given data. The high level of sequence divergence and low bootstrap values for many of these sequences joining the tree around the central branch must impose a cautionary note on any conclusions drawn.

estimates which show bootstrap confidence of even medium levels for the relationships among these deep rooted branches.

The topology of the trees constructed from the alignment of all GSTs shows a clear dichotomy formed by an internal branch. And this was also reflected in the CLUSTALV alignment. Confidence levels for the internal branch producing this dichotomy are 100%.[1] The robustness of the internal branch resulting in the two fan-like sectors, A and B, suggests a very early duplication of an ancestral gene. This may have occurred before the radiation of the eukaryotes and created at least two paralogous genes in the descendant lineages. Each of these hypothetical A and B genes has undergone successive duplications. Both A and B genes are represented in several species, thus mammalian Theta sequences are found in sector B, whereas the majority of known mammalian sequences belong to sector A. There are, as yet, no plant representatives in sector A, however, dipteran sequences are represented in both domains. This may have been in response to the expanding needs of cells as metabolic pathways became more complex and environmental sources of electrophilic compounds proliferated. Rat and dipteran sequences are represented in both domains and the publication of further sequences will probably expand the number of species represented in both domains. It is possible that the inclusion in the alignment of further sequences from species more closely related to the isolated sequences such as the yeast and fish sequences will resolve their relationships with more certainty.

The plant kingdom is represented by relatively few GST sequences at this time although there are three species with multiple sequences. The degree of divergence of some of these sequences is similar to that between the different classes within the mammals. In addition, sequences from different species show more similarity to each other than to sequences of the same species. Thus it is possible these sequences may be representatives of a class system analogous to the Alpha, Mu and Pi classes. The higher level of divergence exhibited by the carnation sequence,

DCGST1, suggests either that in performing its specific role in senescence, this form of carnation GST may be specialised or, alternatively, that this is currently the sole representative of a more divergent class of GSTs in plants. Availability of additional sequences may clarify which is the more likely.

The level of divergence among the sector B dipteran group is well within the intra-class range but available insect sequences are confined to this order. Toung *et al.* (1990) have suggested a class system in insects analogous to that characterised in mammals. Evidence in support of this suggestion is provided by the present study, as dipteran sequences also exhibit sector A/B dichotomy. Interestingly, the level of similarity of the sector B dipteran sequences with the stringent starvation protein of *E. coli* is is within the range of inter-class differences, despite the long history of divergence between *E. coli* and dipterans.

Traditional classification of fungi (including yeasts) as plants was replaced by the view formalised in the five kingdom system proposed by Whittaker and Margulis (1978) in which fungi were given the status of its own kingdom. A more recent proposal, gaining support, considers fungi more closely related to the animal lineage (Cavalier-Smith 1987, Baldauf and Palmer 1993). In the NJ tree placement of yeast sequences is ambivalent and the level of confidence is low. *Saccharomyces* is placed on the branch leading to plants whereas *Issatchenkis* is on a branch of its own. In the MP tree, however, these sequences cluster together and join the plant branch, albeit with a very deep root. This may be an artefact of the lack of sequences from more closely related species and may be resolved in future.

The level of similarity between microsomal and cytosolic GSTs is not high and there is no significant evidence in this analysis of a relationship between the cytosolic and microsomal GST sequences. A clearer view of any possible relationship must await either the solution of the 3D structure of a microsomal form or more microsomal sequences becoming available. The inclusion of the microsomal sequences, while it does not provide evidence for a common ancestor with the cytosolic GSTs also does not materially affect the results of the comparison among the cytosolic sequences.

The categorisation of the Sigma class of Buetler and Eaton (1992) is supported by the evidence in this study, not, however, with the inclusion of the

28kDa schistosome sequences. The Alpha Mu and Pi classes all contain sequences from species belonging to at least different taxonomic orders and this is also true of the Sigma class. However, using the present definition of a class these schistosome sequences (and the dipteran sequences associated with them) are on the extreme margin of inter-class differences. Lens crystallins are not specifically required for their catalytic activity, instead the characteristics of most importance are optical clarity and stability over time. The level of catalytic activity in the proteins encoded by some of these sequences is quite low (Tomarev *et al.* 1992) reflecting this altered function. Presumably there are metabolic GSTs in these species still to be characterised. It will be of interest to see the relationships of these metabolic GSTs with mammalian GSTs and also with the crystallin genes and these may help resolve the status of the schistosome/dipteran branch relationship with the Sigma class.

From their phylogenetic study, Buetler and Eaton (1992) suggested that the majority of non-mammalian sequences may belong to the Theta class,this suggestion was also made by Pemble and Taylor (1992). However, the present analysis does not support this view. The distances estimated from the alignment are considerably higher than that seen in the other classes. There is also no evidence in the topology of either the NJ or MP trees to suggest these sequences are related at the class level. Pemble and Taylor (1992) also suggested that the sequences in this broader Theta class were of an ancient progenitor gene, however, data from the present study indicate that sequences in both of the sectors delineated here have very ancient origins. The limited evidence available makes it difficult at this time to suggest which of the sequences may be such a representative, or if any are.

6.5 EVOLUTION WITHIN THE GST CLASSES

For each of the three classes, trees were constructed using NJ, MP, branch and bound parsimony (DNApenny) and maximum likelihood algorithms (data not shown) using the alignments listed in Appendices II

(amino acid sequences) and III (nucleotide sequences). Both MP and NJ trees were constructed in two different ways. Either by directly constructing a single NJ or MP tree or by the creation of a series of bootstrap (random sampling with replacement) alignments, each of which was used to construct a single tree and subsequently a consensus tree. The topology of the different types of trees did not differ in any substantial way although some differences in the lengths of the branches was evident. In using multiple sequence sets produced by bootstrapping, distances represented by the branches have been increased in both the NJ and MP consensus trees. This has occurred in the internal branches more so than in the terminal branches. Parsimony trees are not expected to produce accurate estimations of branch lengths. The NJ consensus trees derived from the PHYLIP bootstrap method resembles the parsimony trees. The maximum likelihood trees, on the other hand, resemble more the single alignment NJ trees of their respective groups and these more accurately reflect the distance estimations. The NJ single alignment trees derived from the amino acid sequences are shown in Figure 6.2 and the NJ single alignment trees derived from the nucleotide sequences are shown in Figure 6.3. Bootstrap values are shown at the nodes on the trees.

Distances in intra-class comparisons for the NJ trees are sufficiently low as to allow correction for multiple substitution at a single site (Kimura 1980). Hence corrected distance scores for these comparisons are used here and tabulated in Table 6.6, for amino acid sequences, and Table 6.7 for nucleotide sequences. Correction for multiple substitution has the effect of increasing distance in the more divergent comparisons. Thus distances estimated for a few comparisons, once corrected, are in excess of 100%.

6.5.1 Evidence from amino acid sequences

Within the Alpha, Mu and Pi classes distances among mammalian amino acid sequences can be as little as 5%. Previous estimations of less than 50% divergence as a criterion for class membership are generally confirmed.

Table 6.6 Estimated distances among GST amino acid sequences, calculated using the method of Kimura (1983). **(a)** Alpha class. **(b)** Mu class. **(c)** Pi class.

(a) Alpha class

GGGSTA3	0										
OCGST1	52	0									
OCGST2	51	26	0								
HSGSTA1	49	21	23	0							
HSGSTA2	50	26	25	5	0						
MMGST2	53	30	30	28	31	0					
MMGST1	55	37	37	28	31	37	0				
RNGST2	53	29	30	30	33	16	39	0			
MMGST8	40	60	53	54	55	58	56	59	0		
RRGST8	42	67	56	56	57	59	57	59	10	0	
RNGST1a	57	40	39	29	33	40	5	42	59	57	0
RNGST1b	59	37	37	27	31	40	6	42	59	58	4	0

(b) Mu class

FHGST1	0															
MAGSTM1	88	0														
HSGSTM4	86	24	0													
HSGSTM5	86	22	19	0												
HSGSTM1b	83	22	14	14	0											
HSGSTM2	93	27	19	21	18	0										
HSGSTM3	91	40	35	36	35	41	0									
RNGST4	93	22	21	19	20	17	41	0								
RNGST3	87	25	32	26	25	29	43	26	0							
CLGST1	87	24	30	26	24	26	44	25	12	0						
GGGSTM2	90	43	43	43	46	48	46	44	55	53	0					
MMGST3	90	24	34	28	27	30	45	27	7	12	53	0				
MMGST4	87	12	22	21	19	25	36	22	22	22	48	24	0			
SJGST1	61	99	94	94	99	94	99	100	96	99	100	96	99	0		
SMGST1	63	114	112	108	105	108	105	108	105	107	107	107	112	24	0	

(c) Pi class

CEGSTP1	0					
MMGST7	102	0				
HSGSTP1	110	16	0			
BTGSTP1	112	16	15	0		
RNGST7	109	8	16	16	0	
SSGSTP1	104	16	19	15	17	0

Table 6.7 Estimated distances among GST nucleotide sequences, calculated using the 2-parameter method of Kimura (1983). **(a)** Alpha class. **(b)** Mu class. **(c)** Pi class.

(a) Alpha class

GGGSTA3	0											
OCGSTA1	46	0										
OCGSTA2	42	19	0									
MMGST2	45	26	25	0								
MMGST8	40	49	47	45	0							
RRGST8	39	52	49	46	10	0						
HSGSTA1	45	16	18	27	46	45	0					
HSGSTA2	46	17	18	28	47	46	4	0				
MMGST1	45	25	28	31	45	45	22	23	0			
RNGST1a	46	28	31	33	47	45	24	26	6	0		
RNGST1b	46	27	30	32	47	45	23	25	6	2	0	
RNGST2	44	23	22	12	49	48	23	24	28	30	31	0

b) Mu class

FHGST1	0													
SJGST1	58	0												
SMGST1	56	22	0											
MAGSTM1	77	80	87	0										
HSGSTM2	70	73	81	24	0									
CLGST1	70	80	85	17	18	0								
HSGSTM1b	70	79	80	20	11	16	0							
HSGSTM3	71	80	83	37	35	33	32	0						
MMGST3	77	82	89	20	22	9	20	37	0					
RNGST3	72	81	87	20	21	9	19	38	6	0				
RNGST4	73	75	81	20	16	17	18	37	19	18	0			
HSGSTM4	72	76	82	20	12	19	8	31	22	23	18	0		
HSGSTM5	72	76	81	21	13	19	10	33	22	21	17	13	0	
GGGSTM2	82	100	100	40	45	41	44	45	43	43	40	41	44	0
MMGST4	73	81	90	10	19	13	16	33	16	16	18	18	18	43	0

(c) Pi class

CEGSTP1	0					
BTGSTP1	77	0				
MMGST7	78	20	0			
RNGST7	80	20	7	0		
RNGSTPPS	84	22	8	6	0	
HSGSTP1	76	13	19	20	22	0

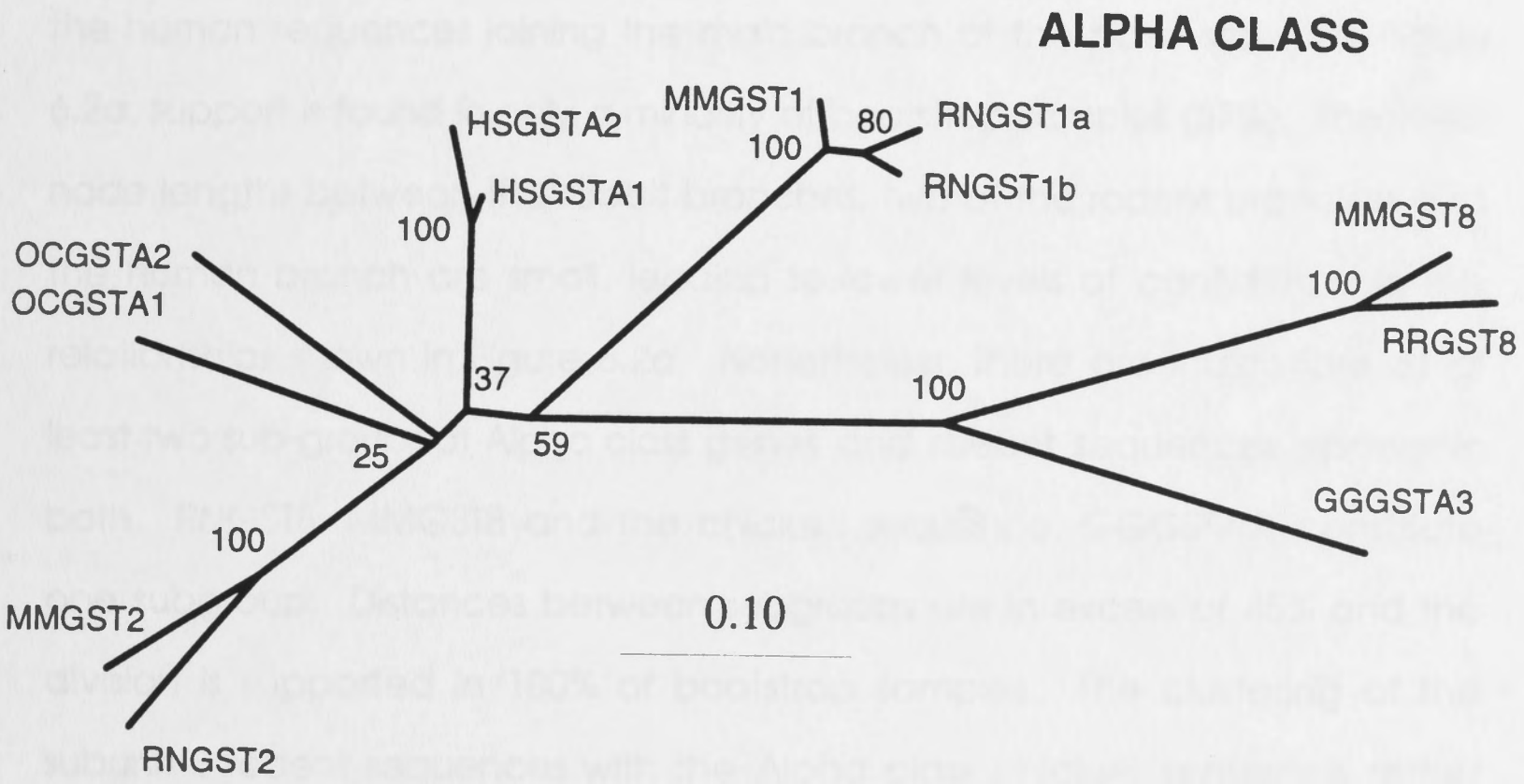
Inclusion of the Theta class increases the level of divergence within a class, as the two known sequences for this class have a corrected distance of 68%. Between classes, uncorrected sequence differences are between 65% and 80%, the Theta class being the most divergent with differences of 76%-83%. Actual distances would be greater due to multiple substitutions at a single site, correction by a method such as Kimura's (1980), adjust for this but cannot be used in cases such as inter-class comparisons of GSTs, where levels of divergence are high. Table 6.5 shows the percentage difference (uncorrected) between the mammalian classes for representative sequences. Inclusion in the alignment of sequences other than those from mammalian species has meant that the percentage differences within the classes generally have a broader range. Some of these distance estimates are higher, when corrected, than those previously used to limit class membership (Mannervik *et al.* 1992).

The branching order found in all types of trees is represented by the NJ trees in Figure 6.2, the numbers at the branches in the trees indicate the percentage of bootstrap samples (of 1000 samples) where the indicated branch or cluster appeared.

In the Alpha class most of the terminal branches are found with a high probability in the bootstrap sampling, however, the relationships among the more deeply rooted branches are not as clearly defined. In the case of the human sequences joining the main branch at the point shown in Figure 6.2a, support is found in only a minority of bootstrap samples (37%). The inter-node lengths between the rabbit branches, two of the rodent branches and the human branch are small, leading to lower levels of confidence in the relationships shown in Figure 6.2a. Nonetheless, there are indications of at least two sub-groups of Alpha class genes and rodent sequences appear in both. RNGST8, MMGST8 and the chicken sequence, GGGSTA3, constitute one subgroup. Distances between subgroups are in excess of 45% and the division is supported in 100% of bootstrap samples.

Figure 6.2 Neighbor-joining trees constructed from corrected distances between amino acid sequences from the Alpha, Mu and Pi class GSTs. Bootstrap values are shown at each node.

Bootstrap values of the final pair are always 100% and are not printed.



In the Alpha class most of the terminal branches are found with a high probability in the bootstrap sampling, however, the relationships among the more deeply rooted branches are not as clearly defined. In the case of the human sequences joining the main branch at the point shown in Figure 6.2a, support is found in only a minority of bootstrap samples (37%). The inter-node lengths between the rabbit branches, two of the rodent branches and the human branch are small, leading to lower levels of confidence in the relationships shown in Figure 6.2a. Nonetheless, there are indications of at least two sub-groups of Alpha class genes and rodent sequences appear in both. RNGST8, MMGST8 and the chicken sequence, GGGSTA3, constitute one subgroup. Distances between subgroups are in excess of 45% and the division is supported in 100% of bootstrap samples. The clustering of the subunit 8 rodent sequences with the Alpha class chicken sequence rather than with any other rodent, or even other mammalian, sequences provides the only difference in a tree that is broadly consistent with accepted systematics. The position of lagomorphs (rabbits) compared to primates or rodents is somewhat ambivalent in the literature and this is again reflected in the Alpha class sequences. The two rabbit sequences appear no more closely related to each other than they are to either of the human sequences. The tree constructed from amino acid distances suggests a marginally closer relationship between rabbits and the primate sequences available (human) compared to the nearest rodent sequences, but this is not strongly supported in the bootstrap sampling.

The Mu class currently contains representatives of a wider variety of mammalian species than the other classes and also includes several non-vertebrate species (Figure 6.2b). Unlike the Alpha class the majority of branches are found in a high percentage of bootstrap samples, except for the human sequences. This generally reflects the high level of similarity among the human sequences. All but one human sequence, HSGSTM3, cluster together, suggesting the possibility of relatively recent duplication events. Rat subunit 4 is also included on this branch but only in 76% of

1 Thus previous suggestions (section 1.3 and 6.1.3)that class variability could be encompassed within 50% similarity is generally confirmed for mammalian sequences but may need more flexibility with the inclusion of non-mammalian sequences.

The Pi class is unique among the GST classes as only one functioning representative enzyme has been characterised for each species (Figure 6.2c). The other classes, even the recently identified Theta class, generally appear to have multiple enzymes in each species. When comparing amino acid sequences in the Pi class it is clear that there is a high level of divergence between the *C. elegans,* CEGST1, sequence and the remaining Pi class sequences, currently all mammalian. All other comparisons within the class provide divergence estimates below 20%. Nonetheless, CEGST1 clusters with the Pi class at a very high confidence level (100%) in the major tree.[1] Pig (*Sus scrofa*, SSGSTP1) is included in this comparison, but not in the nucleotide comparison because only the amino acid sequence is known at present. The inter-node distances in this tree are small among the mammalian sequences, creating low bootstrap sampling scores for these relationships. Within the Pi class, the cattle (*Bos taurus*, BTGSTP1) sequence clusters with the human sequence rather than with the other artiodactyl, pig. In the construction of this tree, the pig sequence joins the nematode sequences before joining with any other mammalian, but only at 51% confidence level. Both rodent sequences also intervene between the two artiodactyl sequences, but again only at the 59% level.

There are as yet only two mammalian sequences available for the Theta class and so a comparison is not possible.

Sinning *et al.* (1993) have reported 27 residues conserved in HSGSTA1, SSGSTP1 and RNGST3 when the 3D structures are taken into account. Other residues in each of the mammalian classes have also been implicated in GSH-binding but these are not conserved across the three classes. Inclusion in an alignment of all the currently available mammalian sequences in these three classes decreases the number of conserved residues to 19, as in some cases conservation is not evident even within the same class as the enzyme whose structure has been solved. If the Theta class mammalian sequences are also added, the number of conserved residues decreases to 10. This number decreases further, to 8 residues, if sequences in the extended

mammalian classes are used. Of these 8 residues only 3 are implicated in GSH binding by the crystallographic data. However, if the Theta class is removed from the comparison there are 13 residues conserved in the extended classes. This lends further support for the suggestion of a pre-animal gene duplication as mammalian sequences in sector A appear to have higher levels of conservation of amino acid residues with sequences from non-mammalian species than when compared to sector B mammalian species.

6.5.2 Evidence from nucleotide sequences

The level of divergence within the Alpha, Mu and Pi classes is such that it is possible to align and compare the nucleotide as well as the amino acid sequences. Kimura's 2-parameter correction for multiple substitutions was used and corrected distances are shown in Table 6.7

As with the amino acid sequences comparisons, the Alpha class nucleotide sequences exhibit the broadest range of divergence levels and the relationships within the class are generally unaltered. Again two subgroups within the Alpha class are apparent and the level of confidence between these is 100% (Figure 6.3a). However, the relationships of the two rabbit sequences are problematic. In the tree based on the nucleotide sequences both rabbit sequences join the branch leading to humans, with a somewhat increased level of confidence, compared to the tree based on amino acid sequences.

The Mu class contains the largest number of sequences because of the recent publication of four new human cDNAs. The branching pattern evident among the large number of human sequences, all with high levels of similarity do not necessarily show correspondingly high levels of confidence (Figure 6.3b). Nonetheless all human sequences cluster on the one branch, it is the order of branching within this group that exhibits low levels of confidence. On several of the mammalian sub-branches in this class multiple species are represented, suggesting that at least one of the duplication

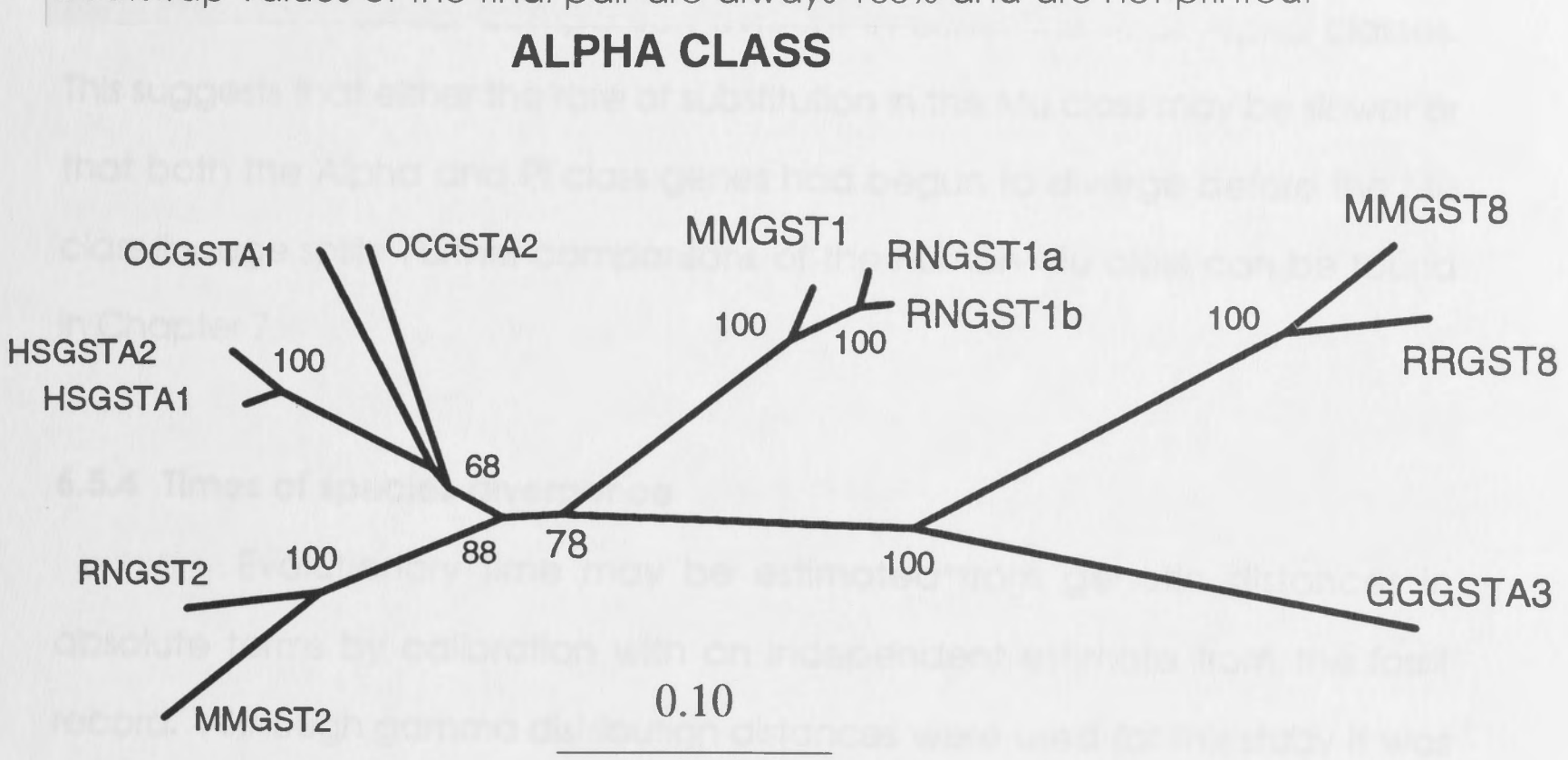
events giving rise to these sequences, the one that gave rise to the ancestor of RNGST4, occurred prior to the mammalian radiation. The other duplication events are represented only by rodent sequences at present. HSGSTM3 is again on a distant branch of its own, however, the chicken sequence is more closely related to the remaining mammalian sequences than HSGSTM3 in this tree, the reverse of the relationship evident in the tree based on amino acid sequences. This is a reflection of the short internode distance between GGGSTM2 and HSGSTM3.

A pseudogene sequence, RNGSTPPS, is included in the Pi class nucleotide alignment, it was not included in the amino acid comparison as frame shift mutations prevent meaningful translation. Despite these mutations, divergence between the pseudogene and the functioning rat Pi class sequence is only 6%, suggesting a relatively recent origin. In the tree based on nucleotide sequences, the branching patterns shown in Figure 6.3c among the rodent sequences are found in a high percentage of bootstrap samples. The clustering of the cattle with the human sequence lends some support to the view that artiodactyls are a more closely related group to primates than are rodents, a contentious point in systematics. The pig nucleotide sequence is currently unknown.

6.5.3 Relationship of GSTM4 to other GSTs

Characterisation and hybridization studies indicated the existence of an extensive range of human GSTs and led to the identification of a novel and alternatively spliced GST (Chapter 3). Amino acid sequence comparison of the novel GSTM4 sequence reported in Chapter 4 and with the 72 other GST sequences included in this study, show that GSTM4 has a high level of similarity with human GSTM1, GSTM2 and GSTM5, clearly placing it within the Mu class. Corrected distances among these human Mu class amino acid sequences range from 13% to 17%. When nucleotide sequences are compared, distances decrease between GSTM1, GSTM2, GSTM4 and GSTM5 (to between 8% and 14%).

Figure 6.3 Neighbor-joining trees constructed from corrected distances between nucleotide sequences from the Alpha, Mu and Pi class GSTs. Bootstrap values are shown at each node.
Bootstrap values of the final pair are always 100% and are not printed.



Distances between GSTM4 and the rat sequence, RNGST4, and its mouse counterpart, MMGST4, are only 17% and 18%, a considerably lower level than for a similar comparison evident in either the Pi or Alpha classes. This suggests that either the rate of substitution in the Mu class may be slower or that both the Alpha and Pi class genes had begun to diverge before the Mu class lineage split. Further comparisons of the human Mu class can be found in Chapter 7.

6.5.4 Times of species divergence

Evolutionary time may be estimated from genetic distances in absolute terms by calibration with an independent estimate from the fossil record. Although gamma distribution distances were used for this study it was found that there was very little difference between gamma and Kimura 2-parameter distances, which are shown in Table 6.7. Both Alpha and Mu classes contain chicken sequences and the minimum distance between these and any mammalian sequences within both classes is 40%. Divergence of mammal-like reptiles from the bird-like reptile lineage is thought to have occurred much earlier, at approximately 300 Ma ago, as inferred from the fossil record (Carrol 1988).

Using the bird/mammal split to calibrate estimations (section 6.3.7) it can be estimated when the Mu class Platyhelminth lineage (Schistosomes, SJGST1 and SMGST1) diverged from the mammalian lineage. Minimum distances are 73% and 80%, respectively, providing estimates of 553 ±83 Ma ago and 607 ±92 Ma (millions of years) ago. Although there is no fossil record of this phylum for comparison, this estimate does not conflict with accepted views. In the Mu class there is also a nematode sequence, from *Fasciola* (FHGST1), which has a similar estimated genetic distance (70%) and, hence, a divergence time of 534 ±81 Ma ago. In the Pi class, another nematode sequence, from *C. elegans*, provides a distance of 76% between it and any mammal sequence, placing it within a similar time frame of the divergences within the Mu class (577 ±87 Ma ago). Fossil records of these animals are very

sketchy indeed, however, there are specimens from a class of the *Nematoda*, found in Lower Carboniferous sediments (Laverack and Dando 1979), providing a minimum age of approximately 350 Ma ago.

The separation of the rodent and primate lineages is commonly assumed to have been during the late Cretaceous, about 80 Ma ago (Carrol 1988). Minimum distances between human and rodent sequences in the Alpha class are 22% between HSGSTA1 and MMGST1 or 23% between HSGSTA1 and RNGST1b; in the Mu class, 16% between HSGSTM1b and MMGST4 or HSGSTM2 and RNGST4; and in the Pi class 19% and 20%. This results in rodent/primate divergence dates of between 165 ±28 to 176 ±30 Ma ago in the Alpha class, 122 ±23 Ma ago in the Mu class and between 147±26 and 153 ±27 Ma ago in the Pi class. The differences between the classes is encompassed by the standard errors quoted, but all show a trend towards dates earlier than estimated in other studies (Easteal 1990, Li *et al.* 1990).

The time of divergence between rat and mouse lineages is a long-standing controversy fuelled by the paucity of fossil records, but by this means is estimated at 8-14 Ma ago (Jacobs and Pilbeam 1980) and by molecular data to be between 20 to 35 Ma (Wilson 1987) or, more recently, 20 to 29 Ma ago (O'hUigin and Li 1992, Chevret *et al.* 1993). The minimum estimated distance is between rat, RRGST3, and mouse, MMGST3 sequences and is 6% in both the Alpha and Mu classes and 7% in the Pi class, suggesting a divergence date of 46 ±12 to 53 ±12 Ma ago. These estimates agree with each other, within the standard errors and the Alpha and Mu class dates are also within the range of estimates indicated by previous molecular studies, although they show a similar trend for early dates as seen in the rodent/primate data above.

In the Alpha class the two rabbit sequences have a lower level of divergence from human sequences (minimum 16%) compared with rodent sequences (minimum 22%), indicating a human/rabbit divergence of 120 ±22 Ma ago in the Mu class. The lagomorph/rodent lineage split is estimated at a minimum of 170 ±29 Ma ago, a similar estimate to the primate/rodent split in the

Alpha class reported above. This suggests that the rodent lineage may have separated from the human/lagomorph lineage very close to the time of the lagomorph/primate split. The dates estimated in this study for the more recent divergences, although consistent are significantly older than those proposed from the fossil record (Carrol 1888). Other molecular studies have suggested somewhat earlier dates than the fossil record (eg Wilson *et al.* 1977, Wu and Li 1985, Wolfe *et al.* 1989, Li *et al.* 1990), but not as early as those consistently derived from this analysis.

6.5.4.1 Relative rate tests

Using sequence data as the basis for the estimation of times of divergence is only possible because of the assumption of evolutionary rate constancy among lineages. Evolutionary rates were compared using the relative rate test of Wu and Li (1985). Controversy over the divergence times and relationship between rabbits and humans led to the choice of sequences for comparison from the Alpha class, using a rat sequence as reference.

The analysis was repeated for rodent and human sequences from the Mu class, in this case the highly divergent human sequence, HSGSTM3 and the chicken sequence, GGGSTM2 were used for reference. Results are shown in Table 6.8.

The substitution rate per year (K) can vary extensively between genes (O'hUigin and Li 1992) but is assumed to be constant in comparisons of paralogous genes. The differences in relative rates between the sequences tested, rabbits and humans in the Alpha class, and mouse and humans in the Mu class, are minimal and not significant. The rate for nondegenerate sites is 13% between human and rodent in the Mu class, irrespective of the use of rat or chicken sequence as reference. Between rabbit and human in the Alpha class, the relative rate at non-degenerate (nonsynonymous changes) sites is 12%.

Table 6.8 Differences in the number of nucleotide substitutions/100 sites between rabbits (OC), rats (RN), mice (MM) and humans (HS).

				Non-Degenerate Sites				4-fold-Degenerate			
CLASS	SPECIES 1	SPECIES 2	OUTGROUP SPECIES 3	length	K12	K13-K23	S.E.	length	K12	K13-K23	S.E.
Alpha	HSGSTA1	OCGSTA1	RNGST2	418	0.10299	0.01101	0.01602	86	0.42748	-0.0356	0.07204
Alpha	HSGSTA2	OCGSTA1	RNGST2	418	0.12789	0.01974	0.01527	86	0.40315	-0.0483	0.07309
Alpha	HSGSTA1	OCGSTA2	RNGST2	421	0.11882	0.00199	0.01567	89	0.46421	0.03437	0.06295
Mu	HSGSTM4	RNGST4	HSGSTM3	416	0.11472	0.03519	0.01618	82.5	0.38486	0	0.06669
Mu	HSGSTM2	RNGST4	HSGSTM3	416	0.08882	0.01096	0.01625	82.5	0.36678	0.00429	0.06862
Mu	HSGSTM4	MMGSTM4	HSGSTM3	417.5	0.1227	-0.0357	0.01953	77	0.45078	-0.1193	0.12041
Mu	HSGSTM2	MMGSTM4	HSGSTM3	417.5	0.13577	0.00271	0.02054	77	0.4433	0.01249	0.13375
Mu	HSGSTM2	MMGSTM4	GGGSTM2	413.5	0.13577	0.01014	0.02371	89	0.4433	0.02192	0.24942

length = number of sites. K 12 = distance between spp. 1 and 2, K13 = distance between spp. 1 and 3. K23 = distance between spp. 2 and 3. S.E. =standard errors.

The rates at 4-fold-degenerate (synonymous changes) sites are 42% to 46% between rabbit and human and 37% to 45% between human and mouse. Thus in both comparisons, for non-degenerate and for 4-fold-degenerate sites, the data indicates the relative rates of evolution in two of the classes of GSTs is very similar.

6.5.5 Discussion

Nj trees of each of the three classes have been constructed in an attempt to investigate the phylogenetic relationships within each of the classes. Single alignment NJ trees were used in preference to the other methods trialed as relationships were not significantly different and these provided the most accurate representation of the distance data available.

Previous studies have suggested that although the Alpha Mu and Pi classes were originally defined in mammals they are not exclusively mammalian and include birds (chicken) and several genera of invertebrates (Buetler and Eaton 1993). Such a proposition is clearly supported by these studies. The clustering of the platyhelminth and schistosome sequences within the mammalian classes indicates that the divergence of the classes significantly predates the vertebrate radiation. This proposition is supported by the inclusion of *C. elegans*, a free living nematode, into the Pi class, with a high level of confidence. The evolution of GSTs in this species are unlikely to have been influenced by the complexities of the mammalian host/parasite interactions that, it could be argued, may have played a role in the evolution of the parasitic forms.

Alpha class sequences, chicken GGGSTA3, rat RNGST8, and mouse MMGST8, constitute a sub-group within the class and may represent an early duplication of the ancestral Alpha gene, apparently occurring before the bird/mammal split. Interestingly, the two rabbit sequences, OCGST1 and 2, in the Alpha class have been placed on the human lineage, at the 68% confidence level in the nucleotide comparison, rather than with any of the rodent sequences. Traditional taxonomy included lagomorphs and rodents

in the same order but separation of these two has been accepted since early this century (Romer 1966). More recently Van Valen (1971), on morphometric grounds, suggested that lagomorphs may be more closely related to primates than to rodents. Genetic sequence analysis has also tended to support the closer relationship of primates and lagomorphs (Easteal 1990, Long and Gillespie 1991). The GST data presented here supports the latter view.

As in the Alpha class, each species apparently requires multiple genes of the Mu class. Rats and mice are represented by two sequences in the Mu class and the topology of the tree (Figure 6.2b) suggests that the initial duplications occurred at least prior to the rodent/human split, as each is associated more closely with sequences from other species rather than with the alternative sequence from the same species. Each of the hamster genera is represented by a single sequence at present, it is quite possible other Mu class genes occur in these genera and have yet to be characterised.

Within the Mu class the relationships among the large number of human sequences is problematical. Although work on rats and mice has been widespread there may be more rodent genes to be characterised. The appearance at present is that there have been a series of duplication events apparently confined to humans. Characterisation of GST sequences from other primates, particularly our close relatives, chimpanzees (*Pan*) and gorillas (*Gorilla*), may clarify this. HSGSTM3 is significantly more distant from the other mammalian sequences but there is as yet no paralogous gene known from the rodents. HSGSTM3 is a similar distance from the Mu class chicken sequence than is evident between the Alpha class chicken sequence and the nearest mammalian sequences. These Alpha sequences appear to constitute a sub-group within the class. It is therefore possible there may be a parallel between this subgroup in Alpha and the chicken/HSGSTM3 in Mu. An investigation to find human Alpha class genes of the chicken subgroup-type

and possible Mu class genes in rodents of the HSGSTM3-type may be illuminating.

In the Pi class the relationships between the pig amino acid sequence and other mammalian sequences is also anomalous. Cattle and pigs are both artiodactyls and most published genetic evidence contradicts some palaeontological phylogenies by placing rodents as more distant relatives of primates than are artiodactyls (Easteal 1990, Bulmer 1991, Honeycutt 1993). However, the Pi class GST sequences available suggest that cattle are more closely related to humans than to pigs, thus splitting the artiodactyl order (Figure 6.2c). Rodents also appear to intervene between the two artiodactyl sequences. Support for these relationships is not high due to the similarities among all mammalian sequences in this class. In any other class of GSTs it would be reasonable to suggest that the pig sequence may be a paralogue rather than a homolog, however, there is little evidence for multiple genes within a species in the Pi class. Inclusion of sequences from other artiodactyls and the availability of the pig cDNA sequence may resolve these relationships with more certainty. It may be more suitable to look at the artiodactyl/human/rodent relationships in another class where the differences may be a little greater.

The comparison between the *C. elegans* and the mammalian sequences is also of note. *C. elegans* is a free living nematode, whereas in the Mu class the non-vertebrate sequences are all from parasitic species. The parasitic life style is thought to affect evolutionary processes, due to the necessity of co-evolution of host and parasite. However, the distances estimated between the parasitic nematode and mammalian sequences from the Mu class and the free living form and mammalian sequences in the Pi class are very similar.

The presence of chicken sequences clearly belonging to both the Alpha and Mu classes and the nematode and schistosome species from Mu and Pi classes provides a strong indication that the origins of these classes is very ancient indeed, probably predating the animal radiation.

Some of the problems with the use of molecular data to provide absolute dates of divergence concern relative evolutionary rates and are reviewed Britten (1986), Easteal (1992). The fossil record itself, however, is not without its limitations. It is the subject of regular, if infrequent, reinterpretation. It can, at best, only give a minimum time for the emergence of a species, the oldest fossil recorded for any one species is unlikely to have been the progenitor, or even one of the early representatives of the species in question. Finally, divergence of species in the fossil record is limited to the appearance of morphological differences, which probably followed behavioural, genetic and biochemical species differences by a significant span of time. Nonetheless, the independence of the fossil record makes it a valuable tool for the estimation of evolutionary time.

Dates calculated in this study rely on calibration with the fossil record. The bird/mammal split is a very ancient one and regular reinterpretations of the fossil record occurs. Necessarily the older the record of interest, generally the less accurate are the estimations. Dates for the divergence of non-vertebrate species from the mammalian lineage are sufficiently old as to not be inconsistent with opinion derived from the fossil record. But the trend in the dates for more recent divergences, those of the rodent/primate and the rat/mouse, appear to be consistently older than either palaeontological or other molecular data would suggest. Two explanations for the bias are possible; either the calibration date from the fossil record is inaccurate. This may be due to an over-estimation in the age of the bird/mammal split in the fossil record, it would need to be younger than the present view. Alternatively, in using a single gene family a bias may be introduced. Most such studies use a large number of genes.

The neutral theory of evolution has assumed a constant mutation rate thus predicting that the rate of molecular evolution is uniform among lineages. By comparing the distance between two species using the distance between each of them and a third, more divergent, species the hypothesis of rate constancy can be evaluated by the relative rate test.

Among the mammalian GSTs tested relative rates are not significantly different (Table 6.8). Previous studies in other genes have also demonstrated this (Wilson *et al.* 1977, Li *et al.* 1990, Easteal 1988, 1990), although alternative studies have indicated a higher rate in the rodent lineage than in humans. This has been attributed to shorter generation times, metabolic rates, change in G+C content of the genomes or differences in DNA repair efficiencies (e.g. Wu and Li 1985, Mouchiroud *et al.* 1988, Saccone *et al.* 1989, Wolfe *et al.* 1989, Martin and Palumbi 1993). Relative rate tests of molecular evolution within the mammalian species represented in this analysis show no evidence of a difference in substitution rates among the lineages tested, however, this data is derived from only one gene family.

6.6 SUMMARY

Multiple alignment of the available GST sequences, incorporating evidence from the recent studies of 3D structure, has provided the basis for construction of phylogenetic trees and evaluation of their reliability. Two regions of increased similarity were found, both containing residues concerned with structural relationships such as subunit interactions in the dimer and with domain interactions within the subunit. The N-terminal region also contained some of the residues involved in the GSH-binding site.

The phylogenetic analysis shows some evidence of an early duplication of an ancestral GST gene and subsequent radiation of GSTs both within and between species. This duplication apparently occurred prior to the plant/animal split as sequences from animals appear in both sectors of the tree. Evidence is also provided for a pre-vertebrate origin for the Alpha, Mu and Pi classes which were originally defined in mammalian species. Currently within the animal kingdom a fifth class of GSTs, containing eye lens crystallin sequences and designated Sigma, is supported by this analysis. There are indications of an analogous class system within the plants. Clearly there are also at least two classes of GSTs in insects, as suggested by Toung *et al.* (1990). Conversely, the inclusion of most of the non-mammalian

sequences into the Theta class, as suggested by Buetler and Eaton (1992) and Pemble and Taylor (1992) is not supported. Contrary to the suggestion by Pemble and Taylor (1992) that Theta sequences represent an ancient progenitor gene, this analysis indicates that both sectors of the phylogenetic tree have very ancient origins. On current evidence it would be difficult to suggest which sequence may be representative of a progenitor gene.

Within the Alpha class the relationship of rabbit and primate sequences is not fully resolved, although indications are that rabbits are more closely related to primates than to rodents, supporting recent molecular and some morphologic studies. There is also evidence of a subgroup of sequences within the Alpha class, consisting of rodent subunit 8 and chicken sequences.

There are two artiodactyl sequences in the Pi class, however, only one is represented in the nucleotide comparison. Artiodactyls are now thought to be more closely related to primates than are rodents but in the Pi class the relationships are ambivalent, the cattle sequence clustering with the human, whereas the pig is on a separate branch. These relationships are not strongly supported, however, and further investigations are warranted.

GSTs may provide an estimation of divergence times among species, however, there appears to be a consistent over-estimation of the dates of divergence when compared to dates determined from the fossil record or from other molecular studies. This may be due to inaccuracies in the fossil record and the calibration date of the bird/mammal split or due to a bias in the use of a single gene family rather than many genes. Relative rate tests of molecular evolution within the mammalian species represented suggests that substitution rates have been similar in the rodent and human lineages.

CHAPTER 7 POSSIBLE ORIGINS OF THE HUMAN GLUTATHIONE *S*-TRANSFERASE M1 POLYMORPHISM

7.1 INTRODUCTION

The genetic deficiency of GSTM1-1 (formerly GST1) is widespread and occurs in similar frequencies in most population groups (Board 1990, Chapter 1.7.2). This deficiency appears to be the result of the absence of the GSTM1 gene which is observed as the lack of an approximately 8kb *Eco*RI fragment in Southern blot experiments (Chapter 5). No clear molecular explanation for this apparent deletion has yet been offered.

As discussed in Chapter 5 there is now good evidence that most of the Mu class genes are clustered at 1p13 (Taylor *et al.* 1990, Zhong *et al.* 1992, Pearson *et al.* 1993, Chapter 5). Previously, the assumption has been that the absence of the *GSTM1* gene sequence was the result of a deletion. However, GSTM1 may have been inserted in the Mu class gene cluster by a chromosomal rearrangement. The human alpha globin genes may be a useful model, as both insertions and deletions have occurred. In some cases of alpha thalassemia the misalignment of duplicated alpha globin genes has resulted in an unequal crossing over causing, on one chromosome, the effective deletion of one gene copy and the creation of a triplicated, chimeric locus on the reciprocal chromosome (Goosens *et al.* 1980). Subsequent selection by Malaria of either chromosomes with alpha globin deletions or insertions has led to their frequent occurence in some populations (Yenchitsomanus *et al.* 1986). As in the alpha globin gene family, high levels of sequence similarity are evident between the clustered human GST Mu class genes. Thus it is possible that the *GSTM1* polymorphism could have arisen by unequal crossing over leading to either the deletion of the pre-existing GSTM1 gene from the Mu class

gene.cluster, or, by the generation *de novo,* of GSTM1 as a chimera of two adjacent genes in the cluster.

7.2 AIM

This study was initiated to explore possible origins of the human Mu class GSTM1 gene and mechanisms to explain the GSTM1 polymorphism. Evidence available that the human Mu class genes are apparently clustered at 1p13 (Chapter 5, Pearson *et al.* 1993) in conjunction with comparisons of the human Mu class sequences provides the basis for the testing of models that may account for the sequence similarity among the Mu class genes and the absence of GSTM1 in 40% of the human population.

7.3 MATERIALS AND METHODS

7.3.1 Alignment of human GST sequences

Sequences used in the alignment of the human Mu class and their GENBANK accession numbers are shown in Table 6.1. HSGSTM3, although clearly a Mu class sequence, is not included in this analysis due to its high level of divergence from all the other mammalian sequences. For the detailed analysis of the relationships among the human Mu class sequences, alignment included all of the coding region and approximately 300bp of the 3′ non-coding region. Beyond this point GSTM5 is highly divergent, containing an additional 98bp when compared to the other Mu class sequences and no successful alignment was achieved. The 5′ region available for several of the sequences was extremely short and was not considered in the analysis. However, intron 6 is available for all human Mu class sequences and therefore was included. Figure 7.1 shows the alignment of coding and noncoding sequences used for the human Mu class analysis. Exon-specific analyses of the sequences was undertaken to identify any heterogeneity in the patterns of phylogenetic relationship among exons.

Figure 7.1 Alignment of Human Mu class sequences for M1, M2, M4, M5 and the rat sequence closest to them. Sequences include complete coding region (upper case), intron 6 and 300bp of the 3′ non-translated region (lower case).

```
HSGSTM1b    ATGCCCATGA  TACTGGGGTA  CTGGGACATC  CGCGGGCTGG  CCCACGCCAT
HSGSTM2     ATGCCCATGA  CACTGGGGTA  CTGGAACATC  CGCGGGCTGG  CCCATTCCAT
HSGSTM4     ATGTCCATGA  CACTGGGGTA  CTGGGACATC  CGCGGGCTGG  CCCACGCCAT
HSGSTM5     ATGCCCATGA  CTCTGGGGTA  CTGGGACATC  CGTGGGCTGG  CCCACGCCAT
RNGST4      ATGCCTATGA  CACTGGGTTA  CTGGGACATC  CGTGGGCTGG  CTCACGCCAT

HSGSTM1b    CCGCCTGCTC  CTGGAATACA  CAGACTCAAG  CTATGAGGAA  AAGAAGTACA
HSGSTM2     CCGCCTGCTC  CTGGAATACA  CAGACTCAAG  CTACGAGGAA  AAGAAGTACA
HSGSTM4     GCGCCTGCTC  CTGGAATACA  CAGACTCAAG  CTACGAGGAA  AAGAAGTATA
HSGSTM5     CCGCTTGCTC  CTGGAATACA  CAGACTCAAG  CTATGTGGAA  AAGAAGTACA
RNGST4      TCGCCTGTTC  CTGGAGTATA  CAGACACAAG  CTATGAGGAC  AAGAAGTACA

HSGSTM1b    CGATGGGGGA  CGCTCCTGAT  TATGACAGAA  GCCAGTGGCT  GAATGAAAAA
HSGSTM2     CGATGGGGGA  CGCTCCTGAT  TATGACAGAA  GCCAGTGGCT  GAATGAAAAA
HSGSTM4     CGATGGGGGA  CGCTCCTGAC  TATGACAGAA  GCCAGTGGCT  GAATGAAAAA
HSGSTM5     CGATGGGGGA  CGCTCCTGAC  TATGACAGAA  GCCAGTGGCT  GAATGAAAAA
RNGST4      GCATGGGGGA  TGCTCCCGAC  TATGACAGAA  GCCAGTGGCT  GAGTGAGAAG

HSGSTM1b    TTCAAGCTGG  GCCTGGACTT  TCCCAATCTG  CCCTACTTGA  TTGATGGGGC
HSGSTM2     TTCAAGCTGG  GCCTGGACTT  TCCCAATCTG  CCCTACTTGA  TTGATGGGAC
HSGSTM4     TTCAAGCTGG  GCCTGGACTT  TCCCAATCTG  CCCTACTTGA  TTGATGGGGC
HSGSTM5     TTCAAGCTGG  GCCTGGACTT  TCCCAATCTG  CCCTACTTGA  TTGATGGGGC
RNGST4      TTCAAACTGG  GCCTGGACTT  CCCCAATCTG  CCCTACTTAA  TTGATGGGTC

HSGSTM1b    TCACAAGATC  ACCCAGAGCA  ACGCCATCTT  GTGCTACATT  GCCCGCAAGC
HSGSTM2     TCACAAGATC  ACCCAGAGCA  ACGCCATCCT  GCGGTACATT  GCCCGCAAGC
HSGSTM4     TCACAAGATC  ACCCAGAGCA  ACGCCATCCT  GTGCTACATT  GCCCGCAAGC
HSGSTM5     TCACAAGATC  ACCCAGAGCA  ATGCCATCCT  GCGCTACATT  GCCCGCAAGC
RNGST4      ACACAAGATC  ACCCAGAGCA  ATGCCATCCT  GCGCTACCTT  GGCCGGAAGC

HSGSTM1b    ACAACCTGTG  TGGGGAGACA  GAAGAGGAGA  AGATTCGTGT  GGACATTTTG
HSGSTM2     ACAACCTGTG  TGGGGAATCA  GAAAAGGAGC  AGATTCGCGA  AGACATTTTG
HSGSTM4     ACAACCTGTG  TGGGGAGACA  GAAGAGGAGA  AGATTCGTGT  GGACATTTTG
HSGSTM5     ACAACCTGTG  TGGGGAGACA  GAAGAGGAGA  AGATTCGTGT  GGACATTTTG
RNGST4      ACAACCTTTG  TGGGGAGACA  GAGGAGGAGA  GGATTCGTGT  GGACGTTTTG

HSGSTM1b    GAGAACCAGA  CCATGGACAA  CCATATGCAG  CTGGGCATGA  TCTGCTACAA
HSGSTM2     GAGAACCAGT  TTATGGACAG  CCGTATGCAG  CTGGCCAAAC  TCTGCTATGA
HSGSTM4     GAGAACCAGG  CTATGGACGT  CTCCAATCAG  CTGGCCAGAG  TCTGCTACAG
HSGSTM5     GAGAACCAGG  TTATGGATAA  CCACATGGAG  CTGGTCAGAC  TGTGCTATGA
RNGST4      GAGAACCAGG  CTATGGACAC  CCGCCTACAG  TTGGCCATGG  TCTGCTACAG

HSGSTM1b    TCCAGAATTT  GAGAAACTGA  AGCCAAAGTA  CTTGGAGGAA  CTCCCTGAAA
HSGSTM2     CCCAGATTTT  GAGAAACTGA  AACCAGAATA  CCTGCAGGCA  CTCCCTGAAA
HSGSTM4     CCCTGACTTT  GAGAAACTGA  AGCCAGAATA  CTTGGAGGAA  CTTCCTACAA
HSGSTM5     CCCAGATTTT  GAGAAACTGA  AGCCAAAATA  CTTGGAGGAA  CTCCCTGAAA
RNGST4      CCCTGACTTT  GAGAGAAAGA  AGCCAGAGTA  CTTAGAGGGT  CTCCCTGAGA

HSGSTM1b    AGCTAAAGCT  CTACTCAGAG  TTTCTGGGGA  AGCGGCCATG  GTTTGCAGGA
HSGSTM2     TGCTGAAGCT  CTACTCACAG  TTTCTGGGGA  AGCAGCCATG  GTTTCTTGGG
HSGSTM4     TGATGCAGCA  CTTCTCACAG  TTCCTGGGGA  AGAGGCCATG  GTTTGTTGGA
HSGSTM5     AGCTAAAGCT  CTACTCAGAG  TTTCTGGGGA  AGCGGCCATG  GTTTGCAGGA
RNGST4      AGATGAAGCT  TTACTCCGAA  TTCCTGGGCA  AGCAGCCATG  GTTTGCAGGG

HSGSTM1b    AACAAGgtaa  aggaggagtg  atatgggga-  a-t-gagatc  tgttt-tgct
HSGSTM2     GACAAGgtaa  tgggggcgtg  tgatggggac  accacagatt  tgtca-tact
HSGSTM4     GACAAGgtaa  tgggggcatg  tgatgaggac  actagagatt  tgcca-taca
HSGSTM5     GACAAGgtaa  aggaggagtg  atatgggga-  a-t-gagatc  tgttt-tact
RNGST4      AACAAGgtaa  aggcagcggg  ---tggggag  a--ag-gatt  tgccatttct

HSGSTM1b    t--cacgtgt  tatgaggttc  cagcccacat  attcttggcc  ttctgcagAT
HSGSTM2     tccta--tat  tatgaggttt  cagcccacat  atccttggcc  ttatccagAT
HSGSTM4     tccta--tgt  tacgagattc  cagcccacac  attcttggcc  ttctgcagAT
HSGSTM5     t--catgtgt  ttcgggtttt  cagcccacac  attcttggcc  ttctccagAT
RNGST4      tcccaggttg  -tcaa--ttc  tagcac--tc  accettggct  tcctgcagAT

HSGSTM1b    CACTTTTGTA  GATTTTCTCG  TCTATGATGT  CCTTGACCTC  CACCGTATAT
HSGSTM2     CACCTTTGTG  GATTTCATCG  CTTATGATGC  CCTTGAGAGA  AACCAAGTAT
HSGSTM4     CACCTTTGTA  GATTTCCTCG  CCTATGATGT  CCTTGACCTC  CACCGTATAT
HSGSTM5     CACCTTTGTG  GATTTCCTTG  CCTATGATGT  CCTTGACATG  AAGCGTATAT
RNGST4      TACGTATGTG  GATTTTCTTG  TTTACGATGT  CCTTGATCAA  CACCGTATAT

HSGSTM1b    TTGAGCCCAA  CTGCTTGGAC  GCCTTCCCAA  ATCTGAAGGA  CTTCATCTCC
HSGSTM2     TTGAGCCCAG  CTGCCTGGAT  GCCTTCCCAA  ACCTGAAGGA  CTTCATCTCC
HSGSTM4     TTGAGCCCAA  CTGCTTGGAC  GCCTTCCCAA  ATCTGAAGGA  CTTCATCTCC
HSGSTM5     TTGAGCCCAA  GTGCTTGGAC  GCCTTCCTAA  ACTTGAAGGA  CTTCATCTCC
RNGST4      TTGAACCCAA  GTGCCTGGAC  GCCTTCCCAA  ACCTGAAGGA  CTTCGTGGCT
```

```
HSGSTM1b    CGCTTTGAGG    GCTTGGAGAA    GATCTCTGCC    TACATGAAGT    CCAGCCGCTT
HSGSTM2     CGATTTGAGG    GCTTGGAGAA    GATCTCTGCC    TACATGAAGT    CCAGCCGCTT
HSGSTM4     CGCTTTGAGG    GCTTGGAGAA    GATCTCTGCC    TACATGAAGT    CCAGCCGCTT
HSGSTM5     CGCTTTGAGG    GTTTGAAGAA    GATCTCTGCC    TACATGAAGT    CCAGCCAATT
RNGST4      CGGTTTGAGG    GCCTGAAGAA    GATATCTGAC    TACATGAAGA    GCGGCCGCTT

HSGSTM1b    CCTCCCAAGA    CCTGTGTTCT    CAAAGATGGC    TGTCTGGGGC    AACAAGTAGg
HSGSTM2     CCTCCCAAGA    CCTGTGTTCA    CAAAGATGGC    TGTCTGGGGC    AACAAGTAGg
HSGSTM4     CCTCCCAAAA    CCTCTGTACA    CAAGGGTGGC    TGTCTGGGGC    AACAAGTAAt
HSGSTM5     CCTCCGAGGT    CTTTTGTTTG    GAAAGTCAGC    TACATGGAAC    AGCAAATAGg
RNGST4      CCTCTCCAAG    CCAATCTTTG    CAAAGATGGC    CTTTTGGAAC    CCAAAGTAGc

HSGSTM1b    gccttga-ag    gc-aggaggt    gggagtgagg    agcccatact    -cagcctgct
HSGSTM2     gccttga-ag    gc-aggaggt    gggagtgagg    agcccatact    -cagcctgct
HSGSTM4     gccttga-ag    gccaggaggt    gggagtgagg    agcccatact    -cagcctgct
HSGSTM5     gcccagtgat    gccagaagat    gggagggagg    agccaacctt    gctgcctgcg
RNGST4      acc-acaaag    tccag-acct    ggg---gat-    actcatgagt    gc--cctgct

HSGSTM1b    gcccaggctg    tgcagc-gca    gctggactct    gcatcccagc    acctgcctcc
HSGSTM2     gcccaggctg    tgcagc-gca    gctggactct    gcatcccagc    acctgcctcc
HSGSTM4     gcccaggctg    tgcagc-gca    gctggactct    gcatcccagc    acctgcctcc
HSGSTM5     accctgg---    ---agg-aca    gcctgactc-    -----cctgg    acctgccttc
RNGST4      -----ggctg    tgggcctaga    gcatggctct    ggcgcccacc    acatgcagc-

HSGSTM1b    tcgttccttt    ctcctgttta    ttcccatctt    tactcccaag    acttcattgt
HSGSTM2     tcgttccttt    ctcctgttta    ttcccatctt    tactcccaag    acttcattgt
HSGSTM4     tcgttccttt    ctcctgttta    ttcccatctt    taccccaag    actttattgg
HSGSTM5     ttccttttc     cttctttcta    ctctcttctc    ttc-cccaag    gcctcattgg
RNGST4      ---tttct-c    ctcctttcca    ttccc---tg    ttcctcca--    ---tc--t--

HSGSTM1b    ccctcttcac    tccccctaaa    ccccctgtcc-    catgcaggcc    ctttgaagcc
HSGSTM2     ccctcttcac    tccccctaaa    ccccctgtcc-    catgcaggcc    ctttgaagcc
HSGSTM4     gcctcttcac    ttccccctaaa    ccccctgtcc-    catgcaggcc    ctttgaagcc
HSGSTM5     cttcctttct    tctaacatca    tcccctcccg    catcgaggct    ctttaaagct
RNGST4      -cctcttccc    ---------a    gcccttgcct    cagtcaagc-    ctcagttccc

HSGSTM1b    tcagctaccc    actatccttc    gtgaacatcc    cc-tcccatc    at-taccctt
HSGSTM2     tcagctaccc    actatccttc    gtgaacatcc    cc-tcccatc    at-taccctt
HSGSTM4     tcagctaccc    actttccttc    atgaacatcc    ccctcccaac    ac-taccctt
HSGSTM5     tcagctcccc    actgtcctcc    atcaaagtcc    ccctcctaac    gtcttcctttt
RNGST4      tggtctctcc    attt--cttc    atta--gtcc    cc-tccc-tt    gtct---ctg

HSGSTM1b    ccctgcacta    aagccagcct    gaccttcctt    cctgttagtg    gttgtgtctg
HSGSTM2     ccctgcacta    aagccagcct    gaccttcctt    cctgttagtg    gttgtgtctg
HSGSTM4     ccctgcacta    aagccagcct    gaccttcctt    cctgttagtg    gttgtatctg
HSGSTM5     ccctgcacta    acgccaacct    gactgctttt    cctgtcagtg    cttttctctt
RNGST4      ccctgca-t-    ---ccaacc-    -------ctt    ccc-tcactg    attttcggag

HSGSTM1b    cttt
HSGSTM2     cttt
HSGSTM4     cttt
HSGSTM5     cttt
RNGST4      gact
```

7.3.2 Estimation of phylogenetic relationships

Distances between the Human Mu class sequences were calculated for synonymous and nonsynonymous sites within the coding regions using the methods Li *et al.* (1985) and Li (1993). This permits estimations of the numbers of synonymous differences per synonymous site and the numbers of nonsynonymous differences per nonsynonymous site. Kimura 2-parameter distances (Kimura 19 80) were calculated for the noncoding regions in CLUSTALV (Higgins 1993). A weighted mean was then derived from the two scores and a weighted standard error calculated. These are shown in Table 7.1.

To estimate the phyletic relationships among the GSTs from nucleotide distances unrooted phylogenetic trees were constructed by the neighbor-joining (NJ) method (Saitou and Nei 1987) as detailed in Chapter 6.3.5. Phylogenetic trees were tested for reliability by the bootstrap method described in Chapter 6.3.6.

For the estimation of divergence times among the human Mu class sequences, distances used were those described above, using the methods Li *et al.* (1985) and Li (1993).

TABLE 7.1 Nucleotide substitutions/100 synonymous and noncoding sites for pairwise comparisons between human Mu class GSTM1, GSTM2, GSTM4 and GSTM5 and rat subunit 4, the nearst outgroup sequence.

COMPARISONS		EXONS 1-4+7 and 3′ INTRON 6		EXONS 5-6 + 5′ INTRON 6		EXON8 coding & untranslated region	
		wt mean	wt S.E.	wt mean	wt S.E.	wt mean	wt S.E.
HSGSTM1	HSGSTM2	0.199	0.065	0.126	0.752	0	0
	HSGSTM4	0.078	0.0399	0.123	0.365	0.039	0.013
	HSGSTM5	0.191	0.069	0.046	0.145	0.384	0.053
	RNGST4	0.648	0.026	0.113	0.474	0.462	0.059
HSGSTM2	HSGSTM4	0.196	0.068	0.070	0.306	0.036	0.012
	HSGSTM5	0.224	0.073	0.116	0.357	0.384	0.053
	RNGST4	0.663	0.045	0.103	0.668	0.462	0.059
HSGSTM4	HSGSTM5	0.152	0.058	0.116	0.359	0.406	0.052
	RNGST4	0.547	0.119	0.107	0.439	0.479	0.056
HSGSTM5	RNGST4	0.601	0.168	0.110	0.638	0.544	0.068

7.4. RESULTS

Genetic deficiency of GSTM1-1 is widespread in the human population (Chapter 1.7.2 and Chapter 5, Board 1980, Seidegård *et al.* 1988). Although there has been no clear molecular explanation of this polymorphism there has been an underlying assumption in the literature that the absence of the *GSTM1* gene is the result of a deletion, thus the ancestral chromosome would have contained the *GSTM1* gene. Following on from the model proposed for the human alpha globin gene family and a comparison of the human Mu class GST sequences, an insertion model has also been suggested to explain the *GSTM1* polymorphism (Board *et al.* 1993). If the polymorphism is

the result of an insertion event then the *GSTM1* gene could be considered a chimeric product and the ancestral chromosome would not have contained the *GSTM1* gene.

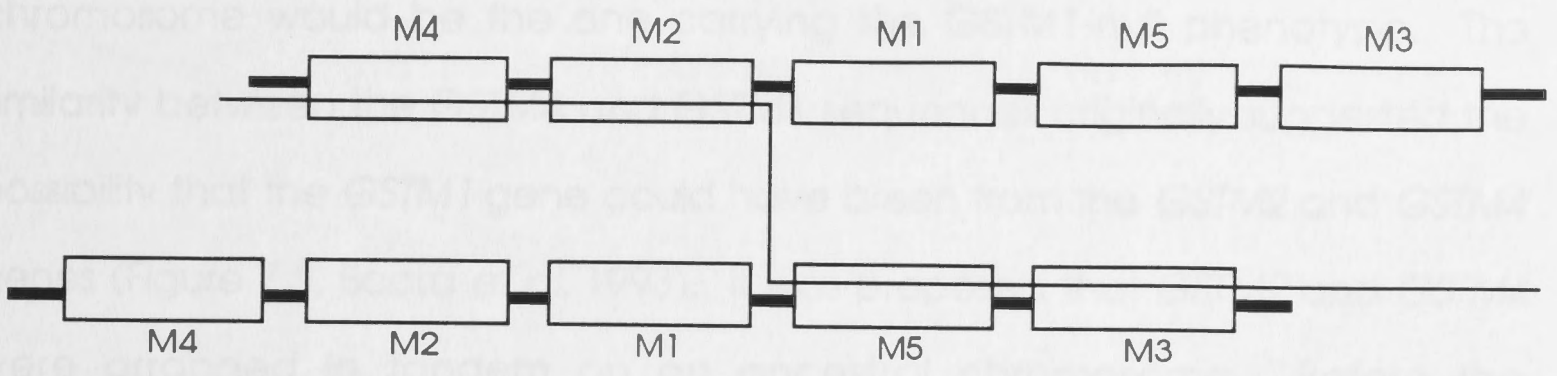
It is possible to propose and test models that may account for the observed similarities and differences between the Mu class sequences, and the absence of *GSTM1* in some individuals. Unequal crossover events need to be invoked as the mechanism for both the deletion and the insertion models. Such events are thought to require a high degree of sequence similarity to promote non-homologous pairing (Metzenberg *et al.* 1991).

In the deletion model a crossover event could have occurred either *between* (inter-locus) Mu class gene loci or *within* (intra-locus) Mu class genes (Figure 7.2a and b). If the former is the case then a simple deletion of the locus would be the result and evidence for this would need to be sought outside the *GSTM1* gene (Figure 7.2a). This type of deletion model makes no predictions as to the arrangement of the gene loci on the chromosome or of similarities or differences among the Mu class sequences beyond the probability of a high level of sequence similarity for the promotion of an unequal crossover event.

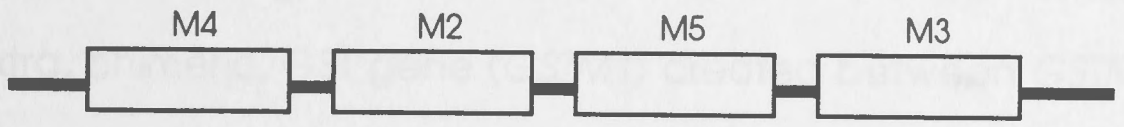
In the latter case, where an intra-locus unequal crossover event occurred, then a chimeric gene would be the result (Pearson *et al.* 1993). In the simplest version of this model the chimeric gene (GSTM2/M1) would be functional and substantially equivalent to the non-GSTM1 donor gene (GSTM2). Thus in one step the GSTM1-null phenotype is produced (Figure 7.2b). This model predicts that there will be a strong similarity between the 3′ end of the chimeric GSTM2/GSTM1 and GSTM2. It is also predicted that the functional chimeric gene sequence (GSTM2/GSTM1) may be different in GSTM1-null chromosomes to that on GSTM1-positive chromosomes (GSTM2), the ancestral state, and that the intervening sequence between the 5′ donor gene (*GSTM2*) and *GSTM1* would be missing on GSTM1-null chromosomes.

Figure 7.2 Schematics showing proposed deletion models for the origin of the GSTM1 polymorphism in humans **(a)** between loci, **(b)** within loci.

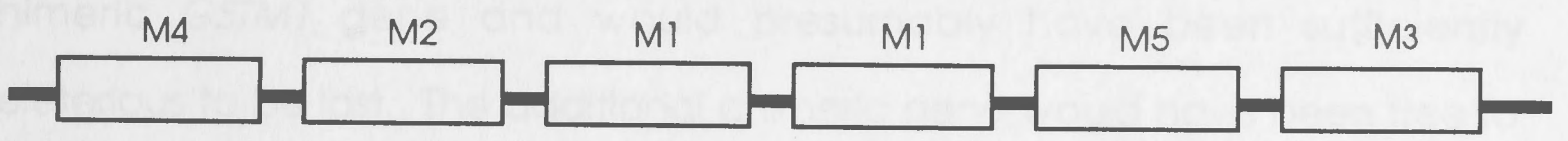
(a) between loci



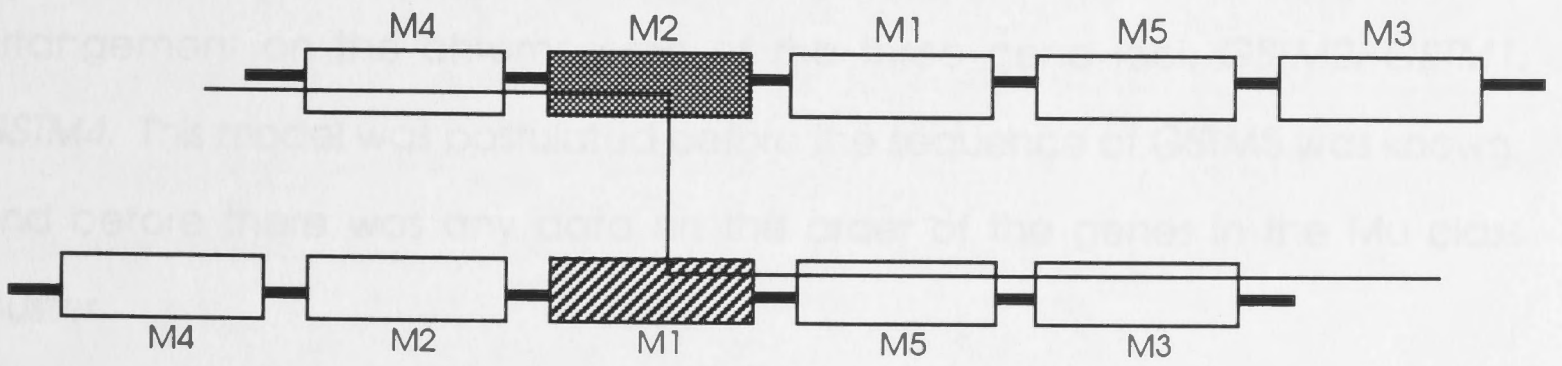
products



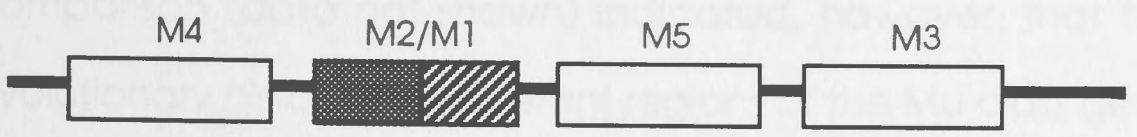
lost product ?



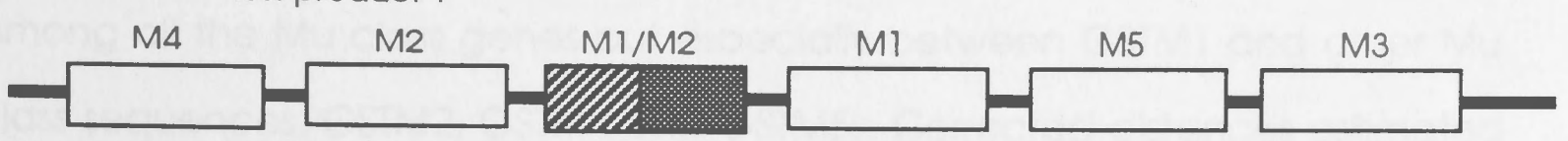
(b) within loci



products



lost product ?

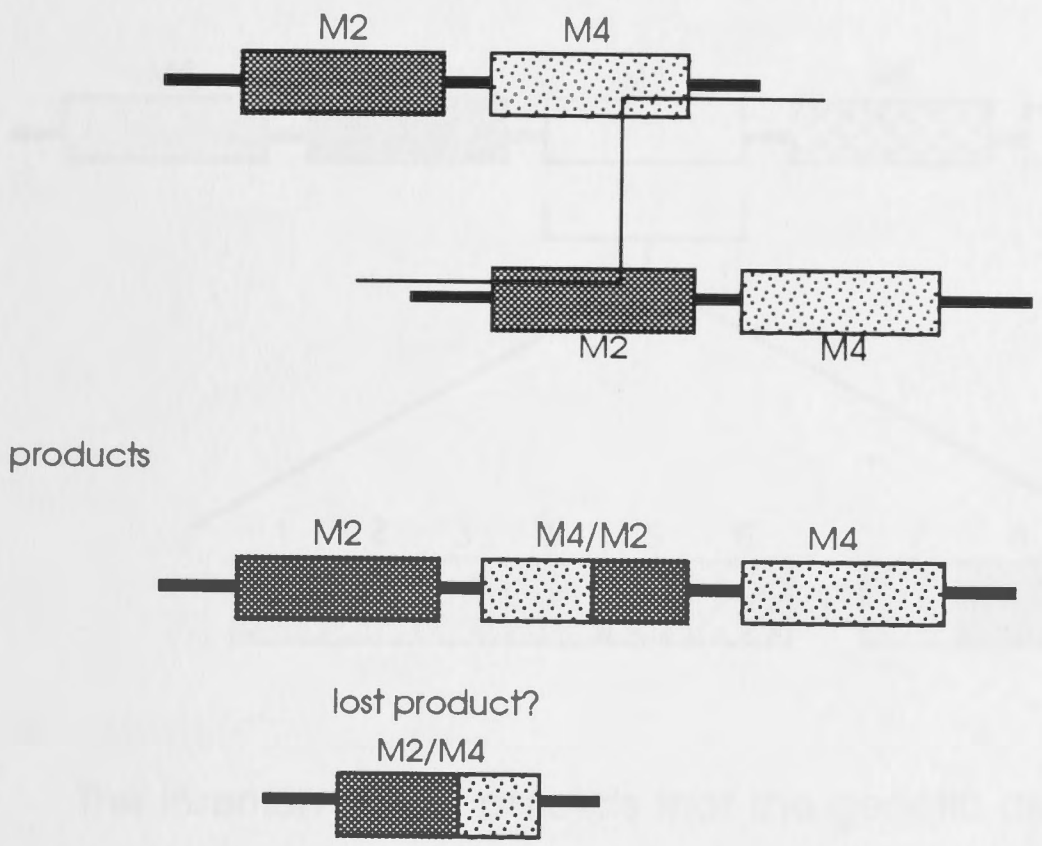


In the insertion models, *GSTM1* would also have resulted from a chimera due to an intra-locus unequal crossover event. The ancestral chromosome would be the one carrying the GSTM1-null phenotype. The similarity between the GSTM4 and GSTM1 sequences originally suggested the possibility that the *GSTM1* gene could have arisen from the *GSTM2* and *GSTM4* genes (Figure 7.3, Board *et al.* 1993). It was proposed that *GSTM2* and *GSTM4* were arranged in tandem on an ancestral chromosome. Before the divergence of the major human population groups an unequal crossover occurred. This generated two new and different chromosomes, one had an extra, chimeric, GST gene (*GSTM1*) created between *GSTM2* and *GSTM4*. The other chromosome would have lost both GSTM2 and GSTM4 in creating the chimeric *GSTM1* gene and would presumably have been sufficiently deleterious to be lost. The additional chimeric gene would have been free to accumulate mutations because it was contained in a chromosome having the full ancestral compliment of GST genes. This model predicts that there will be a strong similarity between the 5′ end of GSTM1 and GSTM4, and a strong similarity between the 3′ end of GSTM1 and GSTM2. It also predicts the arrangement on the chromosome of the three gene loci, *GSTM2, GSTM1, GSTM4*. This model was postulated before the sequence of GSTM5 was known, and before there was any data on the order of the genes in the Mu class cluster.

The validity of the models may be tested in part by the comparison and analysis of the nucleotide sequences (Figure 7.1). Exon specific sequence comparison (data not shown) indicated, however, that there were different evolutionary histories for different regions of the Mu class genes. Alignment and corrected nucleotide distance estimations clearly show complex relationships among all the Mu class genes but especially between GSTM1 and other Mu class sequences, GSTM2, GSTM4 and GSTM5. Corrected distances estimated for specific subregions among all the sequences are shown in Table 7.1 and

indicate the possibility of gene conversion events having occurred among all genes.

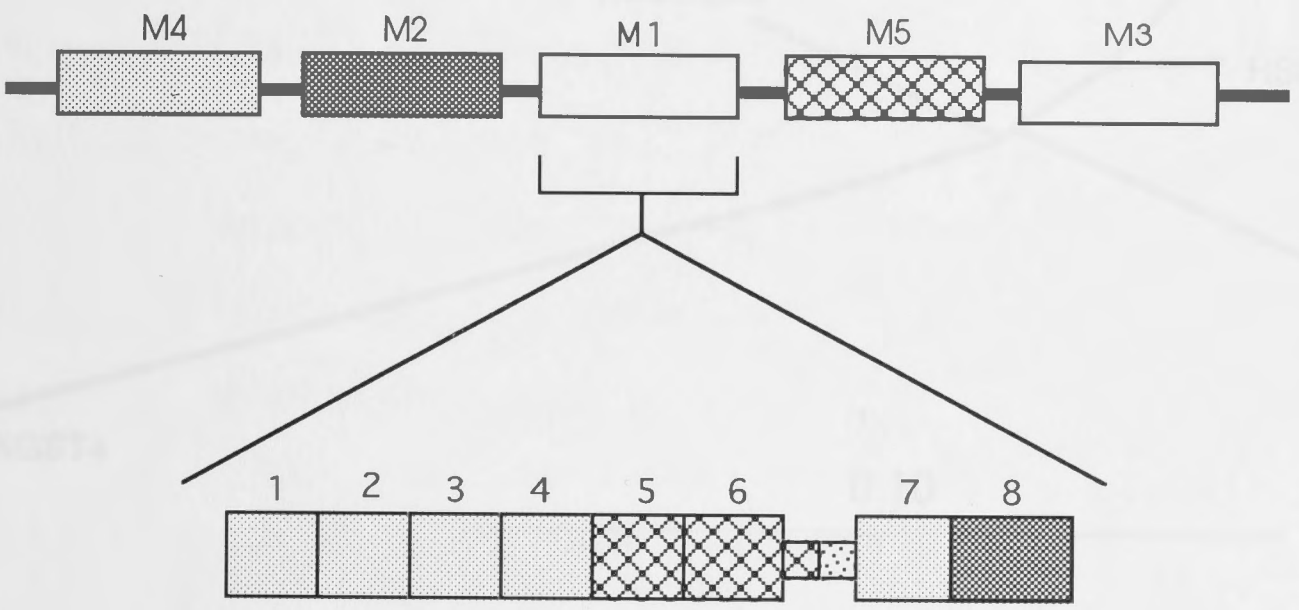
Figure 7.3 Original insertion model proposed GSTM4 and GSTM2 as the donor genes in producing a chimeric GSTM1 product, Board *et al.* (1993).



Exons 1 to 4, and 7 (and including the 3′ of intron 6) in GSTM1 have a greater similarity to GSTM4, whereas exons 5 and 6 and the 5′ of intron 6 have a higher similarity to GSTM5. Exon 8, containing the 3′ tail of the coding region, and the 3′ noncoding region shows a marked similarity to GSTM2 (Figure 7.4). A series of phylogenetic trees were generated from the distance estimates to reflect these relationships and are shown in Figure 7.5.

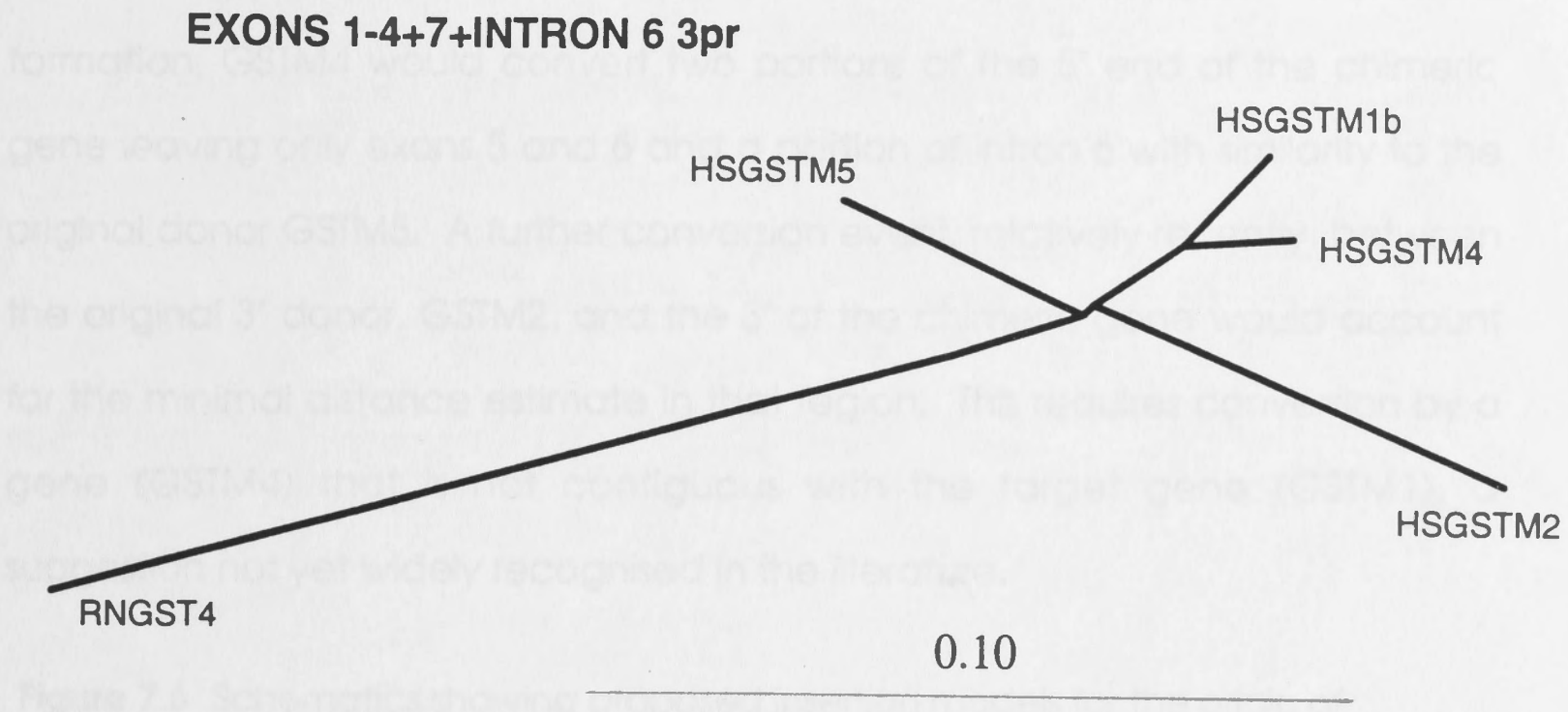
The complex pattern evident between GSTM1 and the other three Mu class sequences (Figure 7.4) does not provide evidence in support of a simple model for the origin of GSTM1. While differences in similarities between the 5′ or 3′ ends of the cDNAs are in general agreement with the basic insertion model, to account for the more complex relationships observed requires GSTM4 and GSTM2 as donor genes and, in addition, a subsequent gene conversion event by GSTM5 encompassing exons 5, 6 and the 5′ portion of intron 6 as shown in Figure 7.4.

Figure 7.4 Schematic showing arrangement of Mu class loci on human chromosome 1p13 and the relationships between GSTM1 and the other Mu class genes.



The insertion model predicts that the genetic distances calculated between the respective segments of GSTM1 and the donor genes should be similar and the conversion event by GSTM5 should, at the least, be of equivalent nucleotide distance and probably less distant. The corrected distances between the relevant regions, shown in Table 7.1, indicate that the divergence of GSTM1 from the proposed 3′ donor gene, GSTM2, is considerably less than the distance between GSTM1 and the proposed 5′ donor gene, GSTM4. Additionally, the divergence between GSTM1 and GSTM5 in the exons 5 and 6 region is somewhat greater than either of the distances estimated for the other two regions, implying this relationship predates the GSTM1/M2 and GSTM1/M4 relationships in the 3′ and 5′ regions. Further, Pearson *et al.* (1993) has reported evidence of the chromosomal arrangement of the GST Mu class genes from a Yac clone, suggesting that the order is not as suggested in the first insertion model, but is GSTM4, GSTM2, GSTM1, GSTM5, GSTM3 as shown in Figure 7.4. This means it is unlikely that the chimeric GSTM1 was formed using GSTM2 and GSTM4 as donor genes, the arrangement of GSTM1 between GSTM2 and GSTM5 means those two genes are the more likely donors.

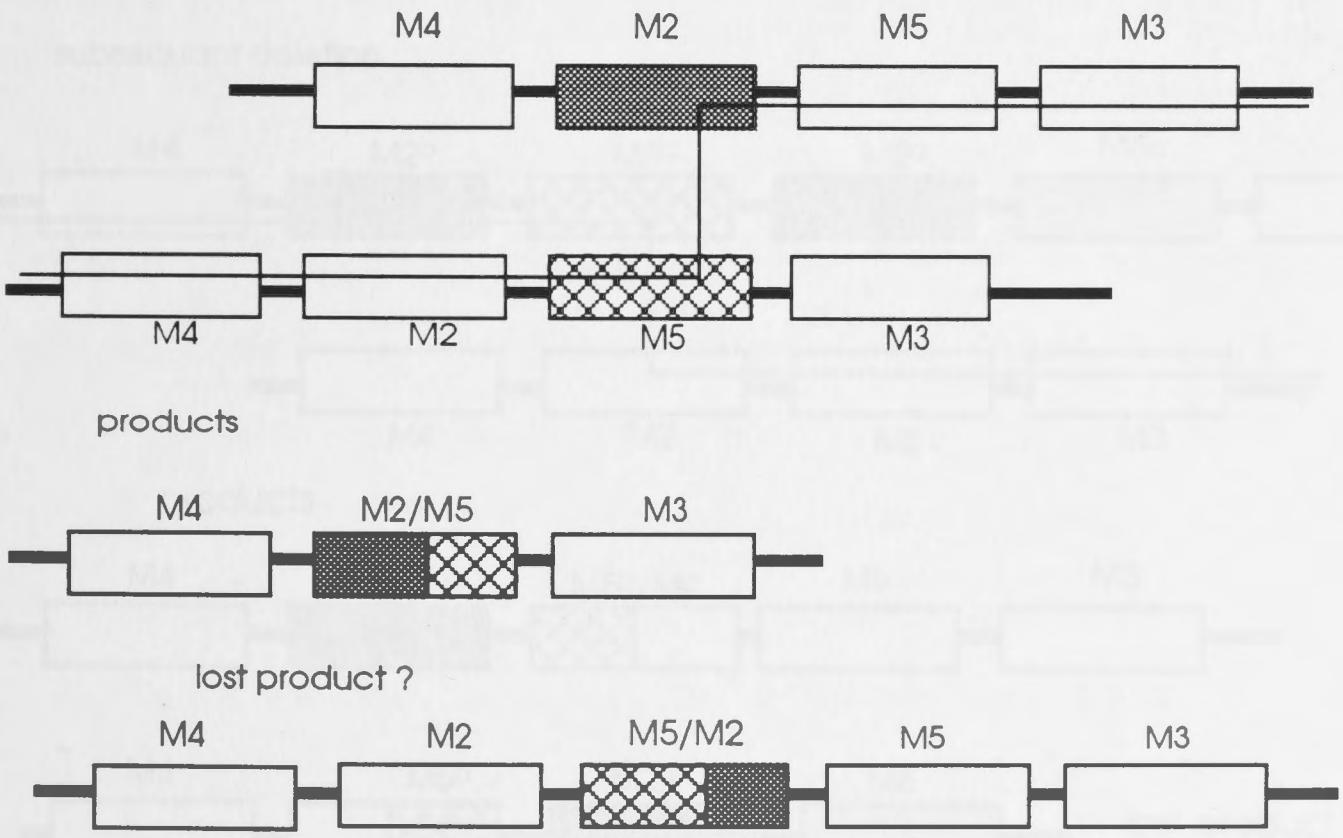
Figure 7.5 Neighbor-joining trees showing the relationships among the human Mu class seuqences in the three regions deltailed.



To use GSTM5 and GSTM2 as the original donor genes to create a chimera of 5′ GSTM5-like sequence and a 3′ GSTM2-like sequence (Figure 7.6a) a more complex insertion model can be formulated. Subsequent to chimera formation, GSTM4 would convert two portions of the 5′ end of the chimeric gene leaving only exons 5 and 6 and a portion of intron 6 with similarity to the original donor GSTM5. A further conversion event, relatively recently, between the original 3′ donor, GSTM2, and the 3′ of the chimeric gene would account for the minimal distance estimate in that region. This requires conversion by a gene (GSTM4) that is not contiguous with the target gene (GSTM1), a suggestion not yet widely recognised in the literature.

Figure 7.6 Schematics showing proposed insertion models for the origin of the GSTM1 polymorphism in humans **(a)** GSTM1 as a simple chimera between GSTM2 and GSTM5, **(b)** as a duplication followed by a deletion.

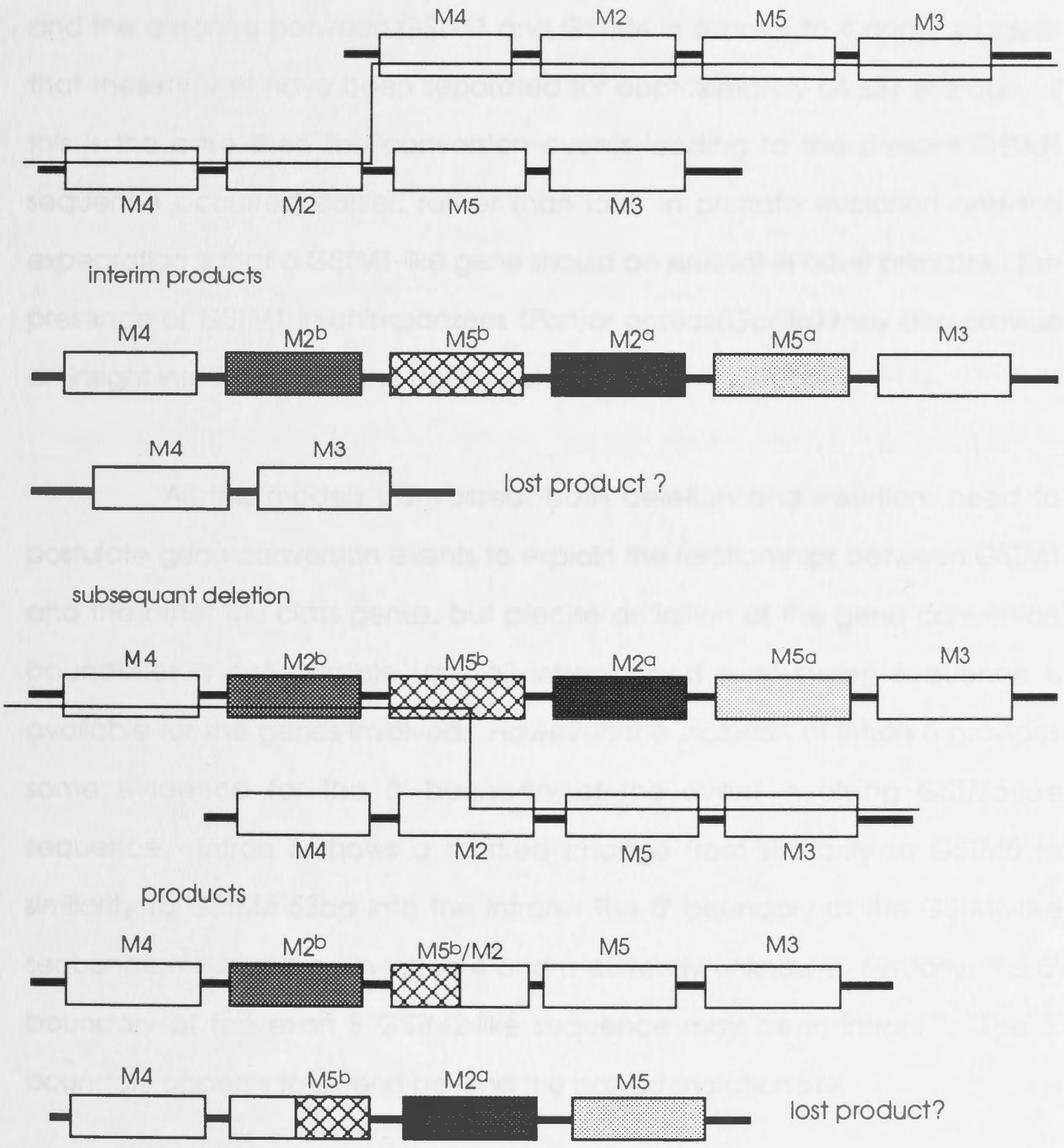
(a) GSTM1 as a simple chimera between GSTM2 and GSTM5



A third model incorporating a chimeric gene requires an initial duplication of a block of sequence including GSTM5 and GSTM2 (Figure 7.6b). The duplication would be followed by an unequal crossover event and creation of the chimeric gene by deletion of the 3′ of GSTM5a and the 5′ of

GSTM2a forming the chimeric gene (Figure 7.6b). Subsequent gene conversion events must be also postulated, as in all models.

Figure 7.6b Schematics showing a proposed model for the origin of the GSTM1 polymorphism in humans as a duplication followed by a deletion.



It is possible to date approximately the events described in the evolution of GSTM1 by the same means used to estimate species divergences in section 6.4.5. However, a more suitable date for calibration of these

estimates is the rodent/primate split the date of this divergence was estimated in Chapter 6.5.4 to be 122 Ma ago in the Mu class GSTs. However, the distances calculated suggest that the conversion of the GSTM1 exon 8 region by GSTM2 is so recent as to be undatable by this means. The separation between GSTM1 and GSTM5 in the exon 5 and 6 region is estimated to have occurred approximately 116 ±48 Ma ago and the distance between GSTM1 and GSTM4 in exons 1 to 4 and 7 suggests that these regions have been separated for approximately 64 ±37 Ma ago. If this is the case then the conversion events leading to the present GSTM1 sequence occurred earlier, rather than later in primate evolution and the expectation is that a GSTM1-like gene should be present in other primates. The presence of GSTM1 in chimpanzees (*Pan*)or gorillas*(Gorilla)* may also provide an insight into the age of the polymorphism.

All the models canvassed, both deletion and insertion, need to postulate gene conversion events to explain the relationships between GSTM1 and the other Mu class genes, but precise definition of the gene conversion boundaries is not possible until all intronic and surrounding sequence is available for the genes involved. However, the inclusion of intron 6 provides some evidence for the 3′ boundary of the event involving GSTM5-like sequence. Intron 6 shows a marked change from similarity to GSTM5 to similarity to GSTM4 53bp into the intron. The 5′ boundary of the GSTM5-like sequence is likely to be in intron 4 and is currently unknown. Similarly, the 5′ boundary of the exon 8 GSTM2-like sequence may be in intron 7. The 3′ boundary appears to extend beyond the polyadenalation site.

7.5 DISCUSSION

Considering the level of gene structure and sequence conservation evident within the human Mu class, it seems clear that gene duplication has probably taken place. In attempting to explain the origin of the GSTM1

polymorphism exon-specific sequence analysis has brought to light a complex array of relationships among the human Mu class. The apparent regionality of the sequence similarities between GSTM1 and three other Mu class sequences, GSTM2, GSTM4 and GSTM5, suggests that gene conversion may also have played a significant role in its evolutionary history. Gene conversion is also evident among the other Mu class sequences although to a lesser extent than is apparent in GSTM1

Two basic models have been proposed for the origin of the GSTM1 polymorphism, a deletion model, widely assumed in the literature and an insertion model. The simplicity of the GSTM2/GSTM4 chimeric model, first proposed by Board *et al.* (1993), has not been substantiated by detailed analysis. The evidence, from Yac clones, presented by Pearson *et al.* (1993) of the order in which the Mu class genes are placed on the chromosome suggests the proposed production of a simple chimera between GSTM4 and GSTM2, the suggested donor genes, is unlikely. The more complex GSTM2/GSTM5 chimeric model can be supported by the mapping and genetic distance data but requires a gene conversion event by a non-contiguous gene. This may be possible but has not yet been widely recognised in the literature, more compelling evidence would be necessary before it is considered a likely possibility. The final insertion model proposed is more complex, requiring the duplication of a block of sequence before the creation of the chimera and subsequent gene conversion. This model is capable of accounting for the relationships seen in the analysis of the sequences and for the GSTM1 polymorphism, however, it is a very complex model, requiring several steps thus diminishing its probability.

Both deletion models are simpler, neither the inter-locus nor the intra-locus models are required to explain any of the complex relationships observed between GSTM1 and the other Mu class sequences, conversion events could then be assumed to have occurred independently, either before or after the deletion event. The second deletion model, proposed by Pearson *et al.* (1993)

is a chimeric model, however, the chimera is a substitute for the original GSTM2 while GSTM1 is deleted in the same event. This model can account for the similarity at the 3' between GSTM1 and GSTM2 and also the deletion, providing the deletion is a relatively recent event. The relationship of GSTM1 with GSTM4 in exons 1 to 4 and 7 and with GSTM5 in exons 5 and 6 could be explained as separate, unconnected conversion events. Conversion between GSTM4 and GSTM1 in this model also, however, involves non-contiguous genes. The model also predicts differences in the GSTM2 sequence between the GSTM1-null chromosome and the GSTM1-positive chromosme but there is no evidence so far for this.

Sequencing of the genes and surrounding regions may help resolve these issues. In the inter-locus deletion model, the length of sequence missing is not predicted but may not be substantially longer than the actual gene. However, it is possible that at least a small segment of inter-genetic sequence may be missing in the GSTM1-null that is present in the GSTM1-positive. Predictions from an intra-locus deletion model include that the functional chimeric gene sequence (GSTM2/GSTM1) may be different in GSTM1-null chromosomes to that on GSTM1-positive chromosomes (GSTM2), the ancestral state, and that the intervening sequence between the 5' donor gene (*GSTM2*) and *GSTM1* should be missing on GSTM1-null chromosomes. The insertion models suggest the possibility of some sequence repetition extending beyond the GSTM1 locus that would not be present on the GSTM1-null chromosome. Thus sequence beyond the 5' of GSTM1 should resemble the intergenic sequence in front of GSTM5 5' and similarly the 3' of GSTM1 should resemble 3' intergenic sequence beyond GSTM2.

The insertion models for the origin of GSTM1 predict the order in which the Mu class genes are arranged in the cluster at chromosome 1p13 and the predicted arrangements are incorporated in Figures 7.2, 7.4 and 7.6. Concurrent studies have now determined the order of the Mu class genes at chromosome 1p13 (Pearson *et al.* 1993). In terms of the insertion models this

evidence is more consistent with the second, more complex, model which uses GSTM5 and GSTM2 as the original donor genes. Deletion models are unaffected by the arrangement of the genes along the chromosome as no predictions of order can be made from them.

There is an almost endless array of complex models that could be formulated to explain the relationships among the human Mu class sequences only a few of those possible have been analysed here. Without further corroborative evidence the deletion models require the least number of steps and are therefore the more parsimonious explanations.

Under a deletion model one possible explanation of the increase in GSTM1-null frequency in the human population is that in certain circumstances the end product of a conjugation reaction is more reactive and deleterious than the initial substrate. There may be selection for the deletion of GSTM1 if it is an efficient catalyst of such a reaction. The high specificity of GSTM1-1 for *trans*-stilbene oxide (Table 3.2) shows that GSTM1-1 does indeed have significantly altered substrate specificities.

Alternatively, there is evidence of the GSTM1-null phenotype being over-represented in patients with several types of cancer in comparison to the GSTM1-positive phenotype (Strange *et al.* 1989, Heagerty *et al.* 1994). Thus it appears that in certain circumstances it is advantageous to carry the GSTM1 gene. Under an insertion model this could be accounted for by suggesting the increase in frequency of the chromosome containing the triplicated gene arrangement possibly may have been due to a changed substrate specificity in the chimeric GSTM1-1of benefit in reducing cell damage leading to cancer.

It is unknown at this point whether the GSTM1-null phenotype is the subject of concerted selection, or could be considered a neutral allele, or whether it is being maintained by balancing selection. It has been generally thought that polymorphisms under either positive or negative selection reach fixation relatively quickly. There are also suggestions that neutral or near-neutral polymorphisms would also reach fixation in a relatively short time, given

the small effective population size of the human species for the majority of its existence (Nei 1983). Thus, unless the GSTM1 polymorphism has been the subject of balancing selection it is unlikely to be a very old polymorphism.

Dating the origin of the polymorphism is not possible at this point, however, the relative uniformity of the GSTM1-null phenotype throughout human populations indicates that the event causing the polymorphism probably occurred, at least, before the world-wide spread of our species. Origin of the polymorphism as an inter-locus deletion event would allow a relatively recent origin, as the gene conversion events apparent from this analysis could have occurred prior to the deletion. Whereas in an intra-locus deletion model it may be possible to date the event leading to the polymorphism from the genetic distance between the version of GSTM2 predicted on the GSTM1-null chromosome and the version predicted on the GSTM1-positive chromosome, if and when a second type of GSTM2 is sequenced. In the insertion model the event leading to the polymorphism may be datable from the distances between the donor genes and the chimeric gene. Estimates of divergence among the human Mu class sequences give a strong indication that GSTM1 is not a recent creation, suggesting that if an insertion model is correct then the polymorphism is probably pre-hominid in origin and is being maintained under balancing selection.

The null allele may well promote even more gene mispairing in this multigene region than the similarity between the genes would probably normally promote and thus gene duplication and gene conversion events may be further increased. Taylor et al. (1991) have proposed that there may be somatic gene conversion and rearrangements in the Mu class gene cluster, but definitive evidence has yet to be presented.

It is clear that given the complex relationships found between GSTM1 and three other Mu class sequences a simple explanation of the GSTM1 polymorphism may not be possible. At this point the complexity of the relationships among the human Mu class prohibit the formulation of a simple

explanation, further sequencing information and population studies are required. Even with the additional information it may not be possible to determine the origin of the GSTM1 polymorphism with more certainty, but such studies will certainly shed more light on some of the complex relationships of the human Mu class genes.

7.6 SUMMARY

Two basic models have been proposed for the origin of the GSTM1 polymorphism, a deletion model, widely assumed in the literature and an insertion model. Exon-specific sequence analysis has brought to light a complex array of relationships among the human Mu class. The simplicity of the GSTM2/GSTM4 chimeric insertion model has not been substantiated by the detailed analysis. To account for the relationships evident in the comparison of the human Mu class sequences it is necessary to invoke the process of gene conversion. Several models were tested for both an insertion or deletion origin for the GSTM1 polymorphism. An insertion model capable of accounting for the observed relationships is more complex than the deletion models. However, the deletion models do not attempt to explain all the relationships observed.

CHAPTER 8 FINAL DISCUSSION AND CONCLUSIONS

This thesis describes several investigations, including the identification of a new human Mu class GST, the comparative characterisation of the human Mu class GSTs, their chromosomal localisation and an analysis of the evolutionary relationships of the GST super family. The results of the analysis of each facet of this study will be summarised, then discussed and directions for future work will be considered.

8.1 Cloning and characterisation of GSTM4

Chapter 3 presents the detailed sequence analysis of two novel Mu class cDNA clones encoding a new Mu class glutathione S-transferase. Both clones are incomplete, one is missing the sequence encoding exon 4 and the other is missing exon 8. This appears to be the result of alternative splicing, however, the complete sequence has been deduced from the two cDNA clones.

Alternative splicing has previously been unknown in a GST transcript. It is possible that the transcripts of GSTM4 cloned from the cDNA library may represent either 1. An incomplete form of illegitimate transcription, a phenomenon recently described in which genes not normally transcribed in a tissue are, in fact, found in low copy number. 2. Secondly, these incomplete transcripts may represent experimental alternative splicing as part of a selection process. It has been postulated that alternative splicing may be an intermediate step in an evolutionary path eventually leading to novel protein isoforms or alterations in regulation allowing transcription in different cell types. 3. Finally, the incomplete transcripts may represent a novel form of regulation in this multigene family. Such transcripts have been observed to be a part of feedback mechanisms to regulate transcription.

This study has therefore provided evidence of a novel human Mu class cDNA and also of the possible alternative splicing of that gene. At the initiation of this study only one human Mu class sequence was known but this has changed rapidly over the last three years. The characterisation of a Yac clone containing all the human Mu class genes now characterised in separate studies, suggests that it is unlikely more Mu class GSTs will be found.

Apart from the present report, alternative splicing is unknown in the GSTs. It is necessary to determine if this is a general phenomenon or whether it is a specific property of the GSTM4 gene. The pattern of expression of GSTM4 in different tissues has not been clearly defined. Messenger RNA northern blots of the GSTM4 transcripts found in tissues may be able to detect if transcripts are all full length or if alternatively spliced forms are present in any significant quantity. In an initial survey of tissue specific mRNA, Comstock *et al.* (1993), observed hybridization to a GSTM4 probe in several tissues but the presence or absence of a shorter mRNA was not noted. Additionally, it was not clear if the GSTM4 probe was gene specific. However, even if alternatively spliced transcripts of GSTM4 are not found by these methods, this does not entirely preclude the possibility of alternative splicing in GSTM4. As has been detailed in Chapter 3, some examples of alternatively spliced transcripts have been difficult to detect because of either instability or extremely low copy number (Cohen *et al.* 1989,Chou *et al.* 1987, Zachar *et al.* 1987).

On a broader level, the cloning of associated genomic sequences for all the human Mu class is a distinct possibility. Evidence from the Yac clone reported by Pearson *et al.*(1993) suggests that this is a relatively small region. Very little is known about gene regulation in this family but in conjunction with tissue expression studies and analysis of the promoter regions, considerable advances could be expected.

8.2 Characterisation of GST isoenzymes

A combination of components from the alternatively spliced cDNA clones allowed the construction of a complete cDNA. The encoded novel GST protein, GSTM4, was expressed in *E. Coli*. Similar constructs were also made of three other human Mu class GST isoenzymes, GSTM1, GSTM2 and GSTM3. The catalytic activity and substrate specificity of the new isoenzyme, GSTM4, were ascertained and for the first time it was possible to compare the activity and characteristics of multiple recombinant human Mu class isoenzymes under identical conditions (Chapter 4). In general, the recombinant GSTM4-4 enzyme has relatively low activity with all the substrates tested and this may explain why it has not been detected previously. This study confirmed evidence presented in other studies that recombinant GSTM1 has significant catalytic activity with *trans*-stilbene oxide (Seidegård *et al.* 1984, Seidegård and Pero,1985, Seidegård *et al.* 1988). It also provided evidence that the other human Mu class isoenzymes tested had activity levels several orders of magnitude lower. Thus the substrate *trans*-stilbene oxide, is confirmed as a specific test for GSTM1 activity.

It is possible that the low activity exhibited by the recombinant GSTM4-4 is a factor of the choice of substrates in these experiments. The very high activity of GSTM1-1 with *trans*-stilbene oxide when compared to the other four enzymes suggests that specific substrates are yet to be found for many GST isoenzymes. It would also be of interest to test rodent Mu class isoenzymes for *trans*-stilbene oxide activity, to see if this is a species-specific trait.

This study was confined to the characterisation of the full length GSTM4 protein, however, it is possible that the shorter transcripts originally cloned still have a function within the cell, apart from a possible function in the regulation of transcription. These transcripts may have binding capabilities rather than catalytic activity, as the latter has been found to be seriously

reduced by removal of the residues of the C-terminal (as in the clone T7.3) and it is now clear from the 3D structure that exon 4 contains some of the residues required for substrate binding (as in the clone GST-T). Construction of expression clones of the alternatively spliced forms of GSTM4 may provide evidence of the level of catalytic activity and/or GSH binding ability of these forms.

The obvious differences among the human Mu class isoenzymes in substrate specificity and their generally low levels of catalytic activity suggest that a broader survey of substrates, and especially endogenous substrates, may provide a fuller understanding of the roles played by these isoenzymes within the cell. The further investigation of possible specific substrates for each isoenzyme warrants attention both for its own sake and because of the light such evidence may shed on the evolution of the second substrate binding site. At this time the human Mu class may be particularly well suited for the purpose. The cloning of the cDNAs of five of this class have now been reported and the characterisation of four of these in Chapter 3 suggests that very little sequence difference is required for large shifts in catalytic activity to be evident. It has been noted (Chapter 4.6) that there are very few amino acid differences that make GSTM1 unique within the class, in fact there are only three residues that are not conserved and are thought to take part in hydrophobic substrate binding. Site-directed mutagenesis studies are now under way to test these possibilities. These residues may be involved in defining the high specificity for *trans*-stilbene oxide.

The patterns of expression for most of the Mu class genes are not well defined. GSTM4 mRNA may be expressed in several tissues (Comstock *et al.* 1993) but it was not clear whether the probe was gene specific, a constant difficulty with the Mu class. A broader survey of tissue and developmental expression patterns of human Mu class isoenzymes, with the gene specific probes now available, is a reasonable prospect. This may provide insights into specific substrates and the reasons for such a multiplicity

of enzymes. It may also be profitable to use the gene specific probes to explore the rodent species for an equivalent array of Mu class genes.

8.3 Chromosomal mapping of the human Mu class

Chapter 5 describes the *in situ* hybridisation experiments aimed at determining the location of the human Mu class GSTs. The chromosomal localisation of the human Mu class GST genes has been complicated by three major factors, the uncertainty as to the total number of genes, the high level of sequence similarity among those cDNAs so far described and the polymorphism in GSTM1 which results from the presence or absence of the *GSTM1* gene. In this study a probe derived from GSTM4 has been shown to cross-hybridise with other members of the human Mu class. The GSTM1-1 status of individuals previously studied in gene mapping experiments has generally been unclear and led to some uncertainty as to the hybridizing locus. Importantly, the probe used in the present study was shown to be capable of distinguishing the *GSTM1*-null allele.

In situ hybridisation with the GSTM4 probe unequivocally localised the human Mu class GSTs to a cluster on chromosome band 1p13 contrary to previous reports of the possible dispersal of members of the class on several different chromosomes. The identical *In situ* hybridisation patterns in individuals with or without the *GSTM1* gene suggests that this locus is also a component of the Mu class GST gene cluster. This has been confirmed in a concurrent study by Pearson *et al.* (1993) of a Yac clone containing all five Mu class genes.

The use of proven locus specific probes for each of the known human Mu class in chromosomal hybridisation studies is, potentially, the best way of providing evidence of unequivocal localisation and this has now been reported by Pearson *et al.* (1993). There are, however, subsidiary questions remaining, about the relationships between the clustered Mu class locus on chromosome 1 and the peak found in this study on the short arm of

chromosome 6, an association also found in a previous study (DeJong *et al.* (1991). Although the present study found no evidence of an additional locus on chromosome 3, some cross-hybridization has been reported (Islam *et al.* 1989, Pearson *et al.* 1993). The low level of hybridization at these locations suggests the genes present may be pseudogenes or perhaps even more distantly related to the Mu class GSTs.

8.4 Analysis of evolutionary relationships among the GST multigene family

Multiple alignment of the GST sequences available from all species, incorporating evidence from the recent studies of 3D structure, has provided the basis for constructing phylogenetic trees and an evaluation of their reliability. By comparing all known full length amino acid sequences, two regions of increased similarity were found. These regions contain residues concerned with structural relationships such as subunit interactions and with domain interactions within the subunit, rather than being specifically concerned with the substrate binding site. However, of the two regions, the N-terminal region also contains some residues involved in the GSH-binding site.

The phylogenetic analyses in Chapter 6 show clear evidence of an early duplication of an ancestral GST gene and subsequent radiation of GSTs both within and between species. This duplication apparently occurred prior to the plant/animal split as sequences from animals appear in both sector A and sector B of the tree. Evidence is also provided for a pre-vertebrate origin for the Alpha, Mu and Pi classes which were originally defined in mammalian species. Within sector A (primarily containing mammalian sequences) a fifth class of GSTs, Sigma, proposed by Buetler and Eaton (1992), is also well supported by the data. There is less support, however, for the inclusion of the 28kDa schistosome sequences in the new, Sigma, class, as Buetler and Eaton (1992) originally proposed. Dipteran sequences segregate into two well defined groups, one in each sector of the trees,

providing some confirmation of the suggestion by Toung *et al.* (1990), that in insects a class system may be present analogous to that characterised in mammals. There is some suggestion from the present study that there may also be a class system within the plants.

GSTs may provide an estimation of divergence times among the species, however, there appears to be a consistent over-estimation of the dates of divergence when compared to dates determined from the fossil record or from other molecular studies. The former may be due to inaccuracies in the fossil record of the species divergence under investigation, and is a common trend in molecular studies whereas the latter may be due to either inaccuracies of the fossil record of the calibration date used or a bias created by using a single gene family rather than the more common practice of using many different genes. Relative rate tests of molecular evolution within the mammalian species represented show no evidence of a difference in substitution rates between the rodent and human lineages.

There is a definite need for the sampling bias in the characterisation of GSTs to be addressed for the purposes of evolutionary studies. Acquisition of more plant sequences and more intense study of individual plant species would provide evidence to test the proposition of an analogous class system within the plant kingdom. There are several anomalous relationships within the current trees, such as the fish sequence, PPGST1, and the yeast affinities, both of which may be clarified by the inclusion of additional sequences. The solution of one or more crystallographic structures of enzymes from within sector B of the major trees could greatly assist in the alignment of these very diverse sequences.

Within the Pi class, the relationships of the artiodactyl GST sequences, SSGSTP1 and BTGST1 (pig and cattle), are anomalous as they do not cluster together. The cattle sequence clusters with the human whereas the pig sequence (unfortunately available only as an amino acid sequence)

is more distant. the addition of other artiodactyl sequences and especially the pig nucleotide sequence may clarify this. GST data from the analysis of the Alpha class sequences supports the suggestion that rabbits are more closely related to primates than to rodents, still a somewhat controversial point in systematics. Rabbit sequences from the other classes could further test this proposition.

Representation in the mammalian Theta class is too small at present to provide any data on relationships within it, a situation that is bound to change rapidly in the near future, and this very divergent class may provide some interesting insights from its position within sector B.

Recently determined 3D structures in three GSTs have revealed a high degree of structural similarity (Reinemer *et al.* 1991, 1992, Ji *et al.* 1992, Sinning *et al.* 1993) that is not immediately obvious in a comparison of sequences across classes. Solution of the 3D structure of some sector B GSTs would provide significant data on which to improve the alignment of these divergent sequences. Analysis of the conserved positions in the light of the 3D structures may provide a basis for understanding their conservation. Conversely, inspection of the non-conserved regions may provide information about the parts involved in substrate binding in this super gene family, which differ greatly in substrate specificities. The sequences can represent a natural pool of mutant proteins which preserve catalytic function. Information gained through such comparisons is thus complimentary to mutagenesis studies. A specific example is the identification of residues which may be significant in the GSTM1 specificity for *trans*-stilbene oxide from the analysis of the human Mu class sequences (Chapter 4.6 and section 8.2).

Several studies have indicated that 3D structural conservation is not always evident in sequence comparisons (Lesk and Clothia 1980, Ollis *et al.* 1992, Cygler *et al.* 1993). Residues that play a role in stabilising the 3D structure of the protein and in orientation of the substrate and its binding site may be more conserved than those directly involved in substrate binding (Cygler *et*

al. 1993). The regions of most conservation in the GSTs include residues that are considered to be involved in structural stabilisation such as subunit domain and also dimer interactions. The most conserved of all residues in the current alignment (Pro56 as numbered in HSGSTA1) is situated on the junction of a loop and a b-strand (Sinning *et al.* 1993). Thus conservation of 3D structural framework may be more constrained than particular residues, except in specific instances, such as tight turns or difficult conformations.

The functions attributed to GST enzymes may partially explain the high level of diversity among the sequences as the broad range of hydrophobic substrates conjugated by GSTs might be expected to have had some influence. However, the binding of GSH is a constant feature of all these enzymes and thus the binding site for this compound may be under some constraint. Comparison of the substitution rates in the GSH-binding site and the hydrophobic-binding site when compared to the background rate in noncoding regions, may provide evidence to test this proposition, but these are very short regions, leading to large standard errors.

Alternatively, it may be that the scaffolding of the protein on which to correctly orientate residues for GSH-binding and a cleft of hydrophobic residues may be more conserved than the specific residues required for binding, such as a proton donor.

8.5 Origins of the GSTM1 polymorphism

The genetic deficiency of GSTM1-1 is widespread and occurs in similar frequencies in most population groups (Board 1990, Chapter 1.7.2). Southern blot experiments of genomic DNA from individuals with the deficiency show a loss of an approximately 8kb *Eco*RI fragment and appears to be the result of the absence of the GSTM1 gene (Chapter 5). No clear molecular explanation for this apparent deletion has yet been offered. Two basic models have been proposed for the origin of the GSTM1

polymorphism, a deletion model, widely assumed in the literature, and an insertion model (Board *et al.* 1993).

Both models assumed a relatively simple relationship among the human Mu class genes, however, exon-specific sequence analysis indicated a complex array of relationships within this class. The simplicity of the original GSTM2/GSTM4 chimeric insertion model and its predictions concerning the arrangement of the loci on the chromosome have not been substantiated by the detailed analysis. To account for the relationships evident in the sequence comparisons, it is necessary to invoke the process of gene conversion. There is evidence of gene conversion among many of the human Mu class, although GSTM1 appears to have the most complex pattern of relationships.

Several models were tested for both an insertion or deletion origin for the GSTM1 polymorphism. An insertion model using GSTM2/GSTM5 as donors in the creation of a chimeric GSTM1 gene and capable of accounting for all the observed phenomena was more complex than the deletion models. However, the deletion models do not explain all the relationships observed.

All but the inter-locus deletion model provide some predictions about the genes involved and his provides a basis for testing the validity of one model compared to another. The intra-locus deletion model suggests, for instance, that the sequence of GSTM2 should be different on a GSTM1-null chromosome to that on a GSTM1-positive chromosome and this could be tested. In the inter-locus deletion model, the length of sequence missing is not predicted but may not be substantially longer than the actual gene. However, it is possible that at least a small segment of inter-genetic sequence may be missing in the GSTM1-null that is present in the GSTM1-positive. The insertion models also make testable predictions. The insertion models suggest the possibility of some sequence repetition extending beyond the GSTM1 locus that would not be present on the GSTM1-null

chromosome. Thus sequence beyond the 5′ of GSTM1 should resemble the intergenic sequence in front of GSTM5 5′ and similarly the 3′ of GSTM1 should resemble 3′ intergenic sequence beyond GSTM2.

Intronic sequence would provide data for further analysis in the dating of the gene conversion events apparent in the Mu class. Noncoding sequence is thought to provide distance measures that more accurately reflect the stochastic substitution rate (Kimura 1983), without the effects of selection processes.

8.6 Conclusions

The results and discussion which constitute this thesis have fulfilled the initial aims of characterisation of further members of the human Mu class GST gene family and an increase in the understanding of evolutionary relationships within the GST multigene family. Comparison of sequences, in the light of 3D structural information may aid investigations of conserved positions and conversely, inspection of the non-conserved regions may provide information on substrate binding in this super gene family, which differ greatly in substrate specificities. The sequences can be used as a natural pool of mutant proteins which preserve catalytic function. Information gained through such comparisons is thus complimentary to mutagenesis studies. An interesting insight into the relationship between structure and function has been provided by the comparison of the human Mu class sequences suggesting the substantially higher *trans*-stilbene oxide activity of GSTM1 may be associated with as few as three amino acid residues.

APPENDIX I Alignment of all available GST amino acid sequences. Sequence designations and species are listed in Table 6.1.

```
MMGST7      M--------- ---------- ---------- ---------- ---------- ----------
RNGST7      M--------- ---------- ---------- ---------- ---------- ----------
HSGSTP1     M--------- ---------- ---------- ---------- ---------- ----------
BTGSTP1     M--------- ---------- ---------- ---------- ---------- ----------
SSGSTP1     M--------- ---------- ---------- ---------- ---------- ----------
CEGSTP1     M--------- ---------- ---------- ---------- ---------- ----------
MAGSTM1     M--------- ---------- ---------- ---------- ---------- ----------
HSGSTM4     M--------- ---------- ---------- ---------- ---------- ----------
HSGSTM5     M--------- ---------- ---------- ---------- ---------- ----------
HSGSTM1b    M--------- ---------- ---------- ---------- ---------- ----------
HSGSTM2     M--------- ---------- ---------- ---------- ---------- ----------
HSGSTM1a    M--------- ---------- ---------- ---------- ---------- ----------
HSGSTM3     M--------- ---------- ---------- ---------- ---------- ----------
RNGST4      M--------- ---------- ---------- ---------- ---------- ----------
RNGST3      M--------- ---------- ---------- ---------- ---------- ----------
CLGST1      M--------- ---------- ---------- ---------- ---------- ----------
GGGSTM2     M--------- ---------- ---------- ---------- ---------- ----------
MMGST3      M--------- ---------- ---------- ---------- ---------- ----------
MMGST4      M--------- ---------- ---------- ---------- ---------- ----------
SJGST1      M--------- ---------- ---------- ---------- ---------- ----------
SMGST1      M--------- ---------- ---------- ---------- ---------- ----------
FHGST1      M--------- ---------- ---------- ---------- ---------- ----------
OCGSTA1     M--------- ---------- ---------- ---------- ---------- ----------
OCGSTA2     M--------- ---------- ---------- ---------- ---------- ----------
HSGSTA1     M--------- ---------- ---------- ---------- ---------- ----------
HSGSTA2     M--------- ---------- ---------- ---------- ---------- ----------
MMGST2      M--------- ---------- ---------- ---------- ---------- ----------
MMGST1      M--------- ---------- ---------- ---------- ---------- ----------
RNGST2      M--------- ---------- ---------- ---------- ---------- ----------
MMGST8      M--------- ---------- ---------- ---------- ---------- ----------
RRGST8      M--------- ---------- ---------- ---------- ---------- ----------
RNGST1a     M--------- ---------- ---------- ---------- ---------- ----------
RNGST1b     M--------- ---------- ---------- ---------- ---------- ----------
GGGSTA3     M--------- ---------- ---------- ---------- ---------- ----------
SHGST1      M--------- ---------- ---------- ---------- ---------- ----------
SBGST1      M--------- ---------- ---------- ---------- ---------- ----------
SMGST2      M--------- ---------- ---------- ---------- ---------- ----------
ASGST1      M--------- ---------- ---------- ---------- ---------- ----------
DMelGST2    M--------- ---------- ---------- ---------- ---------- ----------
MDGST2      M--------- ---------- ---------- ---------- ---------- ----------
OVCRY1      M--------- ---------- ---------- ---------- ---------- ----------
OVCRY2      M--------- ---------- ---------- ---------- ---------- ----------
ODCRY1      M--------- ---------- ---------- ---------- ---------- ----------
ODCRY2      M--------- ---------- ---------- ---------- ---------- ----------
ODCRY3      M--------- ---------- ---------- ---------- ---------- ----------
ODCRY4      M--------- ---------- ---------- ---------- ---------- ----------
OSCRY1      M--------- ---------- ---------- ---------- ---------- ----------
OSCRY2      M--------- ---------- ---------- ---------- ---------- ----------
OSCRY3      M--------- ---------- ---------- ---------- ---------- ----------
FBDH        M--------- ---------- ---------- ---------- ---------- ----------
PPGST1      M--------- ---------- ---------- ---------- ---------- ----------
IOGST1      M--------- ---------- ---------- ---------- ---------- ----------
SacCURE2    MMNNNGNQVS NLSNALRQVN IGNRNSNTTT DQSNINFEFS TGVNNNNNNN SSSNNNNVQN
ZMGST3      M--------- ---------- ---------- ---------- ---------- ----------
ZMGST1      M--------- ---------- ---------- ---------- ---------- ----------
TAGST1      M--------- ---------- ---------- ---------- ---------- ----------
TAGST2      M--------- ---------- ---------- ---------- ---------- ----------
DCGST1      M--------- ---------- ---------- ---------- ---------- ----------
ATGST2      M--------- ---------- ---------- ---------- ---------- ----------
ATGST1      M--------- ---------- ---------- ---------- ---------- ----------
ATGST3      M--------- ---------- ---------- ---------- ---------- ----------
NTGST1      M--------- ---------- ---------- ---------- ---------- ----------
SCGST1      M--------- ---------- ---------- ---------- ---------- ----------
HSGSTMIC    M--------- ---------- ---------- ---------- ---------- ----------
RNGSTMIC    M--------- ---------- ---------- ---------- ---------- ----------
DMelGST1    M--------- ---------- ---------- ---------- ---------- ----------
 DMelGST3   M--------- ---------- ---------- ---------- ---------- ---------- -
ECSSPB      M--------- ---------- ---------- ---------- ---------- ----------
MDGST1      M--------- ---------- ---------- ---------- ---------- ----------
RNGST5      M--------- ---------- ---------- ---------- ---------- ----------
RNGST12     M--------- ---------- ---------- ---------- ---------- ----------
LCGST1      M--------- ---------- ---------- ---------- ---------- ----------
MBDMD       M--------- ---------- ---------- ---------- ---------- ----------
```

```
MMGST7      ---------- ---------- ---------- ---------- ---------- -PPYTIVYFP
RNGST7      ---------- ---------- ---------- ---------- ---------- -PPYTIVYFP
HSGSTP1     ---------- ---------- ---------- ---------- ---------- -PPYTVVYFP
BTGSTP1     ---------- ---------- ---------- ---------- ---------- -PPYTIVYFP
SSGSTP1     ---------- ---------- ---------- ---------- ---------- -PPYTITYFP
CEGSTP1     ---------- ---------- ---------- ---------- ---------- --TLKLTYFD
MAGSTM1     ---------- ---------- ---------- ---------- ---------- --PVTLGYWD
HSGSTM4     ---------- ---------- ---------- ---------- ---------- --SMTLGYWD
HSGSTM5     ---------- ---------- ---------- ---------- ---------- --PMTLGYWD
HSGSTM1b    ---------- ---------- ---------- ---------- ---------- --PMILGYWD
HSGSTM2     ---------- ---------- ---------- ---------- ---------- --PMTLGYWN
HSGSTM1a    ---------- ---------- ---------- ---------- ---------- --PMILGYWD
HSGSTM3     ---------- ---------- ---------- ---------- --------SC ESSMVLGYWD
RNGST4      ---------- ---------- ---------- ---------- ---------- --PMTLGYWD
RNGST3      ---------- ---------- ---------- ---------- ---------- --PMILGYWN
CLGST1      ---------- ---------- ---------- ---------- ---------- --PMILGYWN
GGGSTM2     ---------- ---------- ---------- ---------- ---------- --VVTLGYWD
MMGST3      ---------- ---------- ---------- ---------- ---------- --PMILGYWN
MMGST4      ---------- ---------- ---------- ---------- ---------- --PMTLGYWD
SJGST1      ---------- ---------- ---------- ---------- ---------- --SPILGYWK
SMGST1      ---------- ---------- ---------- ---------- ---------- --APKFGYWK
FHGST1      ---------- ---------- ---------- ---------- ---------- --PAKLGYWK
OCGSTA1     ---------- ---------- ---------- ---------- ---------- ARKPLLHYFN
OCGSTA2     ---------- ---------- ---------- ---------- ---------- AGKPKLHYFN
HSGSTA1     ---------- ---------- ---------- ---------- ---------- AEKPKLHYFN
HSGSTA2     ---------- ---------- ---------- ---------- ---------- AEKPKLHYSN
MMGST2      ---------- ---------- ---------- ---------- ---------- AGKPVLHYFD
MMGST1      ---------- ---------- ---------- ---------- ---------- AGKPVLHYFN
RNGST2      ---------- ---------- ---------- ---------- ---------- PGKPVLHYFD
MMGST8      ---------- ---------- ---------- ---------- ---------- AAKPKLYYFN
RRGST8      ---------- ---------- ---------- ---------- ---------- EVKPKLYYFQ
RNGST1a     ---------- ---------- ---------- ---------- ---------- SGKPVLHYFN
RNGST1b     ---------- ---------- ---------- ---------- ---------- SGKPVLHYFN
GGGSTA3     ---------- ---------- ---------- ---------- ---------- AAKPVLYYFN
SHGST1      ---------- ---------- ---------- ---------- --------MT GDHIKVIYFN
SBGST1      ---------- ---------- ---------- ---------- ---------T GDHIKVIYFN
SMGST2      ---------- ---------- ---------- ---------- ---------A GEHIKVIYFD
ASGST1      ---------- ---------- ---------- ---------- ---------- -PQYKLTYFD
DMelGST2    -ADEAQAP-P AEG---APPA EGEAPPPAEG AEGAVEGGEA APPAEPAEPI KHSYTLFYFN
MDGST2      -ADEAPAAPP AEGEAPAAPA EGEAPPPAEG EAP------- -----PAEPV KNTYTLFYFN
OVCRY1      ---------- ---------- ---------- ---------- ---------- -PSYTLHYFN
OVCRY2      ---------- ---------- ---------- ---------- ---------- -PSYTLHYFN
ODCRY1      ---------- ---------- ---------- ---------- ---------- -PSYTLNYFN
ODCRY2      ---------- ---------- ---------- ---------- ---------- -PSYTLNYFN
ODCRY3      ---------- ---------- ---------- ---------- ---------- -PSYTLHYFN
ODCRY4      ---------- ---------- ---------- ---------- ---------- -PSYTLHYFN
OSCRY1      ---------- ---------- ---------- ---------- ---------- -PNYTLYYFN
OSCRY2      ---------- ---------- ---------- ---------- ---------- -PSYTLYYFN
OSCRY3      ---------- ---------- ---------- ---------- ---------- -PKYTLYYFN
FBDH        ---------- ---------- ---------- ---------- ---------- -PEVSL-YNY
PPGST1      ---------- ---------- ---------- ---------- ---------- AKDMTLLW--
IOGST1      ---------- ---------- ---------- ---------- ---------- --TFATVYIK
SacCURE2    NNSGRNGSQN NDNENNIKNT LEQHRQQQQA FSDMSHVEYS RITKFFQEQP LEGYTL-FSH
ZMGST3      ---------- ---------- ---------- ---------- ---------- -APLKL-YGM
ZMGST1      ---------- ---------- ---------- ---------- ---------- -APMKL-YGA
TAGST1      ---------- ---------- ---------- ---------- ---------- -SPVKV-FGH
TAGST2      ---------- ---------- ---------- ---------- ---------- -SPVKV-FGH
DCGST1      ---------- ---------- ---------- ---------- ------SSSE TQKMQL-YSF
ATGST2      ---------- ---------- ---------- ---------- ---------- -VTVKL-YGM
ATGST1      ---------- ---------- ---------- ---------- ---------- -AGIKV-FGH
ATGST3      ---------- ---------- ---------- ---------- ---------- -AGIKV-FGH
NTGST1      ---------- ---------- ---------- ---------- ---------- --AIKV-HGS
SCGST1      ---------- ---------- ---------- ---------- ---------- --TIKV-HGN
HSGSTMIC    ---------- ---------- --VDLTQVMD DEVFMAFASY ATIILSKMML MSTATAFYRL
RNGSTMIC    ---------- ---------- --ADLKQLMD NEVLMAFTSY ATIILAKMMF LSSATAFQRL
DMelGST1    ---------- ---------- ---------- ---------- ---------- ----VDFYYL
DMelGST3    ---------- ---------- ---------- ---------- ---------- -----DFYYH
ECSSPB      ---------- ---------- ---------- ---------- --AVAAN--K RSVMTL-FSG
MDGST1      ---------- ---------- ---------- ---------- ---------- -----DFYYL
RNGST5      ---------- ---------- ---------- ---------- ---------- --VLEL-YLD
RNGST12     ---------- ---------- ---------- ---------- ---------- --GLEL-YLD
LCGST1      ---------- ---------- ---------- ---------- ---------- -----DFYYL
MBDMD       ---------- ---------- ---------- ---------- -SPNPTNIHT GKTLRLLYH-
```

```
MMGST7      VRGRC----E AMRMLLADQ- --GQS-WKE- --EVVTI--- -----DT--- -WMQGLLKPT
RNGST7      VRGRC----E ATRMLLADQ- --GQS-WKE- --EVVTI--- -----DV--- -WLQGSLKST
HSGSTP1     VRGRC----A ALRMLLADQ- --GQS-WKE- --EVVTV--- -----ET--- -WQEGSLKAS
BTGSTP1     VQGRC----E AMRMLLADQ- --GQS-WKE- --EVVAM--- -----QS--- -WLQGPLKAS
SSGSTP1     VRGRC----E AMRMLLADQ- --DQS-WKE- --EVVTM--- -----ET--- -W--PPLKPS
CEGSTP1     IHGLA----E PIRLLLADK- --QVA-YED- --HRVTY--- -----EQ--- -W--ADIKPK
MAGSTM1     IRGLA----H AIRLLLEYT- --DTS-YEE- -KKYTMGDAP NFDR-SQ--- -WLNEKFKLG
HSGSTM4     IRGLA----H AIRLLLEYT- --DSS-YEE- -KKYTMGDAP DYDR-SQ--- -WLNEKFKLG
HSGSTM5     IRGLA----H AIRLLLEYT- --DSS-YVE- -KKYTMGDAP DYDR-SQ--- -WLNEKFKLG
HSGSTM1b    IRGLA----H AIRLLLEYT- --DSS-YEE- -KKYTMGDAP DYDR-SQ--- -WLNEKFKLG
HSGSTM2     IRGLA----H SIRLLLEYT- --DSS-YEE- -KKYTMGDAP DYDR-SQ--- -WLNEKFKLG
HSGSTM1a    IRGLA----H AIRLLLEYT- --DSS-YEE- -KKYTMGDAP DYDR-SQ--- -WLNEKFKLG
HSGSTM3     IRGLA----H AIRLLLEFT- --DTS-YEE- -KRYTCGEAP DYDR-SQ--- -WLDVKFKLD
RNGST4      IRGLA----H AIRLLLEYT- --DSS-YEE- -KRYTMGDAP DFDR-SQ--- -WLNEKFKLG
RNGST3      VRGLT----H PIRLLLEYT- --DSS-YEE- -KRYAMGDAP DYDR-SQ--- -WLNEKFKLG
CLGST1      VRGLT----N PIRLLLEYT- --DSS-YEE- -KKYTMGDAP DSDR-SQ--- -WLNEKFKLG
GGGSTM2     IRGLA----H AIRLLLEYT- --ETP-YQE- -RRYKAGPAP DFDP-SD--- -WTNEKEKLG
MMGST3      VRGLT----H PIRMLLEYT- --DSS-YDE- -KRYTMGDAP DFDR-SQ--- -WLNEKFKLG
MMGST4      IRGLA----H AIRLLLEYT- --DTS-YED- -KKYTMGDAP DYDR-SQ--- -WLSEKFKLG
SJGST1      IKGLV----Q PTRLLLEYL- --EEK-YEE- -HLYER-D-- EGDK------ -WRNKKFELG
SMGST1      VKGLV----Q PTRLLLEHL- --EET-YEE- -RAYDR-N-- EIDA------ -WSNDKFKLG
FHGST1      IRGLQ----Q PVRLLLEYL- --GEE-YEE- -HLYGR-D-- DREK------ -WFGDKFNMG
OCGSTA1     GRGRM----E SIRWLLAAA- --GEE-FDE- --KFMET--- -----AED-- -LDKLRNDGS
OCGSTA2     ARGRM----E SIRWLLTAA- --GVE-FEE- --KCMKT--- -----RED-- -LEKLRKDGV
HSGSTA1     ARGRM----E STRWLLAAA- --GVE-FEE- --KFIKS--- -----AED-- -LDKLRNDGY
HSGSTA2     IRGRM----E SIRWLLAAA- --GVE-FEE- --KFIKS--- -----AED-- -LDKLRNDGY
MMGST2      GRGRM----E PIRWLLAAA- --GVE-FEE- --KFLKT--- -----RDD-- -LARLRSDGS
MMGST1      ARGRM----E CIRWLLAAA- --GVE-FEE- --KFIQS--- -----PED-- -LEKLKKDGN
RNGST2      GRGRM----E PIRWLLAAA- --GVE-FEE- --QFLKT--- -----RDD-- -LARLRNDGS
MMGST8      GRGRM----E SIRWLLAAA- --GVE-FEE- --EFLET--- -----REQ-- -YEKMQKDGH
RRGST8      GRGRM----E SIRWLLATA- --GVE-FEE- --EFLET--- -----REQ-- -YEKLQKDGC
RNGST1a     ARGRM----E CIRWLLAAA- --GVE-FEE- --KLIQS--- -----PED-- -LEKLKKDGN
RNGST1b     ARGRM----E CIRWLLAAA- --GVE-FDE- --KFIQS--- -----PED-- -LEKLKKDGN
GGGSTA3     GRGKM----E SIRWLLAAA- --GVE-FEE- --VFLET--- -----REQ-- -YEKLLQSGI
SHGST1      GRGRA----E SIRMTLVAA- --GVN-YED- --ERISF--- -----QD--- -WPKIK--PT
SBGST1      GRGRA----E SIRMTLVAA- --GVN-YED- --ERISF--- -----QD--- -WPKIK--PT
SMGST2      GRGRA----E SIRMTLVAA- --GVD-YED- --ERISF--- -----QD--- -WPKIK--PT
ASGST1      IRGLG----E GARLIFHQA- --GVK-FED- --NRLKR--- -----ED--- -WPALK--PK
DMelGST2    VKALP----S PCATC-SDG- --NQE-YED- ----VA---- -----HPR-- RVPALK--PT
MDGST2      VKALA----E PLRYLFAYG- --GIE-YED- --VRVT---- -----RDE-- -WPALK--PT
OVCRY1      HRGRARSAVC CSQLLVSSTM --TAD----- --RVFRM--- -----GQHEK -QDAMSHDAN
OVCRY2      HRGRA----E ICRMLFAAA- --GVQ-YND- --RRVDC--- -----SE--- -WTGMR--NQ
ODCRY1      HRGRA----E ICRMLFAAA- --GVQ-YND- --RRVDC--- -----SE--- -WTGMK--TQ
ODCRY2      HRGRA----E ICRMLFAAA- --GVQ-YND- --RRIET--- -----SE--- -WSNMR--SK
ODCRY3      HRGRA----E ICRMLFAAA- --GVQ-YND- --RRIES--- -----SE--- -WNGMR--NQ
ODCRY4      HRGRA----E ICRMLFAAA- --GVQ-YND- --RRIES--- -----SE--- -WGSMR--SK
OSCRY1      GRGRA----E ICRMLMAAA- --GVQ-YTD- --KRFEF--- -----NE--- -WDKYR--ND
OSCRY2      GRGRA----E ICRMLFAVA- --SVQ-YQD- --KRIEL--- -----AE--- -WTQFK--TK
OSCRY3      SRGRA----E ICRMLFAAA- --NIP-YND- --VRIDY--- -----SE--- -WDIYR--SK
FBDH        TMSICS---M KTRLA--ME- EFGVD-YDD- --KQVDIGFA ----LEN--- -FE-PDYVRL
PPGST1      GSGSP----P CWRVMIVLE- EKNLQAYNS- --KLLSF--- ---EKGEHKS ----AEVMSM
IOGST1      P--------H TPR--GDWLA SLG-Q-YVGL EIKTVDY--- ---KSAEASK ----FEELF-
SacCURE2    RSAP-----N GFKVAIVLS- ELGFH-YNT- --IFLDF--- ---NLGEHRA ----PEFVSV
ZMGST3      PLSP-----N VVRVATVLN- EKGLD-FEIV ---PVDL--- ---TTGAHKQ ----PDFLAL
ZMGST1      VMSW-----N LTRCATALE- EAGSD-YEIV ---PINF--- ---ATAEHKS ----PEHLVR
TAGST1      PMLT-----N VARVLLFLE- EVGAE-YELV ---PMDF--- ---VAGEHKR ----PQHVQL
TAGST2      PMLT-----N VARVLLFLE- EVGAE-YELV ---PVDF--- ---VAGEHKR ----PQHVQL
DCGST1      SLSS-----C AWRVRIALH- LKGLD-FEYK ---AVDL--- ---FKGEHLT ----PEFLKL
ATGST2      AYST-----C TKRVYTTAK- EIGVD-VKIV ---PVDL--- ---MKGEHKE ----PAYLDN
ATGST1      PASI-----A TRRVLIALH- EKNLD-FELV ---HVEL--- ---KDGEHKK ----EPFLSR
ATGST3      PAST-----A TRRVLIALH- EKNVD-FEFV ---HVEL--- ---KDGEHKK ----EPFILR
NTGST1      PMST-----A TMRVAACLI- EKELD-FEFV ---PVDM--- ---ASGEHKK ----HPYLSL
SCGST1      PRST-----A TQRVLVALY- EKHLE-FEFV ---PIDM--- ---GAGGHKQ ----PSYLAL
HSGSTMIC    --TRK-VFAN PEDCVA-FGK --GEN----- A-KKYLRTDD ----RVERVR R---AHLNDL
RNGSTMIC    --TNK-VFAN PEDCAG-FGK --GEN----- A-KKFLRTDE ----KVERVR R---AHLNDL
DMelGST1    --PGS----S PCRSVIMTAK AVGVE----L N-KKLLNLQA -----GEHLK ----PEFLKI
DMelGST3    --PCS----A PCRSVIMTAK ALGVD----L N-MKLLKVMD -----GEQLK ----PEFVKL
ECSSPB      --PTD-IYSH QVR-IVLAEK --GVS-FE-I E--HVEKDNP -----PQDL- -------IDL
MDGST1      --PGS----A PCRSVLMTAK ALGIE----L N-KKLLNLQA -----GEHLK ----PEFLKI
RNGST5      --LLS----Q PCRAIYIFAK KNNIP-FQ-M --HTVELRK- -----GEHLS ----DAFAQV
RNGST12     --LLS----Q PSRAVYIFAK KNGIP-FQ-L --RTVDLLK- -----GQHLS ----EQFSQV
LCGST1      --PGS----T PYHSVLMTAK ALGIE----L T-KKLLNLQA -----GEHLK ----PEFLKI
MBDMD       --PAS----Q PCRSAHQFMY EIDVP-FEE- --EVVDISTD ----ITERQ- -----EFRDK
```

```
MMGST7      C--------- ---LYGQL-- ---PKFED-G D---LT--LY Q-SNAILRHL GRSLG-----
RNGST7      C--------- ---LYGQL-- ---PKFED-G D---LT--LY Q-SNAILRHL GRSLG-----
HSGSTP1     C--------- ---LYGQL-- ---PKFQD-G D---LT--LY Q-SNTILRHL GRTLG-----
BTGSTP1     C--------- ---LYGQL-- ---PKFQD-G D---LT--LY Q-SNAILRHL GRTLG-----
SSGSTP1     C--------- ---LFRQL-- ---PKFQD-G D---LT--LY Q-SNAILRHL GRSFG-----
CEGSTP1     M--------- ---IFGQV-- ---PCLLS-G D---EE--IV Q-SGAIIRHL ARLNG-----
MAGSTM1     L--------- ---DFPNL-- ---PYLID-G S---HK--IT Q-SNAILRYI ARKHD-----
HSGSTM4     L--------- ---DFPNL-- ---PYLID-G A---HK--IT Q-SNAILCYI ARKHN-----
HSGSTM5     L--------- ---DFPNL-- ---PYLID-G A---HK--IT Q-SNAILRYI ARKHN-----
HSGSTM1b    L--------- ---DFPNL-- ---PYLID-G A---HK--IT Q-SNAILCYI ARKHN-----
HSGSTM2     L--------- ---DFPNL-- ---PYLID-G T---HK--IT Q-SNAILRYI ARKHN-----
HSGSTM1a    L--------- ---DFPNL-- ---PYLID-G A---HK--IT Q-SNAILCYI ARKHN-----
HSGSTM3     L--------- ---DFPNL-- ---PYLLD-G K---NK--IT Q-SNAILRYI ARKHN-----
RNGST4      L--------- ---DFPNL-- ---PYLID-G S---HK--IT Q-SNAILRYL GRKHN-----
RNGST3      L--------- ---DFPNL-- ---PYLID-G S---RK--IT Q-SNAIMRYL ARKHH-----
CLGST1      L--------- ---DFPNL-- ---PYLID-G S---HK--IT Q-SNAILRYI ARKHN-----
GGGSTM2     L--------- ---DFPNL-- ---PYLID-G D---VK--LT Q-SNAILRYI ARKHN-----
MMGST3      L--------- ---DFPNL-- ---PYLID-G S---HK--IT Q-SNAILRYL ARKHH-----
MMGST4      L--------- ---DFPNL-- ---PYLID-G S---HK--IT Q-SNAILRYL ARKHN-----
SJGST1      L--------- ---EFPNL-- ---PYYID-G D---VK--LT Q-SMAIIRYI ADKHN-----
SMGST1      L--------- ---EFPNL-- ---PYYID-G D---FK--LT Q-SMAIIRYI ADKHN-----
FHGST1      L--------- ---DLPNL-- ---PYYID-D K---CK--LT Q-SVAIMRYI ADKHG-----
OCGSTA1     L--------- ---MYQQV-- ---PMVEI-D G---MK--LV Q-TRAILNYV ANKHN-----
OCGSTA2     L--------- ---MFQQV-- ---PMVEI-D G---MK--LV Q-TRAIFNYI ADKHN-----
HSGSTA1     L--------- ---MFQQV-- ---PMVEI-D G---MK--LV Q-TRAILNYI ASKYN-----
HSGSTA2     L--------- ---MFQQV-- ---PMVEI-D G---MK--LV Q-TRAILNYI ASKYN-----
MMGST2      L--------- ---MFQQV-- ---PMVEI-D G---MK--LV Q-TKAILNYI ASKYN-----
MMGST1      L--------- ---MFDQV-- ---PMVEI-D G---MK--LV Q-TRAILNYI ATKYD-----
RNGST2      L--------- ---MFQQV-- ---PMVEI-D G---MK--LV Q-TRAILNYI ATKYN-----
MMGST8      L--------- ---LFGQV-- ---PLVEI-D G---MM--LT Q-TRAILSYL AAKYN-----
RRGST8      L--------- ---LFGQV-- ---PLVEI-D G---ML--LT Q-TRAILSYL AAKYN-----
RNGST1a     L--------- ---MFDQV-- ---PMVEI-D G---MK--LA Q-TRAILNYI ATKYD-----
RNGST1b     L--------- ---MFDQV-- ---PMVEI-D G---MK--LA Q-TRAILNYI ATKYD-----
GGGSTA3     L--------- ---MFQQV-- ---PMVEI-D G---MK--LV Q-TRAILNYI AGKYN-----
SHGST1      I--------- ---PGGRL-- ---PAVKITD NHGHVKW-MV E-SLAIARYM AKKHH-----
SBGST1      I--------- ---PGGRL-- ---PAVKITD NHGHVKW-ML E-SLAIARYM AKKHH-----
SMGST2      I--------- ---PGGRL-- ---PAVKVTD DHGHVKW-ML E-SLAIARYM AKKHH-----
ASGST1      T--------- ---PFGQL-- ---PLLEV-D G---EV--LA Q-SAAIYRYL GRQFG-----
DMelGST2    M--------- ---PMGQM-- ---PVLEV-D G---KR--VH Q-SISMARFL AKTVG-----
MDGST2      M--------- ---PMGQM-- ---PVLEV-N G---KR--VH Q-SISMARFL AKTVG-----
OVCRY1      V--------- ---GIGQQ-- --NP-----N S--------P E--YAMARYL AREFG-----
OVCRY2      M--------- ---PCSMM-- ---PMLEI-D N---RHQ-IP Q-SMAIARYL AREFG-----
ODCRY1      M--------- ---PCSMM-- ---PMLEI-D N---RHQ-IP Q-SMAIARYL AREFG-----
ODCRY2      M--------- ---PCSMM-- ---PMLDI-D N---RHQ-IP Q-TMAIARYL AREFG-----
ODCRY3      M--------- ---PCNMM-- ---PMLEL-D N---RTQ-IP Q-SMAMARYL AREFG-----
ODCRY4      M--------- ---PCSMM-- ---PMLEL-D N---KIQ-IP Q-SMAMARYL AREFG-----
OSCRY1      M--------- ---PSMCV-- ---PVLDI-D G---QNK-MP E-TMAIARYL ARENG-----
OSCRY2      M--------- ---PCHML-- ---PILEI-D T---ETQ-VP Q-SMAISRYL AREFG-----
OSCRY3      M--------- ---PGSCL-- ---PVLEI-N D---SIQ-IP Q-TMAIARYL ARQFG-----
FBDH        N--------- ---EKAVV-- ---PTLVV-G D---RV--VT N-SYNIV--L EAANV-GKVG
PPGST1      N--------- ---PRGQL-- ---PSFKH-G S---KV--LN E-SYAACMYL ESQFK-SQ-G
IOGST1      ---------- ---PLKRV-- ---PALVT-P N---GFQ-LT E-LIAIVEYI VAK-G-SK-P
SacCURE2    N--------- ---PNARV-- ---PALIDHG MDN-LS--IW E-SGAILLHL VNKYY-KETG
ZMGST3      N--------- ---PFGQI-- ---PALVD-G D---EV--LF E-SRAINRYI ASKYA-SEGT
ZMGST1      N--------- ---PFGQV-- ---PALQD-G D---LY--LF E-SRAICKYA ARKNK-PE--
TAGST1      N--------- ---PFAKM-- ---PGFQD-G D---LV--LF E-SRAIAKYI LRKYGGTAGL
TAGST2      N--------- ---PFAKM-- ---PGFQD-G E---SL--HI K-SRAIAKYI LRKYGGTAGL
DCGST1      N--------- ---PLGYV-- ---PVLVH-G D---IV--IA D-SLAIIMYL EEKF--PE-N
ATGST2      YH-------- ---PFGVI-- ---PVLED-E DG--TK--IY E-SRAISRYL VAKY--GKGS
ATGST1      N--------- ---PFGQV-- ---PAFED-G D---LK--LF E-SRAITQYI AHRYE-NQGT
ATGST3      N--------- ---PFGKV-- ---PAFED-G D---FK--IF E-SRAITQYI AHEFS-DKGN
NTGST1      N--------- ---PFGQV-- ---PAFED-G D---LK--LF E-SRAITQYI AHVYA-DNGY
SCGST1      N--------- ---PFGQV-- ---PALED-G E---IK--LF E-SRAITKYL AYTHDHQNEG
HSGSTMIC    EN-------- ------II-- ---PFLGI-G LL--YSLSGP DPSTAILHFR L--F----VG
RNGSTMIC    EN-------- ------IV-- ---PFLGI-G LL--YSLSGP DLSTALIHFR I--F----VG
DMelGST1    N--------- ---PQHTI-- ---PTLVD-N G---FA--LW E-SRAIQVYL VEKY--GK-T
DMelGST3    N--------- ---PQHCI-- ---PTLVD-D G---FS--IW E-SRAILIYL VEKY--GA-D
ECSSPB      N--------- ---PNQSV-- ---PTLVD-R E---LT--LW E-SRIIMEYL DERF--PH-P
MDGST1      N--------- ---PQHTI-- ---PTLVD-G D---FA--LW E-SRAIMVYL VEKY--GK-T
RNGST5      N--------- ---PMKKV-- ---PAMKD-G G---FTL--C E-SVAILLYL AHKYK---VP
RNGST12     N--------- ---CLKKV-- ---PVLKD-G S---FVL--T E-STAILIYL SSKYQ---VA
LCGST1      N--------- ---PQHTI-- ---PTLVD-G D---FA--LW E-SRAIMVYL VEKY--GK-N
MBDMD       YN-------- ---PTGQV-- ---PILVD-G E---FT--VW E-SVAIARYV NEKFD-GA-G
```

```
MMGST7      ---LYGKN-- -QREAAQMDM VN---DGVED L---RGKYVT LI----YTN- ----------
RNGST7      ---LYGKD-- -QKEAALVDM VN---DGVED L---RCKYGT LI----YTN- ----------
HSGSTP1     ---LYGKD-- -QQEAALVDM VN---DGVED L---RCKYIS LI----YTN- ----------
BTGSTP1     ---LYGKD-- -QQEAALVDM VN---DGVED L---RCKYVS LI----YTN- ----------
SSGSTP1     ---LYGKD-- -QKEAALVDM VN---DGVED L---RCKYAT LI----YTN- ----------
CEGSTP1     ---LNGSN-- -ETETTFIDM FY---EGLRD L---HTKYTT MI----YRN- ----------
MAGSTM1     ---LCGET-- -EEERIQLDI LE---NQAMD T---RMQLAM VC----YSP- ----------
HSGSTM4     ---LCGET-- -EEEKIRVDI LE---NQAMD V---SNQLAR VC----YSP- ----------
HSGSTM5     ---LCGET-- -EEEKIRVDI LE---NQVMD N---HMELVR LC----YDP- ----------
HSGSTM1b    ---LCGET-- -EEEKIRVDI LE---NQTMD N---HMQLGM IC----YNP- ----------
HSGSTM2     ---LCGES-- -EKEQIREDI LE---NQFMD S---RMQLAK LC----YDP- ----------
HSGSTM1a    ---LCGET-- -EEEKIRVDI LE---NQTMD N---HMQLGM IC----YNP- ----------
HSGSTM3     ---MCGET-- -EEEKIRVDI IE---NQVMD F---RTQLIR LC----YSS- ----------
RNGST4      ---LCGET-- -EEERIRVDI LE---NQLMD N---RMVLAR LC----YNP- ----------
RNGST3      ---LCGET-- -EEERIRADI VE---NQVMD N---RMQLIM LC----YNP- ----------
CLGST1      ---LCGET-- -EEERIRVDI VE---NQAMD T---RMQLIM LC----YNP- ----------
GGGSTM2     ---MCGET-- -EVEKQRVDV LE---NHLMD L---RMAFAR LC----YSP- ----------
MMGST3      ---LDGET-- -EEERIRADI VE---NQVMD T---RMQLIM LC----YNP- ----------
MMGST4      ---LCGET-- -EEERIRVDI LE---NQAMD T---RIQLAM VC----YSP- ----------
SJGST1      ---MLGGC-- -PKERAEISM LE---GAVLD I---RYGVSR IA----YSK- ----------
SMGST1      ---MLGAC-- -PKERAEISM LE---GAVLD I---RMGVLR IA----YNK- ----------
FHGST1      ---MLGTT-- -PEERARISM IE---GAAMD L---RMGFVR VC----YNP- ----------
OCGSTA1     ---LYGKD-- -MKERALIDM YT---EGVAD L---YELVLL LP----LCP- ----------
OCGSTA2     ---LYGKD-- -IKERALIDM YT---EGIVD L---NELILT RP----FLP- ----------
HSGSTA1     ---LYGKD-- -IKERALIDM YI---EGIAD L---GEMILL LP----VCP- ----------
HSGSTA2     ---LYGKD-- -IKEKALIDM YI---EGIAD L---GEMILL LP----FTQ- ----------
MMGST2      ---LYGKD-- -MKERAIIDM YT---EGVAD L---EIMILY YP----HMP- ----------
MMGST1      ---LYGKD-- -MKERALIDM YT---EGILD L---TEMIGQ LV----LCP- ----------
RNGST2      ---LYGKD-- -MKERALIDM YA---EGVAD L---DEIVLH YP----YIP- ----------
MMGST8      ---LYGKD-- -LKERVRIDM YA---DGTQD L---MMMIAV AP----FKT- ----------
RRGST8      ---LYGKD-- -LKERVRIDM YA---DGTQD L---MMMIIG AP----FKA- ----------
RNGST1a     ---LYGKD-- -MKERALIDM YS---EGILD L---TEMIIQ LV----ICP- ----------
RNGST1b     ---LYGKD-- -MKERALIDM YT---EGILD L---TEMIMQ LV----ICP- ----------
GGGSTA3     ---LYGKD-- -LKERALIDM YV---GGTDD L---MGFLLS FP----FLS- ----------
SHGST1      ---MMGGT-- -EEEYYNVEK LI---GQAED L---EHEYYK TL----MKP- ----------
SBGST1      ---MMGET-- -DEEYYNVEK LI---GQVED L---EHEYHK TL----MKP- ----------
SMGST2      ---MMGET-- -DEEYYSVEK LI---GQAED V---EHEYHK TL----MKP- ----------
ASGST1      ---LAGKT-- -PMEEAQVDS IF---DQFKD F---MAEL-R PC----FRVL AG--------
DMelGST2    ---LCGAT-- -PWEDLQIDI VV---DTIND F---RLKIAV VS----YEP- ----------
MDGST2      ---LCGAT-- -PWEDLQVDI VV---DTIND F---RLKIAV VS----YEP- ----------
OVCRY1      ---FHGRN-- -NMEMARVDF IS---DCFYD I---LDDYMR MYQDGNCRMM F---------
OVCRY2      ---FHGRN-- -NLDMARVDF IS---DCFYD I---LDDYLR MYHDKDGRMM F---------
ODCRY1      ---FHGRN-- -NLDMARVDF IS---DCFYD I---MDDYMR MYHDKDGKMM F---------
ODCRY2      ---FHGKN-- -NMEMARVEY IS---DCFYD I---LDDYLR MYQDDNCRMM F---------
ODCRY3      ---YHGKS-- -NMEMARVDF IS---DCFYD I---MDDYMR MYQDGNCRMM F---------
ODCRY4      ---FHGKN-- -NMDMARVDY IS---DSFYD I---LDDYMR MYHDKDGRMM F---------
OSCRY1      ---YYGKN-- -NMDMFRIDY IC---DCFYE I---LHDYMR YFHTKNGR-- ----------
OSCRY2      ---FYGKN-- -NMDMFKVDC LC---DSLFE L---FNDYMA VYNEKDAAKK TEL-------
OSCRY3      ---FYGKH-- -HLDMARVDF IC---DSFYD I---FNDYMR MYHDQKGRVM FELMSQMREW
FBDH        ----IPADP- VENKAAL-D- WF---QKGDQ VN--FQVIT- -YGHKGV--- ----------
PPGST1      NK-LIPDCP- -AEQ-AMMY- -----QRMFE GL-----TL- -AQK--MA-- DVI--YYSW-
IOGST1      E--LSGKT-- -TEERATNTR WLS----FFN SD-------- ------FVQA AGGYFM----
SacCURE2    NPLLWSDD-- -LADQSQINA WL-----FFQ TSGH------ ------APMI GQ--------
ZMGST3      D--LLPAT-- --ASAAKLEV WLEVESHHFH PNAS---PLV FQLLVR-PLL GGA-------
ZMGST1      ---LLRE-GN -LEEAAMVDV WIEVEANQYT AALN---PIL FQVLIS-PML GGT-------
TAGST1      D--LLGENSG -IEELAMVDV WTEVEAQQYY PAIS---PVV FECIII-PFI IPGGGAA---
TAGST2      D--LLGENSG -IEELAMVDV WTEVEAQQYY PAIS---PVV FECIII-PFI IPGGGAA---
DCGST1      P--LLPQD-- -LQKRALNYQ AANIV----T SNIQ---PLQ NL-----AVL NY--------
ATGST2      S--LLPSPSD -PKAYGLFEQ AASVEYSSFD PPAS---SLA YERVFA-GMR GLKT------
ATGST1      N--LLQTDSK NISQYAIMAI GMQVEDHQFD PVAS---KLA FEQIFK-SIY GLT-------
ATGST3      N--LLST-GK -D--MAIIAM GIEIESHEFD PVGS---KLV WEQVLK-PLY GMT-------
NTGST1      Q--LILQD-- -PKKMPSMSV WMEVEGQKFE PPAT---KLT WELGIK-PII GMTT------
SCGST1      T--SLIHKE- -KHEMAAQLV WEEVEAHQFD PVAS---KLA WELVFK-GIF GMQ-------
HSGSTMIC    ---------- -----ARI-- ----YH---- -------TI- AYLT------ ----------
RNGSTMIC    ---------- -----ARI-- ----YH---- -------TI- AYLT------ ----------
DMelGST1    DS-LYPKC-- -PKKRAVIN- ----QRLYFD M-----GTL- -YQS--FAN- ---YYY----
DMelGST3    DS-LYPSD-- -PQKKAVVN- ----QRLYFD M-----GTL- -FQS--FVE- ---AIY----
ECSSPB      -P-LMPVY-- -P--VARGE- ----SRLYMH RIEKDWYTL- -------MNT ----------
MDGST1      DS-LFPKC-- -PKKRAVIN- ----QRLYFD M-----GTL- -YKS--FAD- ---YYY----
RNGST5      DH-WYPQD-- -LQARARVD- ----EYLAWQ HT-----TLR RS-CLRTLWH KVM--FPVFL
RNGST12     DH-WYPAD-- -LQARAQVH- ----EYLGWH AD-----NIR GTFGV-LLWT KVL--GPL-I
LCGST1      DS-LFPKC-- -PKKRAVIN- ----QRLYFD M-----GTL- -YKS--FAD- ---YYY----
MBDMD       N--WFGRG-- -TQERAQIN- ----QFLQWY ---------- AYTLR-L--G GGAFHW----
```

```
MMGST7      ---------- ---------- ---------- ---------- ---------- ----------
RNGST7      ---------- ---------- ---------- ---------- ---------- ----------
HSGSTP1     ---------- ---------- ---------- ---------- ---------- ----------
BTGSTP1     ---------- ---------- ---------- ---------- ---------- ----------
SSGSTP1     ---------- ---------- ---------- ---------- ---------- ----------
CEGSTP1     ---------- ---------- ---------- ---------- ---------- ----------
MAGSTM1     ---------- ---------- ---------- ---------- ---------- ----------
HSGSTM4     ---------- ---------- ---------- ---------- ---------- ----------
HSGSTM5     ---------- ---------- ---------- ---------- ---------- ----------
HSGSTM1b    ---------- ---------- ---------- ---------- ---------- ----------
HSGSTM2     ---------- ---------- ---------- ---------- ---------- ----------
HSGSTM1a    ---------- ---------- ---------- ---------- ---------- ----------
HSGSTM3     ---------- ---------- ---------- ---------- ---------- ----------
RNGST4      ---------- ---------- ---------- ---------- ---------- ----------
RNGST3      ---------- ---------- ---------- ---------- ---------- ----------
CLGST1      ---------- ---------- ---------- ---------- ---------- ----------
GGGSTM2     ---------- ---------- ---------- ---------- ---------- ----------
MMGST3      ---------- ---------- ---------- ---------- ---------- ----------
MMGST4      ---------- ---------- ---------- ---------- ---------- ----------
SJGST1      ---------- ---------- ---------- ---------- ---------- ----------
SMGST1      ---------- ---------- ---------- ---------- ---------- ----------
FHGST1      ---------- ---------- ---------- ---------- ---------- ----------
OCGSTA1     ---------- ---------- ---------- ---------- ---------- ----------
OCGSTA2     ---------- ---------- ---------- ---------- ---------- ----------
HSGSTA1     ---------- ---------- ---------- ---------- ---------- ----------
HSGSTA2     ---------- ---------- ---------- ---------- ---------- ----------
MMGST2      ---------- ---------- ---------- ---------- ---------- ----------
MMGST1      ---------- ---------- ---------- ---------- ---------- ----------
RNGST2      ---------- ---------- ---------- ---------- ---------- ----------
MMGST8      ---------- ---------- ---------- ---------- ---------- ----------
RRGST8      ---------- ---------- ---------- ---------- ---------- ----------
RNGST1a     ---------- ---------- ---------- ---------- ---------- ----------
RNGST1b     ---------- ---------- ---------- ---------- ---------- ----------
GGGSTA3     ---------- ---------- ---------- ---------- ---------- ----------
SHGST1      ---------- ---------- ---------- ---------- ---------- ----------
SBGST1      ---------- ---------- ---------- ---------- ---------- ----------
SMGST2      ---------- ---------- ---------- ---------- ---------- ----------
ASGST1      ---------- ---------- ---------- ---------- ---------- ----------
DMelGST2    ---------- ---------- ---------- ---------- ---------- ----------
MDGST2      ---------- ---------- ---------- ---------- ---------- ----------
OVCRY1      ---------- ---------- QRSRDMSSSS ---------- ---------- ----------
OVCRY2      ---------- ---------- QRSYDNGSSS ---------- ---------- ----------
ODCRY1      ---------- ---------- QRPYDNGNSS ---------- ---------- ----------
ODCRY2      ---------- ---------- QRSGDRNGSS ---------- ---------- ----------
ODCRY3      ---------- ---------- QRSRDMNSSS ---------- ---------- ----------
ODCRY4      ---------- ---------- SRSKDMNSSS ---------- ---------- ----------
OSCRY1      ---------- ---FMQGSGT DMSPDMDPTQ MT-------- ---------- ----------
OSCRY2      ---------- ---------- ---------- ---------- ---------- ----------
OSCRY3      YAARNENSGY EECYMQPSMA PSAQMSQEVD NSDTLADCSE MRSQDSMVEP PSQKLSPELE
FBDH        ---------- ---------- ---------- ---------- ---------- ----------
PPGST1      ---------- ---------- ---------- ---------- ---------- ----------
IOGST1      ---------- ---------- ---------- ---------- ---------- ----------
SacCURE2    ---------- ---------- ---------- ---------- ---------- ----------
ZMGST3      ---------- ---------- ---------- ---------- ---------- ----------
ZMGST1      ---------- ---------- ---------- ---------- ---------- ----------
TAGST1      ---------- ---------- ---------- ---------- ---------- ----------
TAGST2      ---------- ---------- ---------- ---------- ---------- ----------
DCGST1      ---------- ---------- ---------- ---------- ---------- ----------
ATGST2      ---------- ---------- ---------- ---------- ---------- ----------
ATGST1      ---------- ---------- ---------- ---------- ---------- ----------
ATGST3      ---------- ---------- ---------- ---------- ---------- ----------
NTGST1      ---------- ---------- ---------- ---------- ---------- ----------
SCGST1      ---------- ---------- ---------- ---------- ---------- ----------
HSGSTMIC    ---------- ---------- ---------- ---------- ---------- ----------
RNGSTMIC    ---------- ---------- ---------- ---------- ---------- ----------
DMelGST1    ---------- ---------- ---------- ---------- ---------- ----------
DMelGST3    ---------- ---------- ---------- ---------- ---------- ----------
ECSSPB      ---------- ---------- ---------- ---------- ---------- ----------
MDGST1      ---------- ---------- ---------- ---------- ---------- ----------
RNGST5      G--------- ---------- ---------- ---------- ---------- ----------
RNGST12     G--------- ---------- ---------- ---------- ---------- ----------
LCGST1      ---------- ---------- ---------- ---------- ---------- ----------
MBDMD       ---------- ---------- ---------- ---------- ---------- ----------
```

```
MMGST7      ---------- ---------- ---------- ---YENG-KN DYVKA----L P--GHLKP-F
RNGST7      ---------- ---------- ---------- ---YENG-KD DYVKA----L P--GHLKP-F
HSGSTP1     ---------- ---------- ---------- ---YEAG-KD DYVKA----L P--GQLKP-F
BTGSTP1     ---------- ---------- ---------- ---YEAG-KE DYVKA----L P--QHLKP-F
SSGSTP1     ---------- ---------- ---------- ---YEAG-KE KYVKE----L P--EHLKP-F
CEGSTP1     ---------- ---------- ---------- ---YEDG-KA PYIKDV---L P--GELAR--
MAGSTM1     ---------- ---------- ---------- --DFEKR-KP EYLEG----L P--EKMKL-Y
HSGSTM4     ---------- ---------- ---------- --DFEKL-KP EYLEE----L P--TMMQH-F
HSGSTM5     ---------- ---------- ---------- --DFEKL-KP KYLEE----L P--EKLKL-Y
HSGSTM1b    ---------- ---------- ---------- --EFEKL-KP KYLEE----L P--EKLKL-Y
HSGSTM2     ---------- ---------- ---------- --DFEKL-KP EYLQA----L P--EMLKL-Y
HSGSTM1a    ---------- ---------- ---------- --EFEKL-KP KYLEE----L P--EKLKL-Y
HSGSTM3     ---------- ---------- ---------- --DHEKL-KP QYLEE----L P--GQLKQ-F
RNGST4      ---------- ---------- ---------- --DFEKL-KP GYLEQ----L P--GMMRL-Y
RNGST3      ---------- ---------- ---------- --DFEKQ-KP EFLKT----I P--EKMKL-Y
CLGST1      ---------- ---------- ---------- --DFEKQ-KP EFLKT----I P--EKMKM-Y
GGGSTM2     ---------- ---------- ---------- --DFEKL-KP AYLEL----L P--GKLRQ-L
MMGST3      ---------- ---------- ---------- --DFEKQ-KP EFLKT----I P--EKMKL-Y
MMGST4      ---------- ---------- ---------- --DFEKK-KP EYLEG----L P--EKMKL-Y
SJGST1      ---------- ---------- ---------- --DFETL-KV DFLSK----L P--EMLKM-F
SMGST1      ---------- ---------- ---------- --EYETL-KV DFLNK----L P--GRLKM-F
FHGST1      ---------- ---------- ---------- --KFEEV-KG DYLKE----L P--TTLKM-W
OCGSTA1     ---------- ---------- ---------- PEQKDAK-VD FIKEK----I R--TRYFPAF
OCGSTA2     ---------- ---------- ---------- PEEQEAK-LA QIKDK----A K--NRYFPAF
HSGSTA1     ---------- ---------- ---------- PEEKDAK-LA LIKEK----I K--NRYFPAF
HSGSTA2     ---------- ---------- ---------- PEEQDAK-LA LIQEK----T K--NRYFPAF
MMGST2      ---------- ---------- ---------- PEEKEAS-LA KIKEQ----T R--NRYFPAF
MMGST1      ---------- ---------- ---------- PDQREAK-TA LAKDR----T K--NRYLPAF
RNGST2      ---------- ---------- ---------- PGEKEAS-LA KIKDK----A R--NRYFPAF
MMGST8      ---------- ---------- ---------- PKEKEES-YD LILSR----A K--TRYFPVF
RRGST8      ---------- ---------- ---------- PQEKEES-LA LAVKR----A K--NRYFPVF
RNGST1a     ---------- ---------- ---------- PDQREAK-TA LAKDR----T K--NRYLPAF
RNGST1b     ---------- ---------- ---------- PDQKEAK-TA LAKDR----T K--NRYLPAF
GGGSTA3     ---------- ---------- ---------- AEDKVKQ-CA FVVEK----A T--SRYFPAY
SHGST1      ---------- ---------- ---------- EEEKQKIIKE ILNGK----V P--VLLDIIC
SBGST1      ---------- ---------- ---------- EEEKQKITKE ILNGK----V P--VLLDIIC
SMGST2      ---------- ---------- ---------- QEEKEKITKE ILNGK----V P--VLLNMIC
ASGST1      ---------- ---------- --------FE EGDKEK--VL KEVAVP---A R--DKHLP-L
DMelGST2    ---------- ---------- ---------E DEIKEKKLVT -LNAE---VI P----FY---
MDGST2      ---------- ---------- ---------E DEIKEKKLVT -LNNE---VI P----FY---
OVCRY1      ---------- ---------- ---------- ----EKRMR- -FQE-----T C--RRILP-F
OVCRY2      ---------- ---------- ---------- ----ERRMR- -FQE-----T C--RRILP-F
ODCRY1      ---------- ---------- ---------- ----ERRSR- -FQE-----T C--RRILP-F
ODCRY2      ---------- ---------- ---------- ----EKRTR- -YQE-----T L--RRILP-F
ODCRY3      ---------- ---------- ---------- ----ESRMR- -FQE-----T C--RRILP-F
ODCRY4      ---------- ---------- ---------- ----EKRMR- -YQE-----T C--RRIFP-Y
OSCRY1      ---------- ---------- ------SYI- ------QNR- -YLD-----T C--RRILS-F
OSCRY2      ---------- ---------- ---------- ------QKR- -FQN-----T C--LRVLP-Y
OSCRY3      SQSSLCSERP QCGPPDPMMG SDFERLSFNE GRMLEMRRR- -YDE-----T C--RRVLP-F
FBDH        ---------P RGDELLIARR ERAKEYAEKY PELRSIYQAA HDRIVEHGNC A--YDADTVA
PPGST1      ---------- ---------- ---------- -KVPEAERHD SAVKRNKENL S--TELKL-W
IOGST1      ---------- ---------- ---------- -------GPN DEIKQ-QS-- LQTMLSL-L-
SacCURE2    -ALHFRYFHS QKIASAVERY TDEVRRVYGV VEMALAERRE ALVMELDTEN AAAYSAGTTP
ZMGST3      ---------- ---------- ---------- ---PD--AAV -VEKHAE--Q L--AKVLDVY
ZMGST1      ---------- ---------- ---------- ---TDQ--KV -VDENLE--K L--KKVLEVY
TAGST1      ---------- ---------- ---------- ---PNQ--TV -VDESLE--R L--RGVLGIY
TAGST2      ---------- ---------- ---------- ---PNQ--TV -VDESLE--R L--RGVLGIY
DCGST1      ---------- ---------- ---------- ---IEEKLG- -SDEKLSWAK HHIKKGFSAL
ATGST2      ---------- ---------- ---------- ---NEEL-AK ---KYVDT-L N--AK-MDGY
ATGST1      ---------- ---------- ---------- ---TDE--AV -VAE-EEA-K L--AKVLDVY
ATGST3      ---------- ---------- ---------- ---TDK--TV -VEE-EEA-K L--AKVLDVY
NTGST1      ---------- ---------- ---------- ---DDA--AV -K-ES-EA-Q L--SKVLDIY
SCGST1      ---------- ---------- ---------- ---TDT--TV -VEE-NEA-K L--AKVLDVY
HSGSTMIC    ---------- ---------- ---------- --------P- LPQPNRA--- ------LS-F
RNGSTMIC    ---------- ---------- ---------- --------P- LPQPNRG--- ------LA-F
DMelGST1    ---------- ---------- ---------- ---PQVFAKA PADPEAFK-K I--E---AAF
DMelGST3    ---------- ---------- ---------- ---PQIRNNH PADPEAMQ-K V--D---SAF
ECSSPB      ---------- ---------- ---------- -----IINGS ASEADAAR-K Q--LRE----
MDGST1      ---------- ---------- ---------- ---PQIFAKA PADPELFK-K I--E---TAF
RNGST5      ---------- ---------- ---------E QIRPEMLAAT LAD-LD---V N--VQVL---
RNGST12     ---------- ---------- ---------V QV-PEEKVER NRNSMV---L A--LQRL---
LCGST1      ---------- ---------- ---------- ---PQIFAKA PADPELYK-K M--E---AAF
MBDMD       ---------- ---------- -----NIFGC LIYGEKPYSP KFTAEQNK-G R--TLLYEAM
```

```
MMGST7      ---ET---LL SQNQ---GGK AFIV-GDQIS FADYNLLDLL LIHQ-VLAPG CLDN--FPLL
RNGST7      ---ET---LL SQNQ---GGK AFIV-GNQIS FADYNLLDLL LVHQ-VLAPG CLDN--FPLL
HSGSTP1     ---ET---LL SQNQ---GGK TFIV-GDQIS FADYNLLDLL LIHE-VLAPG CLDA--FPLL
BTGSTP1     ---ET---LL SQNK---GGQ AFIV-GDQIS FADYNLLDLL RIHQ-VLAPS CLDS--FPLL
SSGSTP1     ---ET---LL SQNQ---GGQ AFVV-GSQIS FADYNLLDLL RIHQ-VLNPS CLDA--FPLL
CEGSTP1     --LEKL--FH TY-K---NGE HYVI-GDKES YADYVLFEEL DIHL-ILTPN ALDG--VPAL
MAGSTM1     --SE----FL G--K-----R SWFA-GDKIT YVDFLIYDVL DQHR-IFAPK CLDA--FPNL
HSGSTM4     --SQ----FL G--K-----R PWFV-GDKIT FVDFLAYDVL DLHR-IFEPN CLDA--FPNL
HSGSTM5     --SE----FL G--K-----R PWFA-GDKIT FVDFLAYDVL DMKR-IFEPK CLDA--FLNL
HSGSTM1b    --SE----FL G--K-----R PWFA-GNKIT FVDFLVYDVL DLHR-IFEPN CLDA--FPNL
HSGSTM2     --SQ----FL G--K-----Q PWFL-GDKIT FVDFIAYDVL ERNQ-VFEPS CLDA--FPNL
HSGSTM1a    --SE----FL G--K-----R PWFA-GNKIT FVDFLVYDVL DLHR-IFEPK CLDA--FPNL
HSGSTM3     --SM----FL W--K-----F SWFA-GEKLT FVDFLTYDIL DQNR-IFDPK CLDE--FPNL
RNGST4      --SE----FL G--K-----R PWFA-GDKIT FVDFIAYDVL ERNQ-VFEAT CLDA--FPNL
RNGST3      --SE----FL G--K-----R PWFA-GDKVT YVDFLAYDIL DQYH-IFEPK CLDA--FPNL
CLGST1      --SE----FL G--K-----R PWFA-GDKVT LCGFLAYDVL DQYQ-MFEPK CLDP--FPNL
GGGSTM2     --SR----FL G--S-----R SWFV-GDKLT FVDFLAYDVL DQQR-MFVPD CPEL--QGNL
MMGST3      --SE----FL G--K-----R PWFA-GDKVT YVDFLAYDIL DQYR-MFEPK CLDA--FPNL
MMGST4      --SE----FL G--K-----Q PWFA-GNKVT YVDFLVYDVL DQHR-IFEPK CLDA--FPNL
SJGST1      --ED----RL C--H-----K TYLN-GDHVT HPDFMLYDAL DVVL-YMDPM CLDA--FPKL
SMGST1      --ED----RL S--N-----K TYLN-GNCVT HPDFMLYDAL DVVL-YMDSQ CLNE--FPKL
FHGST1      --SN----FL G--D-----R HYLT-GSPVS HVDFMVYEAL DCIR-YLAPQ CLED--FPKL
OCGSTA1     ---EK---VL KS-H----GQ DYLV-GNRLS KADILLVELL YNVE-ELDPS AIAS--FPLL
OCGSTA2     ---EK---VL KS-H----GQ DYLV-GNKLS KADILLVELL YNVE-ELNPG ATAS--FPLL
HSGSTA1     ---EK---VL KS-H----GQ DYLV-GNKLS RADIHLVELL YYVE-ELDSS LISS--FPLL
HSGSTA2     ---EK---VL KS-H----GQ DYLV-GNKLS RADIHLVELL YYVE-ELDSS LISS--FPLL
MMGST2      ---EK---VL KS-H----GQ DYLV-GNRLS RADIALVELL YHVE-ELDPG VVDN--FPLL
MMGST1      ---EK---VL KS-H----GQ DYLV-GNRLT RVDVHLLELL LYVE-ELDAS LLTP--FPLL
RNGST2      ---EK---VL KS-H----GQ DYLV-GNRLS RADVYLVQVL YHVE-ELDPS ALAN--FPLL
MMGST8      ---EK---IL KD-H----GE AFLV-GNQLS WADIQLLEAI LMVE-ELSAP VLSD--FPLL
RRGST8      ---EK---IL KD-H----GE AFLV-GNQLS WADIQLLEAI LMVE-EVSAP VLSD--FPLL
RNGST1a     ---EK---VL KS-H----GQ DYLV-GNRLT RVDIHLLELL LYVE-EFDAS LLTS--FPLL
RNGST1b     ---EK---VL KS-H----GQ DYLV-GNKLT RVDIHLLELL LYVE-EFDAS LLTS--FPLL
GGGSTA3     ---EK---VL KD-H----GQ DFLV-GNRLS WADIHLLEAI LMVE-EKKSD ALSG--FPLL
SHGST1      ---ES----L KA-----STG KLAV-GDKVT LADLVLIAVI DHVT-DLDKE FLTG-KYPEI
SBGST1      ---ES----L KA-----STG KLAV-GDKVT LADLVLIAVI DHVT-DLDKE FLTG-KYPEI
SMGST2      ---ES----L KG-----STG KLAV-GDKVT LADLVLIAVI DHVT-DLDKG FLTG-KYPEI
ASGST1      --LEK---FL A--K---SGS EYMV-GKSVT WADLVITDSL ASWE-SLIPD FLSG-H-LQL
DMelGST2    --LEKLEQTV KDND-----G HLAL-G-KLT WADVYFAGIT DYMNYMVKRD LLEP--YPAV
MDGST2      --LEKLEQTV KDND-----G HLAL--NKLT WADVYFAGIL DYMNYMVKRD ILEQ--YPAL
OVCRY1      --MER---TL EMYN---GGS QYFM-GDQMT MADMMCYCAL ENPL-MEEPS MLSS--YPKL
OVCRY2      --MER---TL EMRN---GGN QFFM-GDQMT MADLMCYCAL ENPL-TDDTS MLSS--YPKL
ODCRY1      --MER---TL EMRN---GGN QFSM-GDQMT MADLMCYCAL ENPL-TDEGS LLSS--YPKL
ODCRY2      --MER---TL EMYK---SGG QFFM-GDQMT MADMMCYCAL ENPI-MEESS LLNS--YPKL
ODCRY3      --MER---TL DMHS---GGS KFFM-GDQMT MADMMCYCAL ENPL-MEESS MLSS--YPKL
ODCRY4      --LEK---TL EMRN---GGN QFFM-GDQIT MADMMCFCAL ENPL-MEDQN ILRS--YPKL
OSCRY1      --LER---TL EMRN---GGK EFFM-GDQMM LCDMMCYCCL ENPM-LEDQT TFNN--FPKL
OSCRY2      --MEK---TL EANK---GGA GWFI-GDQIL LCDMMTHAAL ENPI-QENAN LLKE--YPKL
OSCRY3      --LEG---TL KQRY---GGD RYFM-GEYMT MCDLMCYCAL ENPL-LDNAY LLHP--YPKL
FBDH        QAEVDLQKRL DELDVHLADK PFIA-GSNYS IADIMWTVLL ARIE-MLNMT AWISER-PNL
PPGST1      --EE----YL Q--K---TSG SFVA-GKSFS LADVSVFPGV AYLF-RFGLT --EE-RYPQL
IOGST1      ---E----YI D--K-HLSQS KYFT-NNTIL TADIFAFQIF AMAK-QFGVD -F-T-HYPNV
SacCURE2    MS-QS--RFF D--Y-PV--- -WLV-GDKLT IADLAFVPWN NVVD-RIGIN IKIE--FPEV
ZMGST3      ---EA---HL A--R----NK -YLA-GDEFT LADANHALLP ALTSARPPRP GCVAAR-PHV
ZMGST1      ---EA---RL T--K----CK -YLA-GDFLS LADLNHVSVT LCLFATPYAS VLDA--YPHV
TAGST1      ---EA---RL E--K----SR -YLA-GDSIT FADLNHIPFT FYFMTTPYAK VFDD--YPKV
TAGST2      ---EA---RL E--K----SR -YLA-GDSIS FADLNHIPFT FYFMTTPYAK VFDE--YPKV
DCGST1      ---EK---LL KG-H---AGK -YAT-GDEVG LADLFLAPQI IASITGFGMD M-AE--FPLL
ATGST2      ---ER---IL S--K----QK -YLA-GNDFT LADLFHLPYG -AMVAQLEPT VLDS-K-PHV
ATGST1      ---EA---RL K--E----FK -YLA-GETFT LTDLHHIPAI QYLLGTPTKK LFTE-R-PRV
ATGST3      ---EH---RL G--E----SK -YLA-SDHFT LVDLHTIPVI QYLLGTPTKK LFDE-R-PHV
NTGST1      ---ET---QL A--E----SK -YLG-GDSFT LVDLHHIPNI YYLMSSKVKE VFDS-R-PRV
SCGST1      ---EA---RL T--E----SE -YLGANDSFT LVDLHHLPLL GYLMGTQVKK LFEE-R-AHV
HSGSTMIC    ---------- ---------- -FVGYG--VT LSMAY----- ---------- -----R--LL
RNGSTMIC    ---------- ---------- -FVGYG--VT LSMAY----- ---------- -----R--LL
DMelGST1    ---EFLNTFL EG-Q----D- -YAA-GDSLT VADIALVATV STFE-VAKFE IS---KYANV
DMelGST3    ---GHLDTFL ED-Q----E- -YVA-GDCLT IADIALLASV STFE-VVDFD IA---QYPNV
ECSSPB      ---ELL--AI APVF---GQK PYFL-SDEFS LVDCYL-APL L-WR-LPQLG IE----FSGP
MDGST1      ---DFLNTFL KG-H----E- -YAA-GDSLT VADLALLASV STFE-VASFD -F-S-KYPNV
RNGST5      ---EDQ--FL -QDK----D- -FLV-GPHIS LADVVAITEL MHPV-GGGCP VFEG-R-PRL
RNGST12     ---EDK--FL ---R----DR AFIA-GQQVT LADLMSLEEL IQPV-ALGCN LFEG-R-PQL
LCGST1      ---DFLNTFL EG-H----Q- -YVA-GDSLT VADLALLASV STFE-VAGFD -F-S-KYANV
MBDMD       GTLE--NYWL ---R----DR EYVC-GDEVS YADLAAFHEF VSHEAGKIIP DRVWQGFPKI
```

```
MMGST7      SAYVARLSAR -PKIKA---F LSSPEHVNRP INGNGKQ
RNGST7      SAYVARLSAR -PKIKA---F LSSPDHLNRP INGNGKQ
HSGSTP1     SAYVGRLSAR -PKLKA---F LASPEYVNLP INGNGKQ
BTGSTP1     SAYVARLNSR -PKLKA---F LASPEHMNRP INGNGKQ
SSGSTP1     SAYVARLSAR -PKIKA---F LASPEHVNRP ING----
CEGSTP1     KKFHERFAER -PNIKA---Y LNKRAAINPP VNGNGKQ
MAGSTM1     KDFLARFEGL -KKISD---Y MKSSRFSCKQ IFAKMAV
HSGSTM4     KDFISRFEGL -EKISA---Y MKSSRFLPKP LYTRVAV
HSGSTM5     KDFISRFEGL -KKISA---Y MKSSQFLRGL LFGKSAT
HSGSTM1b    KDFISRFEGL -EKISA---Y MKSSRFLPRP VFSKMAV
HSGSTM2     KDFISRFEGL -EKISA---Y MKSSRFLPRP VFTKMAV
HSGSTM1a    KDFISRFEGL -EKISA---Y MKSSRFLPRP VFSKMAV
HSGSTM3     KAFMCRFEAL -EKIAA---Y LQSDQFCKMP INNKMAQ
RNGST4      KDFIARFEGL -KKISD---Y MKSSRFLPRP LFTKMAI
RNGST3      KDFLARFEGL -KKISA---Y MKSSRYLSTP IFSKLAQ
CLGST1      KDFLARFEGL -KKISA---Y MKTSRFLRRP IFSKMAQ
GGGSTM2     SQFLQRFEAL -EKISA---Y MRSGRFMKAP IFWYTAL
MMGST3      RDFLARFEGL -KKISA---Y MKSSRYIATP IFSKMAH
MMGST4      KDFMGRFEGL -KKISD---Y MKSSRFLSKP IFAKMAF
SJGST1      VCFKKRIEAI -PQIDK---Y LKSSKYIAWP LQGWQAT
SMGST1      VSFKKCIEDL -PQIKN---Y LNSSRYIKWP LQGWDAT
FHGST1      KEFKSRIEDL -PKIKA---Y MESEKFIKWP LNSWIAS
OCGSTA1     KALKTRISSL -PTVKK---F LQPGSQRKPP MDEKNLE
OCGSTA2     QALKTRISNL -PTVKK---F LQPGSQRNPP DDEKCRE
HSGSTA1     KALKTRISNL -PTVKK---F LQPGSPRKPP MDEKSLE
HSGSTA2     KALKTRISNL -PTVKK---F LQPGSPRKPP MDEKSLE
MMGST2      KALRSRVSNL -PTVKK---F LQPGSQRKPF DDAKCVE
MMGST1      KAFKSRISSL -PNVKK---F LQPGSQRKPP LDAKQIE
RNGST2      KALRTRVSNL -PTVKK---F LQPGSQRKPL EDEKCVE
MMGST8      QAFKTRISNI -PTIKK---F LQPGSQRKPP PDGPYVE
RRGST8      QAFKTRISNI -PTIKK---F LQPGSQRKPP PDGHYVD
RNGST1a     KAFKSRISSL -PNVKK---F LQPGSQRKPA MDAKQIE
RNGST1b     KAFKSRISSL -PNVKK---F LQPGSQRKLP MDAKQIE
GGGSTA3     QAFKKRISSI -PTIKK---F LAPGSKRKPI SDDKYVE
SHGST1      --HKHRENLL -ASSPRLAKY LSD-----RA ATPF---
SBGST1      --HKHRENLL -ASSPRLAKY LSD-----RA ATPF---
SMGST2      --HKHRENLL -ASSPRLAKY LSN-----RP ATPF---
ASGST1      KKYIEHVREL -PNIKK---W IAE-----RP KTPY---
DMelGST2    RGVVDAVNAL -EPIKA---W IEK-----RP VTEV---
MDGST2      RGVVDSVNAL -EPIKA---W IEK-----RP QTEV---
OVCRY1      MALRNRVMNH -SKMSS---Y LQR-----RC RTDF---
OVCRY2      QALRNRVMSH -MKMSP---Y LKS-----RS STDF---
ODCRY1      QALRGRVMSH -MKMSP---Y LKS-----RT NTDF---
ODCRY2      QALRTRVMSH -LKMSP---Y LKK-----RS STEF---
ODCRY3      MSLRNRVMSH -PKMCN---Y LKK-----RC RTDF---
ODCRY4      QALRNRVINH -PKMSA---Y LQK-----RS RTEF---
OSCRY1      MSLWKRVASH -PKITP---Y LKK-----RN NTNW---
OSCRY2      AALRTRVAAH -PKIAA---Y IKK-----RN NTAF---
OSCRY3      RGLRDRVSRN -QRINS---Y FTL-----RN YTDF---
FBDH        LAYYQRMKAR ----RS---F ETA-----RV MPNWKGG
PPGST1      TAYYNSLKER -PSIKAS--W PPT-----WL ESP-QGQ
IOGST1      ERFTGEVSQH -P-IIKNM-- ---------- -------
SacCURE2    YKWTKHMMRR -PAVIKAL-- ---------- -RGE---
ZMGST3      KAWWEAIAAR -PAF------ -------QKT VAAIPLP
ZMGST1      KAWWSGLMER -PSV------ -------QK- VAALMKP
TAGST1      KAWWEMLMAR -PAV------ -------QR- VCKHM-P
TAGST2      KAWWEMLMAR -PAV------ -------QR- VCKHM-P
DCGST1      KSLNDAYLKY -QHFRM---- -----RCQR- ISPML--
ATGST2      KAWWAASLRV IPGRLL---- ------RN-- SSK----
ATGST1      NEWVAEITKR -PAS------ -------EK- VQ-----
ATGST3      SAWVADITSR -PSA------ -------QK- VL-----
NTGST1      SAWCADILAR -PAWVKGLE- -----KLQK- -------
SCGST1      SAWCKKILAR -PSWEKTLA- ------LQK- QA-----
HSGSTMIC    KSKL------ ---------Y L--------- -------
RNGSTMIC    RSRL------ ---------Y L--------- -------
DMelGST1    NRWYENAKKV -TPGWEEN-W AGC-----LE FKKYFE-
DMelGST3    ASWYENAKEV -TPGWEEN-W DGV-----QL IKKLVQE
ECSSPB      GAK--ELKGY MTRVFERDSF LAS-----LT EAER-EM
MDGST1      AKWYANLKTV -APGWEEN-W AGC-----LE FKKYFG-
RNGST5      AAWYRRVEAA -VGKDL---F LEA-----HE VILKV-R
RNGST12     TAWRERVEAF -LGAEL---C QEA-----HN PIMSVLG
LCGST1      AKWYANAKTV -APGFDEN-W EGC-----LE FKKFFN-
MBDMD       AAWFKKLSER -PHAKTVSEW QYT--NVGKI IRGELTA
```

APPENDIX II Alignment of GST amino acid sequences. **(a)** Alpha class **(b)** Mu class **(c)** Pi class.

(a) Alpha class

```
GGGSTA3     MAAKPVLYYF  NGRGKMESIR  WLLAAAGVEF  EEVFLETREQ  YEKLLQSGIL  MFQQVPMVEI
OCGSTA1     MARKPLLHYF  NGRGRMESIR  WLLAAAGEEF  DEKFMETAED  LDKLRNDGSL  MYQQVPMVEI
OCGSTA2     MAGKPKLHYF  NARGRMESIR  WLLTAAGVEF  EEKCMKTRED  LEKLRKDGVL  MFQQVPMVEI
MMGST2      MAGKPVLHYF  DGRGRMEPIR  WLLAAAGVEF  EEKFLKTRDD  LARLRSDGSL  MFQQVPMVEI
MMGST8      MAAKPKLYYF  NGRGRMESIR  WLLAAAGVEF  EEEFLETREQ  YEKMQKDGHL  LFGQVPLVEI
RRGST8      MEVKPKLYYF  QGRGRMESIR  WLLATAGVEF  EEEFLETREQ  YEKLQKDGCL  LFGQVPLVEI
HSGSTA1     MAEKPKLHYF  NARGRMESTR  WLLAAAGVEF  EEKFIKSAED  LDKLRNDGYL  MFQQVPMVEI
HSGSTA2     MAEKPKLHYS  NIRGRMESIR  WLLAAAGVEF  EEKFIKSAED  LDKLRNDGYL  MFQQVPMVEI
MMGST1      MAGKPVLHYF  NARGRMECIR  WLLAAAGVEF  EEKFIQSPED  LEKLKKDGNL  MFDQVPMVEI
RNGST1a     MSGKPVLHYF  NARGRMECIR  WLLAAAGVEF  EEKLIQSPED  LEKLKKDGNL  MFDQVPMVEI
RNGST1b     MSGKPVLHYF  NARGRMECIR  WLLAAAGVEF  DEKFIQSPED  LEKLKKDGNL  MFDQVPMVEI
RNGST2      MPGKPVLHYF  DGRGRMEPIR  WLLAAAGVEF  EEQFLKTRDD  LARLRNDGSL  MFQQVPMVEI

GGGSTA3     DGMKLVQTRA  ILNYIAGKYN  LYGKDLKERA  LIDMYVGGTD  DLMGFLLSFP  FLSAEDKVKQ
OCGSTA1     DGMKLVQTRA  ILNYVANKHN  LYGKDMKERA  LIDMYTEGVA  DLYELVLLLP  LCPPEQKDAK
OCGSTA2     DGMKLVQTRA  IFNYIADKHN  LYGKDIKERA  LIDMYTEGIV  DLNELILTRP  FLPPEEQEAK
MMGST2      DGMKLVQTKA  ILNYIASKYN  LYGKDMKERA  IIDMYTEGVA  DLEIMILYYP  HMPPEEKEAS
MMGST8      DGMMLTQTRA  ILSYLAAKYN  LYGKDLKERV  RIDMYADGTQ  DLMMMIAVAP  FKTPKEKEES
RRGST8      DGMLLTQTRA  ILSYLAAKYN  LYGKDLKERV  RIDMYADGTQ  DLMMMIIGAP  FKAPQEKEES
HSGSTA1     DGMKLVQTRA  ILNYIASKYN  LYGKDIKERA  LIDMYIEGIA  DLGEMILLLP  VCPPEEKDAK
HSGSTA2     DGMKLVQTRA  ILNYIASKYN  LYGKDIKEKA  LIDMYIEGIA  DLGEMILLLP  FTQPEEQDAK
MMGST1      DGMKLVQTRA  ILNYIATKYD  LYGKDMKERA  LIDMYTEGIL  DLTEMIGQLV  LCPPDQREAK
RNGST1a     DGMKLAQTRA  ILNYIATKYD  LYGKDMKERA  LIDMYSEGIL  DLTEMIIQLV  ICPPDQREAK
RNGST1b     DGMKLAQTRA  ILNYIATKYD  LYGKDMKERA  LIDMYTEGIL  DLTEMIMQLV  ICPPDQKEAK
RNGST2      DGMKLVQTRA  ILNYIATKYN  LYGKDMKERA  LIDMYAEGVA  DLDEIVLHYP  YIPPGEKEAS

GGGSTA3     CAFVVEKATS  RYFPAYEKVL  KDHGQDFLVG  NRLSWADIHL  LEAILMVEEK  KSDALSGFPL
OCGSTA1     VDFIKEKIRT  RYFPAFEKVL  KSHGQDYLVG  NRLSKADILL  VELLYNVEEL  DPSAIASFPL
OCGSTA2     LAQIKDKAKN  RYFPAFEKVL  KSHGQDYLVG  NKLSKADILL  VELLYNVEEL  NPGATASFPL
MMGST2      LAKIKEQTRN  RYFPAFEKVL  KSHGQDYLVG  NRLSRADIAL  VELLYHVEEL  DPGVVDNFPL
MMGST8      YDLILSRAKT  RYFPVFEKIL  KDHGEAFLVG  NQLSWADIQL  LEAILMVEEL  SAPVLSDFPL
RRGST8      LALAVKRAKN  RYFPVFEKIL  KDHGEAFLVG  NQLSWADIQL  LEAILMVEEV  SAPVLSDFPL
HSGSTA1     LALIKEKIKN  RYFPAFEKVL  KSHGQDYLVG  NKLSRADIHL  VELLYYVEEL  DSSLISSFPL
HSGSTA2     LALIQEKTKN  RYFPAFEKVL  KSHGQDYLVG  NKLSRADIHL  VELLYYVEEL  DSSLISSFPL
MMGST1      TALAKDRTKN  RYLPAFEKVL  KSHGQDYLVG  NRLTRVDVHL  LELLLYVEEL  DASLLTPFPL
RNGST1a     TALAKDRTKN  RYLPAFEKVL  KSHGQDYLVG  NRLTRVDIHL  LELLLYVEEF  DASLLTSFPL
RNGST1b     TALAKDRTKN  RYLPAFEKVL  KSHGQDYLVG  NKLTRVDIHL  LELLLYVEEF  DASLLTSFPL
RNGST2      LAKIKDKARN  RYFPAFEKVL  KSHGQDYLVG  NRLSRADVYL  VQVLYHVEEL  DPSALANFPL

GGGSTA3     LQAFKKRISS  IPTIKKFLAP  GSKRKPISDD  KYVETVRRVL  RMYYDVKPH*
OCGSTA1     LKALKTRISS  LPTVKKFLQP  GSQRKPPMDE  KNLEKAKKIF  ------KIP*
OCGSTA2     LQALKTRISN  LPTVKKFLQP  GSQRNPPDDE  KCREEAKIIF  --------H*
MMGST2      LKALRSRVSN  LPTVKKFLQP  GSQRKPFDDA  KCVESAKKIF  ------S--*
MMGST8      LQAFKTRISN  IPTIKKFLQP  GSQRKPPPDG  PYVEVVRIVL  ------K-F*
RRGST8      LQAFKTRISN  IPTIKKFLQP  GSQRKPPPDG  HYVDVVRTVL  ------K-F*
HSGSTA1     LKALKTRISN  LPTVKKFLQP  GSPRKPPMDE  KSLEEARKIF  ------R-F*
HSGSTA2     LKALKTRISN  LPTVKKFLQP  GSPRKPPMDE  KSLEESRKIF  ------R-F*
MMGST1      LKAFKSRISS  LPNVKKFLQP  GSQRKPPLDA  KQIEEARKVF  ------K-F*
RNGST1a     LKAFKSRISS  LPNVKKFLQP  GSQRKPAMDA  KQIEEARKVF  ------K-F*
RNGST1b     LKAFKSRISS  LPNVKKFLQP  GSQRKLPMDA  KQIEEARKIF  ------K-F*
RNGST2      LKALRTRVSN  LPTVKKFLQP  GSQRKPLEDE  KCVESAVKIF  ------S--*
```

(b) Mu class

```
FHGST1      MP----AKLG  YWKIRGLQQP  VRLLLEYLGE  EYEEHLY---  GR-D-DREKW  FGDKFNMGLD
SJGST1      MS----PILG  YWKIKGLVQP  TRLLLEYLEE  KYEEHLY---  ER-D-EGDKW  RNKKFELGLE
SMGST1      MA----PKFG  YWKVKGLVQP  TRLLLEHLEE  TYEERAY---  DR-N-EIDAW  SNDKFKLGLE
MAGSTM1     MP----VTLG  YWDIRGLAHA  IRLLLEYTDT  SYEEKKYTMG  DAPNFDRSQW  LNEKFKLGLD
HSGSTM2     MP----MTLG  YWNIRGLAHS  IRLLLEYTDS  SYEEKKYTMG  DAPDYDRSQW  LNEKFKLGLD
CLGST1      MP----MILG  YWNVRGLTNP  IRLLLEYTDS  SYEEKKYTMG  DAPDSDRSQW  LNEKFKLGLD
HSGSTM1b    MP----MILG  YWDIRGLAHA  IRLLLEYTDS  SYEEKKYTMG  DAPDYDRSQW  LNEKFKLGLD
HSGSTM3     MSCESSMVLG  YWDIRGLAHA  IRLLLEFTDT  SYEEKRYTCG  EAPDYDRSQW  LDVKFKLDLD
HSGSTM1a    MP----MILG  YWDIRGLAHA  IRLLLEYTDS  SYEEKKYTMG  DAPDYDRSQW  LNEKFKLGLD
MMGST3      MP----MILG  YWNVRGLTHP  IRMLLEYTDS  SYDEKRYTMG  DAPDFDRSQW  LNEKFKLGLD
RNGST3      MP----MILG  YWNVRGLTHP  IRLLLEYTDS  SYEEKRYAMG  DAPDYDRSQW  LNEKFKLGLD
RNGST4      MP----MTLG  YWDIRGLAHA  IRLLLEYTDS  SYEEKRYTMG  DAPDFDRSQW  LNEKFKLGLD
HSGSTM4     MS----MTLG  YWDIRGLAHA  IRLLLEYTDS  SYEEKKYTMG  DAPDYDRSQW  LNEKFKLGLD
HSGSTM5     MP----MTLG  YWDIRGLAHA  IRLLLEYTDS  SYVEKKYTMG  DAPDYDRSQW  LNEKFKLGLD
GGGSTM2     MV----VTLG  YWDIRGLAHA  IRLLLEYTET  PYQERRYKAG  PAPDFDPSDW  TNEKEKLGLD
MMGST4      MP----MTLG  YWDIRGLAHA  IRLLLEYTDT  SYEDKKYTMG  DAPDYDRSQW  LSEKFKLGLD

FHGST1      LPNLPYYIDD  KCKLTQSVAI  MRYIADKHGM  LGTTPEERAR  ISMIEGAAMD  LRMGFVRVCY
SJGST1      FPNLPYYIDG  DVKLTQSMAI  IRYIADKHNM  LGGCPKERAE  ISMLEGAVLD  IRYGVSRIAY
SMGST1      FPNLPYYIDG  DFKLTQSMAI  IRYIADKHNM  LGACPKERAE  ISMLEGAVLD  IRMGVLRIAY
MAGSTM1     FPNLPYLIDG  SHKITQSNAI  LRYIARKHDL  CGETEEERIQ  LDILENQAMD  TRMQLAMVCY
HSGSTM2     FPNLPYLIDG  THKITQSNAI  LRYIARKHNL  CGESEKEQIR  EDILENQFMD  SRMQLAKLCY
CLGST1      FPNLPYLIDG  SHKITQSNAI  LRYIARKHNL  CGETEEERIR  VDIVENQAMD  TRMQLIMLCY
HSGSTM1b    FPNLPYLIDG  AHKITQSNAI  LCYIARKHNL  CGETEEEKIR  VDILENQTMD  NHMQLGMICY
HSGSTM3     FPNLPYLLDG  KNKITQSNAI  LRYIARKHNM  CGETEEEKIR  VDIIENQVMD  FRTQLIRLCY
HSGSTM1a    FPNLPYLIDG  AHKITQSNAI  LCYIARKHNL  CGETEEEKIR  VDILENQTMD  NHMQLGMICY
MMGST3      FPNLPYLIDG  SHKITQSNAI  LRYLARKHHL  DGETEEERIR  ADIVENQVMD  TRMQLIMLCY
RNGST3      FPNLPYLIDG  SRKITQSNAI  MRYLARKHHL  CGETEEERIR  ADIVENQVMD  NRMQLIMLCY
```

```
RNGST4      FPNLPYLIDG  SHKITQSNAI  LRYLGRKHNL  CGETEEERIR  VDILENQLMD  NRMVLARLCY
HSGSTM4     FPNLPYLIDG  AHKITQSNAI  LCYIARKHNL  CGETEEEKIR  VDILENQAMD  VSNQLARVCY
HSGSTM5     FPNLPYLIDG  AHKITQSNAI  LRYIARKHNL  CGETEEEKIR  VDILENQVMD  NHMELVRLCY
GGGSTM2     FPNLPYLIDG  DVKLTQSNAI  LRYIARKHNM  CGETEVEKQR  VDVLENHLMD  LRMAFARLCY
MMGST4      FPNLPYLIDG  SHKITQSNAI  LRYLARKHNL  CGETEEERIR  VDILENQAMD  TRIQLAMVCY

FHGST1      NPKFEEVKGD  YLKELPTTLK  MWSNFLGDRH  YLTGSPVSHV  DFMVYEALDC  IRYLAPQCLE
SJGST1      SKDFETLKVD  FLSKLPEMLK  MFEDRLCHKT  YLNGDHVTHP  DFMLYDALDV  VLYMDPMCLD
SMGST1      NKEYETLKVD  FLNKLPGRLK  MFEDRLSNKT  YLNGNCVTHP  DFMLYDALDV  VLYMDSQCLN
MAGSTM1     SPDFEKRKPE  YLEGLPEKMK  LYSEFLGKRS  WFAGDKITYV  DFLIYDVLDQ  HRIFAPKCLD
HSGSTM2     DPDFEKLKPE  YLQALPEMLK  LYSQFLGKQP  WFLGDKITFV  DFIAYDVLER  NQVFEPSCLD
CLGST1      NPDFEKQKPE  FLKTIPEKMK  MYSEFLGKRP  WFAGDKVTLC  GFLAYDVLDQ  YQMFEPKCLD
HSGSTM1b    NPEFEKLKPK  YLEELPEKLK  LYSEFLGKRP  WFAGNKITFV  DFLVYDVLDL  HRIFEPNCLD
HSGSTM3     SSDHEKLKPQ  YLEELPGQLK  QFSMFLWKFS  WFAGEKLTFV  DFLTYDILDQ  NRIFDPKCLD
HSGSTM1a    NPEFEKLKPK  YLEELPEKLK  LYSEFLGKRP  WFAGNKITFV  DFLVYDVLDL  HRIFEPKCLD
MMGST3      NPDFEKQKPE  FLKTIPEKMK  LYSEFLGKRP  WFAGDKVTYV  DFLAYDILDQ  YRMFEPKCLD
RNGST3      NPDFEKQKPE  FLKTIPEKMK  LYSEFLGKRP  WFAGDKVTYV  DFLAYDILDQ  YHIFEPKCLD
RNGST4      NPDFEKLKPG  YLEQLPGMMR  LYSEFLGKRP  WFAGDKITFV  DFIAYDVLER  NQVFEATCLD
HSGSTM4     SPDFEKLKPE  YLEELPTMMQ  HFSQFLGKRP  WFVGDKITFV  DFLAYDVLDL  HRIFEPNCLD
HSGSTM5     DPDFEKLKPK  YLEELPEKLK  LYSEFLGKRP  WFAGDKITFV  DFLAYDVLDM  KRIFEPKCLD
GGGSTM2     SPDFEKLKPA  YLELLPGKLR  QLSRFLGSRS  WFVGDKLTFV  DFLAYDVLDQ  QRMFVPDCPE
MMGST4      SPDFEKKKPE  YLEGLPEKMK  LYSEFLGKQP  WFAGNKVTYV  DFLVYDVLDQ  HRIFEPKCLD

FHGST1      DFPKLKEFKS  RIEDLPKIKA  YMESEKFIKW  PLNSWIASFG  GGDAAPA---
SJGST1      AFPKLVCFKK  RIEAIPQIDK  YLKSSKYIAW  PLQGWQATFG  GGDHPPK---
SMGST1      EFPKLVSFKK  CIEDLPQIKN  YLNSSRYIKW  PLQGWDATFG  GGDTPPK---
MAGSTM1     AFPNLKDFLA  RFEGLKKISD  YMKSSRFSCK  QIFAKMAVW-  ----NSK---
HSGSTM2     AFPNLKDFIS  RFEGLEKISA  YMKSSRFLPR  PVFTKMAVW-  ----GNK---
CLGST1      PFPNLKDFLA  RFEGLKKISA  YMKTSRFLRR  PIFSKMAQW-  ----SNK---
HSGSTM1b    AFPNLKDFIS  RFEGLEKISA  YMKSSRFLPR  PVFSKMAVW-  ----GNK---
HSGSTM3     EFPNLKAFMC  RFEALEKIAA  YLQSDQFCKM  PINNKMAQW-  ----GNKPVC
HSGSTM1a    AFPNLKDFIS  RFEGLEKISA  YMKSSRFLPR  PVFSKMAVW-  ----GNK---
MMGST3      AFPNLRDFLA  RFEGLKKISA  YMKSSRYIAT  PIFSKMAHW-  ----SNK---
RNGST3      AFPNLKDFLA  RFEGLKKISA  YMKSSRYLST  PIFSKLAQW-  ----SNK---
RNGST4      AFPNLKDFIA  RFEGLKKISD  YMKSSRFLPR  PLFTKMAIW-  ----GSK---
HSGSTM4     AFPNLKDFIS  RFEGLEKISA  YMKSSRFLPK  PLYTRVAVW-  ----GNK---
HSGSTM5     AFLNLKDFIS  RFEGLKKISA  YMKSSQFLRG  LLFGKSATW-  ----NSK---
GGGSTM2     LQGNLSQFLQ  RFEALEKISA  YMRSGRFMKA  PIFWYTALW-  ---NNKKE--
MMGST4      AFPNLKDFMG  RFEGLKKISD  YMKSSRFLSK  PIFAKMAFW-  ----NPK---
```

(c) Pi class

```
CEGSTP1     MT-LKLTYFD   IHGLAEPIRL  LLADKQVAYE  DHRVTYEQW-   -ADIKPKMIF  GQVPCLLSGD
BTGSTP1     MPPYTIVYFP   VQGRCEAMRM  LLADQGQSWK  EEVVAMQSWL   QGPLKASCLY  GQLPKFQDGD
MMGST7      MPPYTIVYFP   VRGRCEAMRM  LLADQGQSWK  EEVVTIDTWM   QGLLKPTCLY  GQLPKFEDGD
RNGST7      MPPYTIVYFP   VRGRCEATRM  LLADQGQSWK  EEVVTIDVWL   QGSLKSTCLY  GQLPKFEDGD
HSGSTP1     MPPYTVVYFP   VRGRCAALRM  LLADQGQSWK  EEVVTVETWQ   EGSLKASCLY  GQLPKFQDGD

CEGSTP1     EEIVQSGAII  RHLARLNGLN  GSNETETTFI  DMFYEGLRDL  HTKYTTMIYR  NYEDGKAPYI
BTGSTP1     LTLYQSNAIL  RHLGRTLGLY  GKDQQEAALV  DMVNDGVEDL  RCKYVSLIYT  NYEAGKEDYV
MMGST7      LTLYQSNAIL  RHLGRSLGLY  GKNQREAAQM  DMVNDGVEDL  RGKYVTLIYT  NYENGKNDYV
RNGST7      LTLYQSNAIL  RHLGRSLGLY  GKDQKEAALV  DMVNDGVEDL  RCKYGTLIYT  NYENGKDDYV
HSGSTP1     LTLYQSNTIL  RHLGRTLGLY  GKDQQEAALV  DMVNDGVEDL  RCKYISLIYT  NYEAGKDDYV

CEGSTP1     KDVLPGELAR  LEKLFHTYKN  GEHYVIGDKE  SYADYVLFEE  LDIHLILTPN  ALDGVPALKK
BTGSTP1     KA-LPQHLKP  FETLLSQNKG  GQAFIVGDQI  SFADYNLLDL  LRIHQVLAPS  CLDSFPLLSA
MMGST7      KA-LPGHLKP  FETLLSQNQG  GKAFIVGDQI  SFADYNLLDL  LLIHQVLAPG  CLDNFPLLSA
RNGST7      KA-LPGHLKP  FETLLSQNQG  GKAFIVGNQI  SFADYNLLDL  LLVHQVLAPG  CLDNFPLLSA
HSGSTP1     KA-LPGQLKP  FETLLSQNQG  GKTFIVGDQI  SFADYNLLDL  LLIHEVLAPG  CLDAFPLLSA

CEGSTP1     FHERFAERPN   IKAYLNKRAA  INPPVNGNGK  Q
BTGSTP1     YVARLNSRPK   LKAFLASPEH  MNRPINGNGK  Q
MMGST7      YVARLSARPK   IKAFLSSPEH  VNRPINGNGK  Q
RNGST7      YVARLSARPK   IKAFLSSPDH  LNRPINGNGK  Q
HSGSTP1     YVGRLSARPK   LKAFLASPEY  VNLPINGNGK  Q
```

APPENDIX III Alignment of all available nucleotide sequences in the (a) Alpha, (b) Mu and (c) Pi classes.

(a) Alpha class nucleotide sequences.

```
GGGSTA3     ATGGCTGCAA  AACCTGTACT  CTACTACTTC  AATGGAAGAG  GCAAAATGGA
OCGSTA1     ATGGCAAGAA  AGCCCCTGCT  TCACTACTTC  AATGGAAGGG  GCAGAATGGA
OCGSTA2     ATGGCAGGGA  AGCCCAAGCT  TCACTACTTC  AATGCACGGG  GCAGAATGGA
MMGST2      ATGGCGGGGA  AGCCAGTCCT  TCATTACTTT  GATGGCAGGG  GAAGAATGGA
MMGST8      ATGGCAGCCA  AACCTAAGCT  CTACTACTTT  AATGGCAGGG  GACGGATGGA
RRGST8      ATGGAAGTCA  AACCCAAGCT  CTACTACTTT  CAAGGCAGGG  GAAGGATGGA
HSGSTA1     ATGGCAGAGA  AGCCCAAGCT  CCACTACTTC  AATGCACGGG  GCAGAATGGA
HSGSTA2     ATGGCAGAGA  AGCCCAAGCT  CCACTACTCC  AATATACGGG  GCAGAATGGA
MMGST1      ATGGCCGGGA  AGCCCGTGCT  TCACTACTTC  AATGCCCGGG  GCAGAATGGA
RNGST1a     ATGTCTGGGA  AGCCAGTGCT  TCACTACTTC  AATGCCCGGG  GCAGAATGGA
RNGST1b     ATGTCTGGGA  AGCCAGTGCT  TCACTACTTC  AATGCCCGGG  GCAGAATGGA
RNGST2      ATGCCGGGGA  AGCCAGTCCT  TCACTACTTC  GATGGCAGGG  GGAGAATGGA

GGGSTA3     GTCGATCCGC  TGGCTGCTGG  CTGCAGCTGG  GGTTGAGTTC  GAAGAGGTGT
OCGSTA1     GTCCATTCGG  TGGCTGTTGG  CCGCTGCTGG  AGAAGAGTTT  GATGAGAAAT
OCGSTA2     GTCTATCCGG  TGGCTCCTGA  CTGCAGCTGG  GGTAGAGTTT  GAAGAGAAAT
MMGST2      GCCTATCCGG  TGGCTCTTGG  CTGCAGCTGG  TGTGGAGTTT  GAAGAAAAAT
MMGST8      GTCGATCCGC  TGGCTGCTGG  CTGCGGCTGG  AGTGGAGTTT  GAGGAAGAAT
RRGST8      GTCGATCCGC  TGGCTGCTGG  CTACAGCTGG  AGTGGAGTTT  GAAGAAGAAT
HSGSTA1     GTCCACCCGG  TGGCTCCTGG  CTGCAGCTGG  AGTAGAGTTT  GAAGAGAAAT
HSGSTA2     GTCCATCCGG  TGGCTCCTGG  CTGCAGCTGG  AGTAGAGTTT  GAAGAGAAAT
MMGST1      GTGCATCAGG  TGGCTCCTGG  CTGCAGCAGG  GGTGGAGTTT  GAAGAGAAGT
RNGST1a     GTGCATCCGG  TGGCTCCTGG  CTGCAGCAGG  AGTGGAGTTT  GAAGAGAAGC
RNGST1b     GTGCATCCGG  TGGCTCCTCG  CTGCAGCAGG  AGTGGAGTTT  GATGAGAAGT
RNGST2      GCCCATCCGG  TGGCTCCTGG  CTGCAGCTGG  AGTAGAGTTT  GAAGAACAAT

GGGSTA3     TTTTGGAAAC  ACGAGAGCAG  TATGAGAAGC  TCCTGCAAAG  TGGAATCCTC
OCGSTA1     TTATGGAAAC  TGCAGAGGAT  TTGGACAAGT  TAAGAAATGA  TGGGAGTTTG
OCGSTA2     GTATGAAAAC  TCGAGAAGAC  CTGGAAAAGT  TAAGAAAAGA  TGGGGTATTG
MMGST2      TTCTGAAAAC  TCGGGATGAC  CTGGCAAGGT  TACGAAGTGA  TGGGAGTCTG
MMGST8      TTCTTGAGAC  AAGGGAACAG  TATGAGAAGA  TGCAAAAGGA  TGGACACCTG
RRGST8      TTCTTGAGAC  GAGAGAACAA  TATGAGAAGT  TGCAAAAGGA  TGGATGCCTG
HSGSTA1     TTATAAAATC  TGCAGAAGAT  TTGGACAAGT  TAAGAAATGA  TGGATATTTG
HSGSTA2     TTATAAAATC  TGCAGAAGAT  TTGGACAAGT  TAAGAAATGA  TGGATATTTG
MMGST1      TTATACAGAG  TCCGGAAGAT  TTGGAAAAGC  TAAAAAAAGA  TGGGAATTTG
RNGST1a     TTATACAGAG  TCCAGAAGAC  TTGGAAAAGC  TAAAGAAAGA  CGGGAATTTG
RNGST1b     TTATACAAAG  TCCAGAAGAC  TTGGAAAAGC  TAAAGAAAGA  CGGGAATTTG
RNGST2      TTCTGAAAAC  TCGGGATGAC  CTGGCCAGGC  TAAGGAATGA  TGGGAGTTTG

GGGSTA3     ATGTTCCAGC  AAGTGCCCAT  GGTGGAGATC  GACGGGATGA  AGTTGGTGCA
OCGSTA1     ATGTACCAGC  AAGTGCCCAT  GGTTGAGATT  GATGGGATGA  AGCTGGTGCA
OCGSTA2     ATGTTCCAGC  AAGTGCCCAT  GGTTGAGATT  GATGGGATGA  AGCTGGTGCA
MMGST2      ATGTTCCAGC  AAGTGCCCAT  GGTAGAGATC  GACGGGATGA  AACTGGTGCA
MMGST8      CTTTTCGGCC  AAGTACCCTT  GGTTGAAATC  GATGGGATGA  TGCTGACACA
RRGST8      CTTTTTGGCC  AAGTCCCATT  GGTGGAAATA  GACGGGATGC  TACTGACACA
HSGSTA1     ATGTTCCAGC  AAGTGCCAAT  GGTTGAGATT  GATGGGATGA  AGCTGGTGCA
HSGSTA2     ATGTTCCAGC  AAGTGCCAAT  GGTTGAGATT  GATGGGATGA  AGCTGGTGCA
MMGST1      ATGTTTGACC  AAGTGCCCAT  GGTGGAGATT  GATGGGATGA  AGCTGGTGCA
RNGST1a     ATGTTTGACC  AAGTGCCCAT  GGTGGAGATT  GACGGGATGA  AGCTGGCACA
RNGST1b     ATGTTTGACC  AAGTGCCCAT  GGTGGAGATT  GACGGGATGA  AGCTGGCACA
RNGST2      ATGTTCCAGC  AAGTGCCCAT  GGTGGAGATT  GATGGGATGA  AGCTGGTGCA
GGGSTA3     GACCAGAGCC  ATCCTCAACT  ACATAGCAGG  GAAATACAAT  CTCTATGGGA
OCGSTA1     GACCAGAGCC  ATTCTCAATT  ATGTTGCCAA  CAAACACAAC  CTGTATGGGA
OCGSTA2     GACCAGAGCC  ATTTTCAACT  ACATTGCAGA  CAAGCACAAC  CTGTATGGGA
MMGST2      GACCAAAGCC  ATTCTCAACT  ACATTGCCTC  CAAATACAAC  CTCTATGGGA
MMGST8      GACCAGGGCC  ATCCTCAGCT  ACCTCGCTGC  CAAGTACAAC  TTGTATGGGA
RRGST8      GACCAGAGCC  ATCCTCAGCT  ACCTGGCCGC  CAAGTACAAC  TTGTATGGGA
HSGSTA1     GACCAGAGCC  ATTCTCAACT  ACATTGCCAG  CAAATACAAC  CTCTATGGGA
HSGSTA2     GACCAGAGCC  ATTCTCAACT  ACATTGCCAG  CAAATACAAC  CTCTATGGGA
MMGST1      GACCAGAGCC  ATTCTCAACT  ACATCGCCAC  CAAATATGAC  CTCTATGGGA
RNGST1a     GACCAGAGCC  ATTCTCAACT  ACATCGCCAC  CAAATATGAC  CTCTATGGGA
RNGST1b     GACCAGAGCC  ATTCTCAACT  ACATCGCCAC  CAAATATGAC  CTCTATGGGA
RNGST2      GACCAGAGCC  ATTCTCAACT  ACATTGCCAC  CAAATACAAC  CTCTATGGGA
```

```
GGGSTA3     AAGACCTGAA  GGAGAGAGCC  CTGATTGACA  TGTATGTTGG  GGGAACAGAT
OCGSTA1     AAGACATGAA  AGAGAGAGCC  CTGATTGATA  TGTATACAGA  AGGTGTGGCC
OCGSTA2     AAGACATAAA  GGAGAGAGCC  CTGATTGATA  TGTATACAGA  AGGCATAGTA
MMGST2      AGGACATGAA  GGAGAGAGCC  ATCATTGACA  TGTACACAGA  AGGAGTGGCG
MMGST8      AGGACCTGAA  GGAGAGAGTC  AGGATTGACA  TGTATGCAGA  TGGCACCCAG
RRGST8      AGGACCTGAA  GGAGAGAGTC  AGGATTGACA  TGTATGCCGA  TGGCACCCAG
HSGSTA1     AAGACATAAA  GGAGAGAGCC  CTGATTGATA  TGTATATAGA  AGGTATAGCA
HSGSTA2     AAGACATAAA  GGAGAAAGCC  CTGATTGATA  TGTATATAGA  AGGTATAGCA
MMGST1      AGGACATGAA  GGAGAGAGCC  CTGATTGACA  TGTATACAGA  AGGTATTTTA
RNGST1a     AGGACATGAA  GGAGAGAGCC  CTGATTGACA  TGTATTCAGA  GGGTATTTTA
RNGST1b     AGGACATGAA  GGAGAGAGCC  CTGATTGACA  TGTACACCGA  AGGCATTTTA
RNGST2      AGGACATGAA  GGAGAGAGCC  CTCATCGACA  TGTATGCAGA  AGGAGTGGCG

GGGSTA3     GACCTTATGG  GCTTCTTGTT  GAGTTTCCCG  TTCTTGTCAG  CTGAGGATAA
OCGSTA1     GATTTGTATG  AATTGGTCTT  ATTACTGCCG  CTGTGTCCCC  CAGAGCAAAA
OCGSTA2     GATTTGAATG  AATTGATTCT  TACTCGTCCA  TTCCTTCCAC  CGGAGGAACA
MMGST2      GATCTGGAGA  TAATGATTCT  CTATTACCCC  CACATGCCCC  CTGAGGAGAA
MMGST8      GACCTGATGA  TGATGATTGC  CGTGGCTCCA  TTTAAAACCC  CCAAGGAAAA
RRGST8      GACCTGATGA  TGATGATTAT  CGGGGCTCCA  TTTAAAGCCC  CTCAGGAAAA
HSGSTA1     GATTTGGGTG  AAATGATCCT  CCTTCTGCCC  GTATGTCCAC  CTGAGGAAAA
HSGSTA2     GATTTGGGTG  AAATGATCCT  TCTTCTGCCC  TTTACTCAAC  CTGAGGAACA
MMGST1      GATCTGACTG  AAATGATTGG  GCAATTGGTA  TTATGTCCCC  CAGACCAAAG
RNGST1a     GATCTGACTG  AAATGATTAT  CCAATTGGTA  ATATGTCCCC  CAGACCAAAG
RNGST1b     GATCTGACTG  AAATGATTAT  GCAATTGGTA  ATATGTCCCC  CAGACCAAAA
RNGST2      GATCTGGATG  AAATAGTTCT  CCATTACCCT  TACATTCCCC  CTGGGGAGAA

GGGSTA3     GGTGAAACAA  TGTGCCTTTG  TAGTTGAGAA  GGCTACAAGC  AGGTACTTCC
OCGSTA1     AGATGCCAAG  GTTGACTTCA  TCAAAGAGAA  AATCAGAACC  CGCTACTTTC
OCGSTA2     AGAGGCAAAA  CTTGCTCAGA  TCAAAGATAA  AGCAAAAAAC  CGTTATTTTC
MMGST2      AGAGGCAAGC  CTTGCCAAGA  TCAAGGAACA  AACCAGGAAC  CGTTACTTCC
MMGST8      AGAGGAGAGC  TATGATTTGA  TACTGTCAAG  AGCTAAAACC  CGTTACTTCC
RRGST8      AGAAGAGAGC  CTAGCTTTAG  CAGTGAAGAG  GGCTAAAAAC  CGTTACTTCC
HSGSTA1     AGATGCCAAG  CTTGCCTTGA  TCAAGGAGAA  AATAAAAAAT  CGCTACTTCC
HSGSTA2     AGATGCCAAG  CTTGCCTTGA  TCCAAGAGAA  AACAAAAAAT  CGCTACTTCC
MMGST1      AGAAGCCAAG  ACTGCCTTGG  CAAAAGACAG  GACCAAAAAC  CGTTACTTGC
RNGST1a     AGAAGCCAAG  ACCGCCTTGG  CAAAAGACAG  GACCAAAAAC  CGGTACTTGC
RNGST1b     AGAAGCCAAG  ACCGCCTTGG  CAAAAGACAG  GACCAAAAAC  CGGTACTTGC
RNGST2      AGAGGCAAGT  CTTGCCAAAA  TCAAGGACAA  AGCAAGGAAC  CGTTACTTTC

GGGSTA3     CAGCATATGA  AAAGGTTTTG  AAAGACCATG  GCCAGGACTT  TCTTGTTGGC
OCGSTA1     CTGCCTTTGA  AAAAGTGTTG  AAGAGCCACG  GACAAGACTA  CCTTGTTGGC
OCGSTA2     CTGCCTTTGA  AAAGGTGTTG  AAGAGCCACG  GACAAGACTA  CCTTGTTGGC
MMGST2      CTGCCTTTGA  AAAGGTGTTG  AAGAGCCATG  GACAAGATTA  TCTCGTTGGC
MMGST8      CAGTGTTTGA  AAAGATTTTA  AAAGACCACG  GAGAGGCTTT  TCTCGTTGGC
RRGST8      CAGTGTTTGA  AAAGATTTTA  AAAGACCATG  GAGAGGCATT  TCTTGTTGGC
HSGSTA1     CTGCCTTTGA  AAAAGTCTTA  AAGAGCCATG  GACAAGACTA  CCTTGTTGGC
HSGSTA2     CTGCCTTTGA  AAAAGTCTTA  AAGAGCCACG  GACAAGACTA  CCTTGTTGGC
MMGST1      CTGCCTTTGA  AAAGGTGTTG  AAGAGCCATG  GACAAGACTA  CCTTGTGGGC
RNGST1a     CTGCCTTTGA  AAAGGTGTTG  AAGAGCCATG  GCCAAGACTA  CCTTGTAGGT
RNGST1b     CTGCCTTTGA  AAAGGTGTTG  AAGAGCCATG  GCCAAGACTA  CCTTGTAGGT
RNGST2      CTGCCTTTGA  AAAGGTGTTG  AAGAGCCATG  GACAAGATTA  TCTCGTTGGC

GGGSTA3     AACCGTCTCA  GCTGGGCAGA  TATTCATCTT  CTTGAAGCCA  TTTTAATGGT
OCGSTA1     AACAGGCTGA  GCAAGGCTGA  CATTCTCCTG  GTTGAACTTC  TCTACAATGT
OCGSTA2     AACAAGCTGA  GCAAGGCTGA  CATTCTCCTG  GTTGAACTTC  TCTACAACGT
MMGST2      AACAGGCTGA  GCAGGGCTGA  TATTGCCCTG  GTTGAACTCC  TCTACCATGT
MMGST8      AACCAGCTCA  GTTGGGCAGA  CATCCAGCTC  CTAGAAGCCA  TTTTGATGGT
RRGST8      AACCAACTCA  GTTGGGCAGA  CATACAGCTA  CTAGAAGCCA  TTTTGATGGT
HSGSTA1     AACAAGCTGA  GCCGGGCTGA  CATTCATCTG  GTGGAACTTC  TCTACTACGT
HSGSTA2     AACAAGCTGA  GCCGGGCTGA  CATTCACCTG  GTGGAACTTC  TCTACTACGT
MMGST1      AACAGGCTGA  CCAGGGTGGA  CGTCCACCTG  CTGGAACTTC  TTCTCTATGT
RNGST1a     AACAGGCTGA  CCCGGGTAGA  CATCCACCTG  CTGGAACTTC  TCCTCTATGT
RNGST1b     AACAAGCTGA  CCCGGGTAGA  CATCCACCTG  CTGGAACTTC  TCCTCTATGT
RNGST2      AATAGGCTGA  GCAGGGCTGA  TGTTTACCTA  GTTCAAGTTC  TCTACCATGT
```

```
GGGSTA3     AGAAGAGAAG  AAGTCAGACG  CTCTCTCGGG  ATTTCCTCTG  TTACAGGCAT
OCGSTA1     GGAAGAGCTC  GACCCCAGTG  CGATCGCCAG  CTTCCCTCTG  CTGAAGGCCC
OCGSTA2     GGAAGAGCTC  AACCCCGGCG  CGACTGCCAG  CTTCCCTCTG  CTGCAGGCCC
MMGST2      GGAAGAGCTG  GACCCGGGCG  TTGTGGACAA  CTTCCCTCTC  CTGAAAGCGC
MMGST8      GGAAGAACTC  AGTGCCCCTG  TACTGTCCGA  CTTCCCTCTG  CTGCAGGCAT
RRGST8      GGAAGAAGTC  AGTGCTCCTG  TGTTGTCTGA  CTTCCCTCTG  CTGCAGGCAT
HSGSTA1     CGAGGAGCTT  GACTCCAGTC  TTATCTCCAG  CTTCCCTCTG  CTGAAGGCCC
HSGSTA2     GGAAGAGCTT  GACTCTAGCC  TTATTTCCAG  CTTCCCTCTG  CTGAAGGCCC
MMGST1      TGAAGAGCTT  GATGCCAGCC  TTCTGACCCC  TTTCCCTCTG  CTGAAGGCCT
RNGST1a     TGAAGAGTTT  GATGCCAGCC  TTCTGACCTC  TTTCCCTCTG  CTGAAGGCCT
RNGST1b     TGAAGAGTTT  GATGCCAGCC  TTCTGACCTC  TTTCCCTCTG  CTGAAGGCCT
RNGST2      GGAAGAGCTG  GACCCCAGCG  CTTTGGCCAA  CTTCCCTCTG  CTGAAGGCCC

GGGSTA3     TTAAAAAAAG  GATAAGCAGC  ATCCCCACAA  TCAAGAAGTT  CCTGGCGCCT
OCGSTA1     TGAAAACCAG  AATCAGCAGC  CTCCCCACTG  TGAAGAAGTT  TCTGCAGCCT
OCGSTA2     TGAAAACCAG  GATCAGCAAT  CTCCCCACCG  TGAAGAAGTT  TCTGCAGCCT
MMGST2      TGAGAAGCAG  AGTCAGCAAC  CTCCCCACAG  TGAAGAAGTT  TCTTCAACCT
MMGST8      TTAAGACAAG  AATCAGCAAC  ATTCCTACAA  TTAAGAAGTT  CCTGCAACCC
RRGST8      TTAAGACAAG  AATCAGCAAC  ATTCCTACAA  TTAAGAAGTT  CCTGCAACCT
HSGSTA1     TGAAAACCAG  AATCAGCAAC  CTGCCCACAG  TGAAGAAGTT  TCTACAGCCT
HSGSTA2     TGAAAACCAG  AATCAGTAAC  CTGCCCACAG  TGAAGAAGTT  TCTACAGCCT
MMGST1      TCAAGAGCAG  AATCAGCAGC  CTCCCCAATG  TGAAGAAGTT  CCTACAGCCT
RNGST1a     TCAAGAGCAG  AATCAGCAGC  CTCCCCAATG  TGAAGAAGTT  CCTGCAGCCT
RNGST1b     TCAAGAGCAG  AATCAGCAGC  CTCCCCAATG  TGAAGAAGTT  CCTGCAGCCT
RNGST2      TGAGAACCAG  AGTCAGCAAC  CTCCCCACAG  TGAAGAAGTT  TCTTCAGCCT

GGGSTA3     GGAAGCAAGA  GAAAACCTAT  TTCTGATGAT  AAATACGTGG  AGACTGTGAG
OCGSTA1     GGTAGCCAGA  GGAAGCCTCC  TATGGATGAG  AAAAATTTAG  AAAAAGCAAA
OCGSTA2     GGCAGCCAGA  GGAATCCGCC  TGATGATGAG  AAATGCAGAG  AAGAAGCAAA
MMGST2      GGCAGCCAGA  GGAAGCCTTT  TGATGACGCA  AAATGTGTTG  AGTCAGCAAA
MMGST8      GGAAGTCAGA  GGAAGCCTCC  TCCAGATGGC  CCCTATGTTG  AGGTGGTCAG
RRGST8      GGAAGTCAGA  GGAAGCCACC  TCCGGATGGC  CACTATGTTG  ACGTGGTCAG
HSGSTA1     GGCAGCCCAA  GGAAGCCTCC  CATGGATGAG  AAATCTTTAG  AAGAAGCAAG
HSGSTA2     GGCAGCCCAA  GGAAGCCTCC  CATGGATGAG  AAATCTTTAG  AAGAATCAAG
MMGST1      GGCAGCCAGA  GAAAGCCTCC  CTTGGATGCA  AAACAAATTG  AAGAAGCAAG
RNGST1a     GGCAGTCAGA  GAAAGCCAGC  CATGGATGCA  AAACAAATCG  AAGAAGCAAG
RNGST1b     GGCAGTCAGA  GAAAGCTTCC  CATGGATGCA  AAACAAATCG  AAGAAGCAAG
RNGST2      GGCAGCCAGA  GGAAGCCATT  AGAGGATGAG  AAATGTGTAG  AATCTGCAGT

GGGSTA3     GAGGGTTCTC  CGTATGTATT  ACGATGTAAA  ACCACATTAG
OCGSTA1     GAAGATTTTC  ----------  --------AA  GATACCATAA
OCGSTA2     AATCATTTTC  ----------  ----------  ----CATTAA
MMGST2      GAAGATTTTC  ----------  --------AG  T------TAA
MMGST8      GATTGTCCTG  ----------  --------AA  G---TTCTAG
RRGST8      GACCGTCCTG  ----------  --------AA  G---TTCTAG
HSGSTA1     GAAGATTTTC  ----------  --------AG  G---TTTTAA
HSGSTA2     GAAGATTTTC  ----------  --------AG  G---TTTTAA
MMGST1      GAAGGTTTTC  ----------  --------AA  G---TTTTAG
RNGST1a     GAAGGTTTTC  ----------  --------AA  G---TTTTAG
RNGST1b     GAAGATTTTC  ----------  --------AA  G---TTTTAG
RNGST2      TAAGATCTTC  ----------  --------AG  T------TAA
```

(b) Mu class nucleotide sequences.

```
FHGST1      ATGCCA----  --------GC  CAAACTCGGA  TACTGGAAAA  TAAGAGGGCT
SJGST1      ATGTCC----  --------CC  TATACTAGGT  TATTGGAAAA  TTAAGGGCCT
SMGST1      ATGGCA----  --------CC  TAAGTTCGGT  TATTGGAAAG  TCAAAGGCCT
MAGSTM1     ATGCCT----  --------GT  GACACTGGGT  TACTGGGACA  TCCGTGGGCT
HSGSTM2     ATGCCC----  --------AT  GACACTGGGG  TACTGGAACA  TCCGCGGGCT
CLGST1      ATGCCT----  --------AT  GATACTGGGA  TACTGGAATG  TCCGCGGTCT
HSGSTM1b    ATGCCC----  --------AT  GATACTGGGG  TACTGGGACA  TCCGCGGGCT
HSGSTM3     ATGTCGTGCG  AGTCGTCTAT  GGTTCTCGGG  TACTGGGATA  TTCGTGGGCT
HSGSTM1a    ATGCCC----  --------AT  GATACTGGGG  TACTGGGACA  TCCGCGGGCT
MMGST3      ATGCCT----  --------AT  GATACTGGGA  TACTGGAACG  TCCGCGGACT
RNGST3      ATGCCT----  --------AT  GATACTGGGA  TACTGGAACG  TCCGCGGGCT
RNGST4      ATGCCC----  --------AT  GACACTGGGT  TACTGGGACA  TCCGTGGGCT
HSGSTM4     ATGTCC----  --------AT  GACACTGGGG  TACTGGGACA  TCCGCGGGCT
HSGSTM5     ATGCCC----  --------AT  GACTCTGGGG  TACTGGGACA  TCCGTGGGCT
GGGSTM2     ATGGTG----  --------GT  CACGTTGGGT  TATTGGGACA  TCCGCGGGTT
MMGST4      ATGCCT----  --------AT  GACACTAGGT  TACTGGGACA  TCCGTGGGCT

FHGST1      CCAACAACCC  GTTCGACTCT  TGCTCGAATA  CCTGGGTGAA  GAGTACGAAG
SJGST1      TGTGCAACCC  ACTCGACTTC  TTTTGGAATA  TCTTGAAGAA  AAATATGAAG
SMGST1      TGTACAACCA  ACTCGACTTC  TTTTGGAACA  CCTTGAAGAA  ACTTATGAGG
MAGSTM1     GGCTCATGCC  ATCCGCCTGC  TCTTGGAGTA  CACAGACACA  AGCTATGAGG
HSGSTM2     GGCCCATTCC  ATCCGCCTGC  TCCTGGAATA  CACAGACTCA  AGCTACGAGG
CLGST1      GACAAACCCC  ATCCGCCTGC  TCCTGGAATA  CACAGACTCA  AGCTATGAGG
HSGSTM1b    GGCCCACGCC  ATCCGCCTGC  TCCTGGAATA  CACAGACTCA  AGCTATGAGG
HSGSTM3     GGCGCACGCC  ATCCGCCTGC  TCCTGGAGTT  CACGGATACC  TCTTATGAGG
HSGSTM1a    GGCCCACGCC  ATCCGCCTGC  TCCTGGAATA  CACAGACTCA  AGCTATGAGG
MMGST3      GACACACCCG  ATCCGCATGC  TCCTGGAATA  CACAGACTCA  AGCTATGATG
RNGST3      GACACACCCG  ATCCGCCTGC  TCCTGGAATA  CACAGACTCA  AGCTATGAGG
RNGST4      AGCGCATGCC  ATCCGCCTGC  TCCTGGAATA  CACAGACTCG  AGCTATGAGG
HSGSTM4     GGCCCACGCC  ATCCGCCTGC  TCCTGGAATA  CACAGACTCA  AGCTACGAGG
HSGSTM5     GGCCCACGCC  ATCCGCTTGC  TCCTGGAATA  CACAGACTCA  AGCTATGTGG
GGGSTM2     GGCCCACGCC  ATCCGCCTGC  TGCTGGAGTA  CACCGAGACC  CCCTACCAGG
MMGST4      GGCTCACGCC  ATCCGCCTGC  TCCTGGAATA  CACAGACACA  AGCTATGAGG

FHGST1      AACATCTGTA  C---------  GGTCGT---G  AT---GATAG  GGAGAAATGG
SJGST1      AGCATTTGTA  T---------  GAGCGC---G  AT---GAAGG  TGATAAATGG
SMGST1      AACGTGCGTA  T---------  GATCGC---A  AT---GAAAT  CGATGCCTGG
MAGSTM1     AAAAGAAATA  CACCATGGGG  GACGCTCCCA  ACTTTGACCG  AAGCCAGTGG
HSGSTM2     AAAAGAAGTA  CACGATGGGG  GACGCTCCTG  ATTATGACAG  AAGCCAGTGG
CLGST1      AGAAGAAATA  CACCATGGGG  GACGCTCCTG  ACTCTGACAG  AAGCCAGTGG
HSGSTM1b    AAAAGAAGTA  CACGATGGGG  GACGCTCCTG  ATTATGACAG  AAGCCAGTGG
HSGSTM3     AGAAACGGTA  CACGTGCGGG  GAAGCTCCTG  ACTATGATCG  AAGCCAATGG
HSGSTM1a    AAAAGAAGTA  CACGATGGGG  GACGCTCCTG  ATTATGACAG  AAGCCAGTGG
MMGST3      AGAAGAGATA  CACCATGGGT  GACGCTCCCG  ACTTTGACAG  AAGCCAGTGG
RNGST3      AGAAGAGATA  CGCCATGGGC  GACGCTCCCG  ACTATGACAG  AAGCCAGTGG
RNGST4      AGAAGAGATA  CACCATGGGA  GACGCTCCCG  ACTTTGACAG  AAGCCAGTGG
HSGSTM4     AAAAGAAGTA  TACGATGGGG  GACGCTCCTG  ACTATGACAG  AAGCCAGTGG
HSGSTM5     AAAAGAAGTA  CACGATGGGG  GACGCTCCTG  ACTATGACAG  AAGCCAGTGG
GGGSTM2     AGCGGCGGTA  CAAAGCCGGG  CCAGCCCCCG  ACTTCGACCC  GAGCGATTGG
MMGST4      ACAAGAAATA  CACCATGGGG  GACGCTCCTG  ACTATGACCG  AAGCCAGTGG

FHGST1      TTTGGCGATA  AATTCAACAT  GGGATTGGAT  TTGCCAAATT  TACCATACTA
SJGST1      CGAAACAAAA  AGTTTGAATT  GGGTTTGGAG  TTTCCCAATC  TTCCTTATTA
SMGST1      AGCAACGATA  AATTTAAATT  AGGCCTGGAG  TTCCCAAATC  TTCCTTATTA
MAGSTM1     CTCAATGAGA  AGTTCAAGCT  GGGCCTGGAC  TTCCCCAATC  TGCCCTACTT
HSGSTM2     CTGAATGAAA  AATTCAAGCT  GGGCCTGGAC  TTTCCCAATC  TGCCCTACTT
CLGST1      CTGAATGAGA  AGTTCAAGCT  GGGCCTGGAC  TTTCCCAATC  TGCCCTACTT
HSGSTM1b    CTGAATGAAA  AATTCAAGCT  GGGCCTGGAC  TTTCCCAATC  TGCCCTACTT
HSGSTM3     CTGGATGTGA  AATTCAAGCT  AGACCTGGAC  TTTCCTAATC  TGCCCTACCT
HSGSTM1a    CTGAATGAAA  AATTCAAGCT  GGGCCTGGAC  TTTCCCAATC  TGCCCTACTT
MMGST3      CTGAATGAGA  AGTTCAAGCT  GGGCCTGGAC  TTTCCCAATC  TGCCTTACTT
RNGST3      CTGAATGAGA  AGTTCAAACT  GGGCCTGGAC  TTCCCCAATC  TGCCCTACTT
RNGST4      CTGAATGAGA  AGTTCAAACT  GGGCCTGGAC  TTCCCCAATC  TGCCCTACTT
HSGSTM4     CTGAATGAAA  AATTCAAGCT  GGGCCTGGAC  TTTCCCAATC  TGCCCTACTT
HSGSTM5     CTGAATGAAA  AATTCAAGCT  GGGCCTGGAC  TTTCCCAATC  TGCCCTACTT
```

```
GGGSTM2     ACCAATGAGA  AGGAGAAGCT  GGGCCTCGAC  TTCCCCAACC  TGCCCTATCT
MMGST4      CTGAGTGAGA  AGTTCAAGCT  GGGCCTGGAC  TTTCCCAATC  TGCCCTACTT

FHGST1      CATTGACGAT  AAGTGCAAAC  TGACTCAGTC  GGTGGCCATA  ATGCGGTACA
SJGST1      TATTGATGGT  GATGTTAAAT  TAACACAGTC  TATGGCCATC  ATACGTTATA
SMGST1      TATTGATGGT  GATTTTAAAT  TAACACAATC  TATGGCTATC  ATACGTTATA
MAGSTM1     AATTGATGGA  TCACACAAGA  TCACCCAGAG  CAACGCCATC  CTGCGCTACA
HSGSTM2     GATTGATGGG  ACTCACAAGA  TCACCCAGAG  CAATGCCATC  CTGCGGTACA
CLGST1      AATTGATGGG  TCCCACAAGA  TCACCCAGAG  CAACGCCATC  CTGCGCTACA
HSGSTM1b    GATTGATGGG  GCTCACAAGA  TCACCCAGAG  CAACGCCATC  TTGTGCTACA
HSGSTM3     CCTGGATGGG  AAGAACAAGA  TCACCCAGAG  CAATGCCATC  TTGCGCTACA
HSGSTM1a    GATTGATGGG  GCTCACAAGA  TCACCCAGAG  CAACGCCATC  TTGTGCTACA
MMGST3      GATCGATGGA  TCACACAAGA  TCACCCAGAG  CAATGCCATC  CTGCGCTACC
RNGST3      AATTGATGGA  TCGCGCAAGA  TTACCCAGAG  CAATGCCATA  ATGCGCTACC
RNGST4      AATTGATGGG  TCACACAAGA  TCACCCAGAG  CAATGCCATC  CTGCGCTATC
HSGSTM4     GATTGATGGG  GCTCACAAGA  TCACCCAGAG  CAACGCCATC  CTGTGCTACA
HSGSTM5     GATTGATGGG  GCTCACAAGA  TCACCCAGAG  CAATGCCATC  CTGCGCTACA
GGGSTM2     CATTGATGGC  GACGTCAAAC  TGACCCAGAG  CAACGCCATC  CTGCGCTACA
MMGST4      GATTGATGGA  TCACACAAGA  TCACCCAGAG  CAATGCCATC  CTGCGCTACC

FHGST1      TTGCGGACAA  GCATGGAATG  CTTGGTACCA  CACCCGAGGA  ACGAGCTCGA
SJGST1      TAGCTGACAA  GCACAACATG  TTGGGTGGTT  GTCCAAAAGA  GCGTGCAGAG
SMGST1      TAGCTGACAA  ACACAACATG  TTGGGGGCTT  GTCCAAAAGA  ACGTGCGGAA
MAGSTM1     TCGCCCGCAA  GCACGACCTG  TGTGGGGAGA  CAGAGGAGGA  GAGGATTCAG
HSGSTM2     TTGCCCGCAA  GCACAACCTG  TGCGGGGAAT  CAGAAAAGGA  GCAGATTCGC
CLGST1      TTGCCCGCAA  GCACAACCTG  TGTGGAGAGA  CAGAGGAGGA  GAGGATTCGT
HSGSTM1b    TTGCCCGCAA  GCACAACCTG  TGTGGGGAGA  CAGAAGAGGA  GAAGATTCGT
HSGSTM3     TCGCTCGCAA  GCACAACATG  TGTGGTGAGA  CTGAAGAAGA  AAAGATTCGA
HSGSTM1a    TTGCCCGCAA  GCACAACCTG  TGTGGGGAGA  CAGAAGAGGA  GAAGATTCGT
MMGST3      TTGCCCGAAA  GCACCACCTG  GATGGAGAGA  CAGAGGAGGA  GAGGATCCGT
RNGST3      TTGCCCGCAA  GCACCACCTG  TGTGGAGAGA  CAGAGGAGGA  GCGGATTCGT
RNGST4      TTGGCCGCAA  GCACAACCTG  TGTGGGGAGA  CAGAAGAGGA  GAGGATTCGT
HSGSTM4     TTGCCCGCAA  GCACAACCTG  TGTGGGGAGA  CAGAAGAGGA  GAAGATTCGT
HSGSTM5     TTGCCCGCAA  GCACAACCTG  TGTGGGGAGA  CAGAAGAGGA  GAAGATTCGT
GGGSTM2     TCGCCCGCAA  GCACAACATG  TGTGGGGAGA  CGGAGGTGGA  GAAGCAGCGC
MMGST4      TTGCCCGAAA  GCACAACCTG  TGTGGAGAGA  CAGAGGAGGA  GAGGATTCGT

FHGST1      ATTTCGATGA  TCGAAGGAGC  TGCAATGGAT  CTTCGGATGG  GTTTTGTTCG
SJGST1      ATTTCAATGC  TTGAAGGAGC  GGTTTTGGAT  ATTAGATACG  GTGTTTCGAG
SMGST1      ATTTCGATGC  TTGAAGGAGC  GGTTTTGGAT  ATTAGGATGG  GTGTTTTAAG
MAGSTM1     CTGGACATTC  TGGAGAACCA  GGCTATGGAC  ACCCGCATGC  AGTTGGCCAT
HSGSTM2     GAAGACATTT  TGGAGAACCA  GTTTATGGAC  AGCCGTATGC  AGCTGGCCAA
CLGST1      GTGGACATTG  TGGAGAACCA  GGCTATGGAC  ACCCGCATGC  AGCTCATCAT
HSGSTM1b    GTGGACATTT  TGGAGAACCA  GACCATGGAC  AACCATATGC  AGCTGGGCAT
HSGSTM3     GTGGACATCA  TAGAGAACCA  AGTAATGGAT  TTCCGCACAC  AACTGATAAG
HSGSTM1a    GTGGACATTT  TGGAGAACCA  GACCATGGAC  AACCATATGC  AGCTGGGCAT
MMGST3      GCAGACATTG  TGGAGAACCA  GGTCATGGAC  ACCCGCATGC  AGCTCATCAT
RNGST3      GCAGACATTG  TGGAGAACCA  GGTCATGGAC  AACCGCATGC  AGCTCATCAT
RNGST4      GTGGACATTC  TGGAGAATCA  GCTCATGGAC  AACCGCATGG  TGCTGGCGAG
HSGSTM4     GTGGACATTT  TGGAGAACCA  GGCTATGGAC  GTCTCCAATC  AGCTGGCCAG
HSGSTM5     GTGGACATTT  TGGAGAACCA  GGTTATGGAT  AACCACATGG  AGCTGGTCAG
GGGSTM2     GTGGACGTGT  TGGAGAACCA  CCTGATGGAT  CTGCGCATGG  CCTTCGCGCG
MMGST4      GTGGACATTT  TGGAGAACCA  GGCTATGGAC  ACCCGCATAC  AGTTGGCCAT

FHGST1      TGTTTGTTAC  AACCCAAAAT  TTGAAGAAGT  GAAAGGAGAT  TATCTGAAAG
SJGST1      AATTGCATAT  AGTAAAGACT  TTGAAACTCT  CAAAGTTGAT  TTTCTTAGCA
SMGST1      AATCGCATAC  AATAAGGAAT  ATGAAACCCT  CAAAGTTGAT  TTTCTCAACA
MAGSTM1     GGTTTGCTAC  AGCCCTGACT  TTGAGAAACG  AAAGCCAGAA  TACTTGGAGG
HSGSTM2     ACTCTGCTAT  GACCCAGATT  TTGAGAAACT  GAAACCAGAA  TACCTGCAGG
CLGST1      GCTCTGTTAC  AACCCTGACT  TTGAGAAGCA  GAAGCCAGAG  TTCTTGAAGA
HSGSTM1b    GATCTGCTAC  AATCCAGAAT  TTGAGAAACT  GAAGCCAAAG  TACTTGGAGG
HSGSTM3     GCTCTGTTAC  AGCTCTGACC  ACGAAAAACT  GAAGCCTCAG  TACTTGGAAG
HSGSTM1a    GATCTGCTAC  AATCCAGAAT  TTGAGAAACT  GAAGCCAAAG  TACTTGGAGG
MMGST3      GCTCTGTTAC  AACCCTGACT  TTGAGAAGCA  GAAGCCAGAG  TTCTTGAAGA
RNGST3      GCTTTGTTAC  AACCCCGACT  TTGAGAAGCA  GAAGCCAGAG  TTCTTGAAGA
RNGST4      ACTTTGCTAT  AACCCTGACT  TTGAGAAGCT  GAAGCCAGGG  TACCTGGAGC
HSGSTM4     AGTCTGCTAC  AGCCCTGACT  TTGAGAAACT  GAAGCCAGAA  TACTTGGAGG
HSGSTM5     ACTGTGCTAT  GACCCAGATT  TTGAGAAACT  GAAGCCAAAA  TACTTGGAGG
```

```
GGGSTM2     GCTCTGCTAC  AGCCCCGACT  TCGAGAAGCT  GAAGCCGGCG  TACCTGGAGC
MMGST4      GGTTTGCTAC  AGCCCTGACT  TTGAGAAAAA  GAAGCCAGAG  TACTTAGAGG

FHGST1      AACTGCCAAC  AACATTGAAG  ATGTGGTCCA  ATTTTCTTGG  AGATCGTCAC
SJGST1      AGCTACCTGA  AATGCTGAAA  ATGTTCGAAG  ATCGTTTATG  TCATAAAACA
SMGST1      AACTTCCTGG  GAGGCTGAAA  ATGTTCGAAG  ATCGTTTGTC  TAACAAAACT
MAGSTM1     GTCTTCCCGA  GAAGATGAAG  CTGTACTCGG  AGTTCCTGGG  CAAGAGATCA
HSGSTM2     CACTCCCTGA  AATGCTGAAG  CTCTACTCAC  AGTTTCTGGG  GAAGCAGCCA
CLGST1      CCATCCCTGA  GAAGATGAAG  ATGTACTCTG  AGTTCCTGGG  CAAGCGGCCA
HSGSTM1b    AACTCCCTGA  AAAGCTAAAG  CTCTACTCAG  AGTTTCTGGG  GAAGCGGCCA
HSGSTM3     AGCTACCTGG  ACAACTGAAA  CAATTCTCCA  TGTTTCTGTG  GAAATTCTCA
HSGSTM1a    AACTCCCTGA  AAAGCTAAAG  CTCTACTCAG  AGTTTCTGGG  GAAGCGGCCA
MMGST3      CCATCCCTGA  GAAAATGAAG  CTCTACTCTG  AGTTCCTGGG  CAAGAGGCCA
RNGST3      CCATCCCTGA  GAAGATGAAG  CTCTACTCTG  AGTTCCTGGG  CAAGCGACCA
RNGST4      AACTGCCTGG  AATGATGCGG  CTTTACTCCG  AGTTCCTGGG  CAAGCGGCCA
HSGSTM4     AACTTCCTAC  AATGATGCAG  CACTTCTCAC  AGTTCCTGGG  GAAGAGGCCA
HSGSTM5     AACTCCCTGA  AAAGCTAAAG  CTCTACTCAG  AGTTTCTGGG  GAAGCGGCCA
GGGSTM2     TGCTGCCAGG  GAAACTGAGG  CAGTTGTCGC  GCTTCCTGGG  CTCCCGGTCC
MMGST4      GTCTCCCTGA  GAAGATGAAG  CTCTACTCTG  AGTTTCTGGG  CAAGCAGCCA

FHGST1      TATTTGACAG  GTTCTCCAGT  TAGCCATGTG  GACTTTATGG  TTTACGAAGC
SJGST1      TATTTAAATG  GTGATCATGT  AACCCATCCT  GACTTCATGT  TGTATGACGC
SMGST1      TATTTGAACG  GTAATTGTGT  AACTCATCCT  GACTTTATGT  TATACGATGC
MAGSTM1     TGGTTTGCAG  GGGACAAGAT  CACCTATGTG  GATTTCCTCA  TCTATGATGT
HSGSTM2     TGGTTTCTTG  GGGACAAGAT  CACCTTTGTG  GATTTCATCG  CTTATGATGT
CLGST1      TGGTTTGCAG  GGGACAAGGT  CACCTTATGT  GGATTTCTCG  CTTATGATGT
HSGSTM1b    TGGTTTGCAG  GAAACAAGAT  CACTTTTGTA  GATTTTCTCG  TCTATGATGT
HSGSTM3     TGGTTTGCCG  GGGAAAAGCT  CACCTTTGTG  GATTTTCTCA  CCTATGATAT
HSGSTM1a    TGGTTTGCAG  GAAACAAGAT  CACTTTTGTA  GATTTTCTCG  TCTATGATGT
MMGST3      TGGTTTGCAG  GGGACAAGGT  CACCTATGTG  GATTTCCTTG  CTTATGACAT
RNGST3      TGGTTTGCAG  GGGACAAGGT  CACCTATGTG  GATTTCCTTG  CTTATGACAT
RNGST4      TGGTTTGCAG  GGGACAAGAT  CACCTTTGTG  GATTTCATTG  CTTACGATGT
HSGSTM4     TGGTTTGTTG  GAGACAAGAT  CACCTTTGTA  GATTTCCTCG  CCTATGATGT
HSGSTM5     TGGTTTGCAG  GAGACAAGAT  CACCTTTGTG  GATTTCCTTG  CCTATGATGT
GGGSTM2     TGGTTTGTGG  GGGACAAGCT  CACCTTCGTG  GACTTCCTGG  CTTACGATGT
MMGST4      TGGTTTGCAG  GGAACAAGGT  CACCTATGTG  GATTTTCTTG  TTTATGATGT

FHGST1      ATTGGACTGT  ATTCGTTATT  TGGCACCACA  GTGTCTGGAG  GACTTCCCCA
SJGST1      TCTTGATGTT  GTTTTATACA  TGGACCCAAT  GTGCCTGGAT  GCGTTCCCAA
SMGST1      CCTTGATGTG  GTTTTATACA  TGGACTCACA  GTGCTTGAAC  GAGTTTCCAA
MAGSTM1     CCTTGACCAG  CACCGTATAT  TTGCGCCAAA  GTGCCTGGAT  GCATTTCCAA
HSGSTM2     CCTTGAGAGA  AACCAAGTAT  TTGAGCCCAG  CTGCCTGGAT  GCCTTCCCAA
CLGST1      CCTTGATCAG  TATCAAATGT  TTGAGCCCAA  GTGCCTGGAT  CCCTTCCCAA
HSGSTM1b    CCTTGACCTC  CACCGTATAT  TTGAGCCCAA  CTGCTTGGAC  GCCTTCCCAA
HSGSTM3     CTTGGATCAG  AACCGTATAT  TTGACCCCAA  GTGCCTGGAT  GAGTTCCCAA
HSGSTM1a    CCTTGACCTC  CACCGTATAT  TTGAGCCCAA  GTGCTTGGAC  GCCTTCCCAA
MMGST3      TCTTGACCAG  TACCGTATGT  TTGAGCCCAA  GTGCCTGGAC  GCCTTCCCAA
RNGST3      TCTTGACCAG  TACCACATTT  TTGAGCCCAA  GTGCCTGGAC  GCCTTCCCAA
RNGST4      TCTTGAGAGG  AACCAAGTGT  TTGAGGCCAC  GTGCCTGGAC  GCGTTCCCAA
HSGSTM4     CCTTGACCTC  CACCGTATAT  TTGAGCCCAA  CTGCTTGGAC  GCCTTCCCAA
HSGSTM5     CCTTGACATG  AAGCGTATAT  TTGAGCCCAA  GTGCTTGGAC  GCCTTCCTAA
GGGSTM2     GCTGGACCAG  CAGCGCATGT  TCGTCCCCGA  CTGCCCCGAG  CTGCAGGGCA
MMGST4      CCTTGATCAA  CACCGAATAT  TTGAGCCCAA  GTGCCTGGAT  GCCTTCCCAA

FHGST1      AATTGAAGGA  ATTCAAGAGT  CGTATTGAAG  ATCTTCCAAA  AATCAAGGCA
SJGST1      AATTAGTTTG  TTTTAAAAAA  CGTATTGAAG  CTATCCCACA  AATTGATAAG
SMGST1      AATTAGTTTC  TTTCAAAAAG  TGTATTGAAG  ATTTACCACA  AATCAAGAAC
MAGSTM1     ACCTGAAGGA  CTTCTTGGCT  CGCTTTGAGG  GCCTGAAGAA  GATATCTGAC
HSGSTM2     ACCTGAAGGA  CTTCATCTCC  CGATTTGAGG  GCTTGGAGAA  GATCTCTGCC
CLGST1      ACCTGAAGGA  CTTCTTGGCC  CGCTTTGAGG  GCCTGAAGAA  GATCTCTGCC
HSGSTM1b    ATCTGAAGGA  CTTCATCTCC  CGCTTTGAGG  GCTTGGAGAA  GATCTCTGCC
HSGSTM3     ACCTGAAGGC  TTTCATGTGC  CGTTTTGAGG  CTTTGGAGAA  AATCGCTGCC
HSGSTM1a    ATCTGAAGGA  CTTCATCTCC  CGCTTTGAGG  GCTTGGAGAA  GATCTCTGCC
MMGST3      ACCTGAGGGA  CTTCCTGGCC  CGCTTCGAGG  GCCTCAAGAA  GATCTCTGCC
RNGST3      ACCTGAAGGA  CTTCCTGGCC  CGCTTCGAGG  GCCTGAAGAA  GATCTCTGCC
RNGST4      ACCTGAAGGA  TTTCATAGCG  CGCTTTGAGG  GCCTGAAGAA  GATCTCCGAC
HSGSTM4     ATCTGAAGGA  CTTCATCTCC  CGCTTTGAGG  GCTTGGAGAA  GATCTCTGCC
HSGSTM5     ACTTGAAGGA  CTTCATCTCC  CGCTTTGAGG  GTTTGAAGAA  GATCTCTGCC
```

```
GGGSTM2     ACCTGAGCCA  ATTCCTGCAG  CGCTTCGAGG  CCCTGGAGAA  GATCTCTGCC
MMGST4      ACCTGAAGGA  CTTCATGGGT  CGCTTTGAGG  GCCTGAAGAA  GATATCTGAC

FHGST1      TACATGGAAT  CAGAGAAGTT  CATCAAGTGG  CCTTTGAACT  CGTGGATTGC
SJGST1      TACTTGAAAT  CCAGCAAGTA  TATAGCATGG  CCTTTGCAGG  GCTGGCAAGC
SMGST1      TACTTAAATT  CTAGCAGGTA  CATAAAATGG  CCTCTGCAAG  GTTGGGATGC
MAGSTM1     TACATGAAAA  GCAGCCGCTT  TAGCTGCAAA  CAAATCTTTG  CAAAGATGGC
HSGSTM2     TACATGAAGT  CCAGCCGCTT  CCTCCCAAGA  CCTGTGTTCA  CAAAGATGGC
CLGST1      TACATGAAGA  CCAGCCGCTT  CCTCCGGAGA  CCTATATTTT  CAAAGATGGC
HSGSTM1b    TACATGAAGT  CCAGCCGCTT  CCTCCCAAGA  CCTGTGTTCT  CAAAGATGGC
HSGSTM3     TACTTACAGT  CTGATCAGTT  CTGCAAGATG  CCCATCAACA  ACAAGATGGC
HSGSTM1a    TACATGAAGT  CCAGCCGCTT  CCTCCCAAGA  CCTGTGTTCT  CAAAGATGGC
MMGST3      TACATGAAGA  GTAGCCGCTA  CATCGCAACA  CCTATATTTT  CAAAGATGGC
RNGST3      TACATGAAGA  GCAGCCGCTA  CCTCTCAACA  CCTATATTTT  CGAAGTTGGC
RNGST4      TACATGAAGT  CCAGCCGCTT  CCTCCCAAGA  CCTCTGTTCA  CAAAGATGGC
HSGSTM4     TACATGAAGT  CCAGCCGCTT  CCTCCCAAAA  CCTCTGTACA  CAAGGGTGGC
HSGSTM5     TACATGAAGT  CCAGCCAATT  CCTCCGAGGT  CTTTTGTTTG  GAAAGTCAGC
GGGSTM2     TATATGCGCT  CGGGGCGCTT  TATGAAGGCC  CCCATTTTCT  GGTACACGGC
MMGST4      TACATGAAGA  GCAGCCGCTT  CCTCTCCAAG  CCAATCTTTG  CAAAGATGGC

FHGST1      TTCTTTCGGT  GGTGGAGACG  CTGCACCGGC  T---------  TGA
SJGST1      CACGTTTGGT  GGTGGCGACC  ATCCTCCAAA  A---------  TAA
SMGST1      CACGTTTGGT  GGTGGAGATA  CTCCTCCAAA  A---------  TAG
MAGSTM1     CGTTTGG---  ----------  --AACTCAAA  A---------  TAG
HSGSTM2     TGTCTGG---  ----------  --GGCAACAA  G---------  TAG
CLGST1      ACAGTGG---  ----------  --AGTAACAA  G---------  TAG
HSGSTM1b    TGTCTGG---  ----------  --GGCAACAA  G---------  TAG
HSGSTM3     CCAGTGG---  ----------  --GGCAACAA  GCCTGTATGC  TGA
HSGSTM1a    TGTCTGG---  ----------  --GGCAACAA  G---------  TAG
MMGST3      CCACTGG---  ----------  --AGTAACAA  G---------  TAG
RNGST3      CCAATGG---  ----------  --AGTAACAA  G---------  TAG
RNGST4      TATTTGG---  ----------  --GGCAGCAA  G---------  TAG
HSGSTM4     TGTCTGG---  ----------  --GGCAACAA  G---------  TAA
HSGSTM5     TACATGG---  ----------  --AACAGCAA  A---------  TAG
GGGSTM2     GCTGTGG---  ---------A  ACAACAAGAA  AGAG------  TGA
MMGST4      CTTTTGG---  ----------  --AACCCAAA  G---------  TAG
```

(c) Pi class nucleotide sequences.

```
CEGSTP1     ATGACCCTC-  --AAGCTCAC  GTACTTCGAC  ATCCACGGAC  TCGCTGAGCC
BTGSTP1     ATGCCTCCCT  ACACCATCGT  CTACTTCCCG  GTTCAAGGGC  GCTGCGAGGC
MMGST7      ATGCCACCAT  ACACCATTGT  CTACTTCCCA  GTTCGAGGGC  GGTGTGAGGC
RNGST7      ATGCCGCCGT  ACACCATTGT  GTACTTCCCA  GTTCGAGGGC  GCTGTGAGGC
RNGSTPPS    ATGCCGCCAT  ACACCATTGT  CTACTTCCCA  GTTCGAGGGC  GCTGTGAGGC
HSGSTP1     ATGCCGCCCT  ACACCGTGGT  CTATTTCCCA  GTTCGAGGCC  GCTGCGCGGC

CEGSTP1     AATCCGTCTT  CTTCTCGCGG  ACAAGCAAGT  TGCCTACGAG  GATCATCGTG
BTGSTP1     CATGCGCATG  CTGCTGGCCG  ACCAGGGCCA  GAGCTGGAAG  GAGGAGGTCG
MMGST7      CATGCGAATG  CTGCTGGCTG  ACCAGGGCCA  GAGCTGGAAG  GAGGAGGTGG
RNGST7      CACGCGCATG  CTGCTGGCTG  ACCAGGGCCA  GAGCTGGAAG  GAGGAGGTGG
RNGSTPPS    CATGGCATGC  TGCTGGCTGA  CCAGGGCCAG  AGCTAGAAAG  AGGA-GT--G
HSGSTP1     CCTGCGCATG  CTGCTGGCAG  ATCAGGGCCA  GAGCTGGAAG  GAGGAGGTGG

CEGSTP1     TAACCTATGA  ACAATGG---  ---GCTGATA  TTAAACCAAA  GATGATCTTC
BTGSTP1     TAGCCATGCA  GAGCTGGCTG  CAGGGCCCAC  TCAAGGCCTC  CTGCCTGTAC
MMGST7      TTACCATAGA  TACCTGGATG  CAAGGCTTGC  TCAAGCCCAC  TTGTCTGTAT
RNGST7      TTACCATAGA  TGTCTGGCTT  CAAGGCTCGC  TCAAGTCCAC  TTGTCTGTAT
RNGSTPPS    GTACCATAGA  TGTCTGGCTT  CAAGGCTCAC  TCAAGCCCAC  TCATCTGTAT
HSGSTP1     TGACCGTGGA  GACGTGGCAG  GAGGGCTCAC  TCAAAGCCTC  CTGCCTATAC
```

```
CEGSTP1     GGCCAGGTTC  CATGTCTTCT  ATCCGGAGAC  GAGGAGATTG  TTCAATCTGG
BTGSTP1     GGGCAGCTCC  CCAAGTTCCA  GGACGGAGAC  CTCACGCTGT  ACCAGTCCAA
MMGST7      GGGCAGCTCC  CCAAGTTTGA  GGATGGAGAC  CTCACCCTTT  ACCAATCTAA
RNGST7      GGGCAGCTCC  CCAAGTTTGA  AGATGGAGAC  CTCACCCTTT  ACCAATCTAA
RNGSTPPS    GGGCAGCTAC  CCAAGTTTGA  GGATGGAGAC  CTCACCCTTT  ACCAGTCTAA
HSGSTP1     GGGCAGCTCC  CCAAGTTCCA  GGACGGAGAC  CTCACCCTGT  ACCAGTCCAA

CEGSTP1     AGCTATCATC  CGTCATCTCG  CTCGTCTTAA  TGGGCTCAAT  GGCTCCAACG
BTGSTP1     TGCCATCCTG  CGGCACCTGG  GCCGCACCCT  CGGGCTGTAT  GGGAAGGACC
MMGST7      TGCCATCTTG  AGACACCTTG  GCCGCTCTTT  GGGGCTTTAT  GGGAAAAACC
RNGST7      TGCCATCTTG  AGGCACCTGG  GTCGCTCTTT  AGGGCTTTAT  GGGAAAGACC
RNGSTPPS    TGCCATCTTG  AGGCACCTGG  GCCGCTCTTT  AGGGCTTTAT  GGGAAAGACC
HSGSTP1     TACCATCCTG  CGTCACCTGG  GCCGCACCCT  TGGGCTCTAT  GGGAAGGACC

CEGSTP1     AGACAGAGAC  AACTTTCATC  GACATGTTCT  ACGAAGGACT  TCGTGATTTG
BTGSTP1     AGCAGGAGGC  GGCCCTGGTG  GACATGGTGA  ATGACGGTGT  AGAGGACCTT
MMGST7      AGAGGGAGGC  CGCCCAGATG  GATATGGTGA  ATGATGGGGT  GGAGGACCTT
RNGST7      AGAAGGAGGC  TGCCTTGGTG  GATATGGTGA  ATGATGGGGT  GGAGGACCTT
RNGSTPPS    AGAGGAAGGC  CGCCCTAGTG  GATATCGTGA  ATGATGGAGT  GGAGGACCTT
HSGSTP1     AGCAGGAGGC  AGCCCTGGTG  GACATGGTGA  ATGACGGCGT  GGAGGACCTC

CEGSTP1     CACACCAAGT  ACACCACTAT  GATCTACAGA  AACTACGAAG  ACGGCAAGGC
BTGSTP1     CGCTGCAAAT  ACGTCTCCCT  CATTTACACC  AACTACGAGG  CGGGCAAGGA
MMGST7      CGCGGCAAAT  ATGTCACCCT  CATCTACACC  AACTATGAGA  ATGGTAAGAA
RNGST7      CGATGCAAAT  ATGGTACCCT  CATCTACACT  AACTATGAGA  ATGGTAAGGA
RNGSTPPS    CGCTGCAAAT  ATGGTACCTT  CATCTACACT  AACTATGAGA  ATGGTAAGGA
HSGSTP1     CGCTGCAAAT  ACATCTCCCT  CATCTACACC  AACTATGAGG  CGGGCAAGGA

CEGSTP1     TCCGTACATC  AAGGACGTTC  TTCCAGGAGA  GCTCGCTCGT  CTGGAGAAGC
BTGSTP1     GGACTATGTG  AAGG---CGC  TGCCCCAGCA  CCTGAAGCCT  TTCGAGACCC
MMGST7      TGACTACGTG  AAGG---CCC  TGCCTGGGCA  TCTGAAGCCT  TTTGAGACCC
RNGST7      TGACTATGTG  AAGG---CCC  TGCCTGGGCA  TCTGAAACCT  TTTGAGACCC
RNGSTPPS    TGACTATGTG  AAAT---CCC  TGCCTGGGCA  TCTGAAGCCT  TTGAAACCCT
HSGSTP1     TGACTATGTG  AAGG---CAC  TGCCCGGGCA  ACTGAAGCCT  TTTGAGACCC

CEGSTP1     TTTTCCATAC  CTACAAGAAC  GGAGAGCACT  ACGTTATTGG  AGACAAGGAA
BTGSTP1     TGCTGTCCCA  GAACAAGGGT  GGCCAGGCCT  TCATCGTGGG  CGACCAGATC
MMGST7      TGCTGTCCCA  GAACCAGGGA  GGCAAAGCTT  TCATCGTGGG  TGACCAGATC
RNGST7      TGCTGTCCCA  GAACCAGGGA  GGCAAAGCTT  TCATTGTGGG  TAACCAGATT
RNGSTPPS    GCTGTCCCAG  AACCAGGGAG  GCAAAGCTTT  CATCATGGGT  GACCAGATTT
HSGSTP1     TGCTGTCCCA  GAACCAGGGA  GGCAAGACCT  TCATTGTGGG  AGACCAGATC

CEGSTP1     AGCTATGCGG  ATTATGTGCT  GTTCGAGGAG  CTCGACATTC  ATTTGATTCT
BTGSTP1     TCCTTTGCGG  ACTACAACCT  GCTGGACCTG  CTTCGGATTC  ACCAGGTCCT
MMGST7      TCCTTTGCCG  ATTACAACTT  GCTGGACCTG  CTGCTGATCC  ACCAAGTCCT
RNGST7      TCCTTTGCAG  ATTACAACTT  GCTGGACCTG  CTGCTGGTCC  ACCAAGTCCT
RNGSTPPS    CCTTTACTGA  TTACAACTTG  CTGGACCTGC  TGCTGGTCCA  CCAAGTGCTG
HSGSTP1     TCCTTCGCTG  ACTACAACCT  GCTGGACTTG  CTGCTGATCC  ATGAGGTCCT

CEGSTP1     CACACCAAAT  GCTCTTGATG  GTGTTCCAGC  ACTCAAGAAG  TTCCACGAGA
BTGSTP1     GGCCCCCAGC  TGTCTGGACT  CCTTCCCCCT  GCTCTCAGCC  TACGTGGCCC
MMGST7      GGCCCCTGGC  TGCCTGGACA  ACTTCCCCCT  GCTCTCTGCC  TATGTGGCTC
RNGST7      GGCCCCTGGC  TGCCTGGACA  ACTTCCCCCT  GCTCTCTGCC  TATGTGGCTC
RNGSTPPS    ACCCTGGCTG  CCTGGACAAC  TT---CCCCT  GCTCTCTGCC  TATGTGGCTC
HSGSTP1     AGCCCCTGGC  TGCCTGGATG  CGTTCCCCCT  GCTCTCAGCA  TATGTGGGGC

CEGSTP1     GATTCGCTGA  GCGTCCAAAC  ATCAAGGCAT  ATCTCAACAA  GAGAGCTGCT
BTGSTP1     GTCTCAACTC  CCGGCCCAAG  CTCAAGGCCT  TCCTGGCCTC  CCCCGAGCAC
MMGST7      GCCTCAGTGC  CCGGCCCAAG  ATCAAGGCCT  TTCTGTCCTC  CCCGGAACAT
RNGST7      GCCTCAGTGC  CCGCCCCAAG  ATCAAGGCCT  TTCTGTCCTC  CCCTGACCAT
RNGSTPPS    ACCTCAGTGC  CCGCCCCAAG  ATCAAGGCCT  TTCTGTCCTC  CCCCGACCAT
HSGSTP1     GCCTCAGCGC  CCGGCCCAAG  CTCAAGGCCT  TCCTGGCCTC  CCCTGAGTAC
```

```
CEGSTP1     ATCAACCCAC  CAGTAAATGG  AAATGGAAAA  CAATAA
BTGSTP1     ATGAACCGGC  CCATCAACGG  CAATGGGAAA  CAGTGA
MMGST7      GTGAACCGTC  CCATCAATGG  CAATGGCAAA  CAGTAG
RNGST7      TTGAACCGTC  CCATCAACGG  CAATGGTAAA  CAGTAG
RNGSTPPS    GTGAACCATC  CCATCAACGG  CAATGGTAAG  ------
HSGSTP1     GTGAACCTCC  CCATCAATGG  CAACGGGAAA  CAGTGA
```

Abramovitz, M. and Listowsky, I. (1987) Selective expression of a unique glutathione S-transferase Yb 3 gene in rat brain. *J. Biol. Chem.* **262**:7770-7773.

Aceto, A., Di Ilio, C., Angelucci, S., Felaco, M. and Federici, G. (1989) Glutathione transferase isoenzymes from human testis. *Biochem. Pharmacol.* **38**:3653-2660.

Adachi, Y. and Yamamoto, T. (1987) Partial defect in hepatic glutathione S-transferase activity in a case of Rotor's syndrome. *Gastro. Japonica* **22**:34-38.

Adang, A.E., Brussee, J., Meyer, D.J., Coles, B., Ketterer, B., van der Gen, A. and Mulder, G.J. (1988) Substrate specificity of rat liver glutathione S-transferase isoenzymes for a series of glutathione analogs. *Biochem. J.* **255**:721-.

Adang, A.E., Meyer, D.J., Brussee, J., van der Gen, A., Ketterer, B. and Mulder, G.J.(1989) Interaction of rat glutathione S-transferases 7-7 and 8-8 with glutamyl- or glycyl-modified glutathione S-transferases. *Biochem. J.* **264**:759-764.

Adang, A.E., Morree, W.J., Brussee, J., Mulder, G.J. and van der Gen, A. (1991) Inhibition of glutathioneS-transferase 3-3 by glutathione derivatives that bind covalently to the active site. *Biochem. J.* **278**:63-68.

Ahmad, H., Wilson, D.E., Fritz, R.R., Singh, S.V., Medh, R.D., Nagle, G.T., Awasthi, Y.C. and Kurosky, A. (1990) Primary and secondary sequence structural analyses of glutathione S-transferase π from human placenta. *Arch. Biochem. Biophys.* **278**:398-408.

Akiyama, K. and Abe, K. (1984) The gene frequencies of glutathione S-transferase isoenzymes in Japanese population. *Jpn. J. Hum. Genet.* **29**:271-276.

Ålin, P., Mannervik,B.and Jornvall,H. (1985). Structural evidence for three different types of glutathione transferase in human tissues. *FEBS Lett.* **182**:319-322.

Ames B.N., Profet ,M. and Gold, L.S. (1990a) Dietary pesticides (99.99% all natural). *Proc. Natl. Acad. Sci. USA* **87**:7777-7781.

Ames B.N., Profet ,M. and Gold, L.S. (1990b) Nature's chemicals and synthetic chemicals: Comparative toxicology. *Proc. Natl. Acad. Sci. USA* **87**:7782-7786.

Ames, B.N. (1983) Dietary carcinogens and anticarcinogens. Oxygen radicals and degenerative diseases. *Science* **221**:1256-1264.

Anderson, L. (1981) Identification of mitochondrial proteins and some of their precursors in two-dimensional electrophoretic maps of human cells. *Proc.Natl.Acad.Sci.U.S.A.* **78**:2407-2411.

Andersson, C., Mosialou, E., Adang, A.E., Mulder, G.J., Van der Gen, A., and Morgenstern, R. (1991) Studies on the activity and activation of rat liver microsomal glutathione transferase with a series of glutathione analogues. *J. Biol. Chem.* **266**: 2076-2079.

Andreadis, A., Gallego, M.E., and Nadal-Ginard, B. (1987) Generation of protein isoform diversity by alternative splicing: mechanistic and biological implications. *Annu. Rev. Cell Biol.* **3**: 207-242.

Atchison, M. and Adesnik, M. (1986). Gene conversion in a cytochrome P-450 gene family. *Proc.Natl.Acad.Sci.U.S.A.* **83**:2300-2304.

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K. (1987) *Current Protocols in Molecular Biology* John Wiley and Sons, New York.

Awasthi,Y.C., Singh,S.V., Ahmad,H., Moller,P.C.and Gupta,V. (1988) Expression of glutathione S-transferase isoenzymes in human small cell lung cancer cell lines. *Carcinogenesis* **9**:89-93.

Balabaskaran, S., and Muniandy, N. (1984) Glutathione S-transferase from *Hevea brasiliensis*. *Phytochemistry* **23**:251-256.

Balloul, J.M., Sondermeyer, P., Dreyer, D., Capron, M., Grzych, J.M., Pierce, R.J., Carvallo, D., Lecocq, J.P., and Capron, A. (1987) Molecular cloning of a protective antigen of schistosomes. *Nature* **326**:149-153.

Bass, N.M., Kirsch, R.E., Tuff, S.A., Marks, I. and Saunders, S.J. (1977). Ligandin heterogeneity : evidence that the two non-identical subunits are the monomers of two distinct proteins. *Biochim.Biophys.Acta* **492**:163-175.

Batist, G., Tulpule, A., Sinha, B.K., Katki, A.G., Meyers, C.E. and Cowen, K.H. (1986) Overexpression of a novel anionic glutathione transferase in multi-drug-resistant human breast cancer cells. *J. Biol. Chem.* **261**:15544-15549.

Beall, C., Fyrberg, C., Song, S. and Fyrberg, E. (1992). Isolation of a Drosophila gene encoding glutathione S-transferase. *Biochem.Genet.* **30**:515-527.

Benachenhou-Lahfa, N., Forterre, P. and Labedan, B. (1993). Evolution of glutamate dehydrogenase genes: evidence for two paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life. *J.Mol.Evol.* **36**:335-346.

Bennett, C.F. and Yeoman, L.C. (1985) Co-localization of nonhistone protein BA with U-snRNPs to the same regions of the cell nucleus. *Exp. Cell. Res.* **157**:379-386.

Bennett, C.F., Rosenfeld,B.I., Huang,C.H. and Yeoman,L.C. (1982) Evidence for two conformational forms of nonhistone protein BA which differ in their affinity for DNA. *Biochem. Biophys. Res. Commun.* **104**:649-656.

Bennett, C.F., Spector, D.L. and Yeoman, L.C. (1986) Nonhistone protein BA is a glutathione S-transferase localised to interchromatinic regions of the cell nucleus. *J.Cell.Biol.* **102**:600-609.

Benson, A.M., Talalay, P., Keen, J.H. and Jakoby, W.B. (1977) Relationship between the soluble glutathione-depenent Δ^5-ketosteroid isomerase and the glutathione S-transferases of the liver. *Proc. Natl. Acad. Sci. USA.* **74**:158-162.

Benton, W.D. and Davis, P.W. (1977) Screening kgt recombinant clones by hybridization to single plaques *in situ. Science* **196:** 180-182.

Berhane,K. and Mannervik,B. (1990). Inactivation of the genotoxic aldehyde acrolein by human glutathione transferases of classes alpha, mu, and pi. *Mol.Pharmacol.* **37**:251-254.

Beutler, E. (1975) *Red Cell Metabolism. A Manual of Biochemical Methods.* 2nd ed. Grune and Stratton New York.

Beutler, E., Dunning, D., Dabe, I.B., and Forman, L. (1988) Erythrocyte glutathione S-transferase deficiency and hemolytic anemia. *Blood* **72**:73-77.

Bhargava, M.M., Listowsky, I., and Arias, I.M. (1978) Ligandin. Bilirubin binding and glutathione-S-transferase activity are independent processes. *J. Biol. Chem.* **253**:4112-4115.

Björnestedt, R., Stenburg, G., Widerstan, M., Board, P.G. Sinning, I., Jones, T.A. and Mannervik, B. (1994) Contribution of arginine15 to the catalytic mechanism human class Alpha glutathione transferase A1-1. *J. Biol. Chem.* (in press).

Board, P. G., and Mannervik, B. (1991) The contribution of the carboxy terminal sequence to the catalytic activity of GST2: A human alpha class glutathione transferase. *Biochem. J.* **275**:171-174.

Board, P.G. (1980) A method for the localization of glutathione S-transferase isoenzymes after starch gel electrophoresis. *Anal. Biochem.* **105**:147-149.

Board, P.G. (1981a) Biochemical genetics of glutathione S-transferase in man. *Am. J. Hum. Genet.* **33**:36-43.

Board, P.G. (1981b) Gene deletion and partial deficiency of the glutathione transferase (ligandin) system in man. *FEBS Lett.* **135**:12-14.

Board, P.G. (1990) Genetic polymorphism of glutathione S-transferase in man. *In: Glutathione S-Transferases and Drug Resistance* Hayes, J.D., Pickett, C.B., and Mantle, T.J. (Eds) pp35-45. Taylor and Francis, London.

Booth, J. Boyland,E. and Sims, P. (1961). An enzyme from rat liver catalysing conjugations with glutathione. *Biochem. J.* **79**:516-24.

Board, P.G. and Pierce, K. (1987) Expression of human glutathione S-transferase 2 in Escherichia coli. Immunological comparison with the basic glutathione S-transferase isoenzymes from human liver. *Biochem. J.* **248**:937-941.

Board, P.G. and Webb, G.C. (1987) Isolation of a cDNA clone and localization of human glutathione S-transferase 2 genes to chromosome band 6p12. *Proc. Natl. Acad. Sci. USA.* **84**:2377-2381.

Board, P.G., Coggan, M. and Woodcock, D.M. (1992) The human Pi class glutathione transferase sequence at 12q13-q14 is a reverse-transcribed pseudogene. *Genomics* **14**:470-473.

Board, P.G., Coggan, M., Johnston, P.N., Ross, V., Suzuki, T., and Webb, G. (1990a) Genetic heterogeneity of the human glutathione transferases: A complex of gene families. *Pharmac. Ther.* **48**:357-369.

Board, P.G., Johnston, P.N., Ross, V.L., Webb, G.C., Coggan, M., and Suzuki, T. (1990b) Molecular genetics of the human glutathione S-transferase. *Int. Symp. Princess. Takamatsu. Cancer Res. Fund.* **21**: 199-211.

Board, P.G., Ross, V.L., Coggan, M. and Suzuki, T. (1993) Polymorphic expression of human glutathione transferases. In: *Structure and function in the Glutathione transferases* Tew, K.D., Pickett, C.B., Mantle, T.J. and Hayes, J.D. (Eds.) CRC Press, Boca Raton, Florida.

Board, P.G., Russell, Morano, R.J. and Oakeshott, J.G. (1994) Purification, molecular cloning and heterologous expression of a glutathione S-transferase from the Australian sheep blowfly (*Lucillia cuprina*). *Biochem. J.* **299**:425-430.

Board, P.G., Suzuki, T. and Shaw, D.C. (1988) Human muscle glutathione S-transferase (GST4) shows close homology to human liver GST1. *Biochim. Biophys. Acta.* **953**:214-217.

Board, P.G., Webb, G.C. and Coggan, M. (1989) Isolation of a cDNA clone of a human glutathione S-transferase 3 gene which localizes to chromosome bands 11q13 and 12q13-14. *Ann.Hum.Gen.* **53**:205-213.

Bora, P.S., Bora, N.S., Wu, X. and Lange, L.G. (1991) Molecular cloning, sequencing and expression of human myocardial fatty acid ethyl ester synthase-III cDNA. *J. Biol. Chem.* **266**:16774-7.

Bora, P.S., Spilburg, C.A. and Lange, L.G. (1989a) Identification of a satellite fatty acid ethyl ester synthase from human myocardial fatty acid ethyl ester synthase-III cDNA. *J. Clin. Invest.* **84**:1942-6.

Bora, P.S., Spilburg, C.A. and Lange, L.G. (1989b) Metabolism of ethanol and carcinogens by glutathione transferases. *Proc. Natl. Acad. Sci. USA* **86**:4470-4473.

Boyland, E. and Chasseaud, L.F. (1969) The role of glutathione and glutathione S-transferases in mercapturic acid biosynthesis. *Adv. Enzymol. Rel. Areas Mol. Biol.* **32**:173-219.

Britten, R.J. (1986) Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**:1393-1398.

Brophy, P.M., Southan, C. and Barrett, J. (1989). Glutathione transferases in the tapeworm Moniezia expansa. *Biochem.J.* **262**:939-946.

Buckle,V.J. and Craig, I.W. (1986) *In situ* hyridization. In: *Human Genetic Diseases-A Practical Approach.* Davies K.E. (Ed), pp85-100. IRL Press Oxford, Washington.

Buetler, T.M. and Eaton, D.L (1992) Glutathione S-transferases: Amino acid sequence comparison, classification and phylogenetic relationship. *Environ. carcino. and Ecotox. Rev.* **C10**:181-203.

Bulmer M, Wolfe KH, Sharpe PM. (1991) Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. *Proc. Natl. Acad. Sci. USA* , **8:** 5974-5978.

Burgess, J.R., Chow, N.W.I., Reddy, C. and Tu, C.-P.D. (1989) Amino acid specificities in the prostaglandin endoperoxide conversion pathway. *Biochem. Biophys. Res. Commun.* **158**:497-502.

Burk, R.F., Trumble, M.J. and Lawrence, R.A. (1980) Rat hepatic cytosolic glutathione-dependent enzyme protection against lipid peroxidation in the NADPH-microsomal lipid peroxidatiion system. *Biochim. Biophys Acta* **618**: 35-41.

Campbell, E., Takahashi, Y., Abramovitz, M., Peretz, M. and Listowsky, I. (1990) A distinct human testis and brain class-I glutathione S-transferase; molecular cloning and characterization of a form present even in individuals lacking hepatic type-I isoenzymes. *J. Biol. Chem.* **265**:9188-9193.

Cantatore,P. and Saccone,C. (1987) Organization, structure, and evolution of mammalian mitochondrial genes. *Int.Rev.Cytol.* **108**:149-208.

Carmichael, J., Forrester, L.M., Lewis, A.D., Hayes J.D., Hayes, P.C. and Wolf, C.R. (1988) Glutathione S-transferase isoenymes and glutathione peroxidase activity in normal and tumour samples from human lung . *Carcinogenesis* **9**:31617-41621

Carroll, R.L. (1988) *Vertebrate Paleontology and Evolution*. Freeman, San Franscisco.

Cavalier-Smith, T. (1987) In: *Evolutionary Biology of the Fungi* Rayner, A.D.M., Brasier, C.M. and Moore, D. (Eds), Cambridge University Press pp339-353.

Chang, C., Saltzman, A.G., Sorensen, N.S., Hiipakka, R.A. and Liao, S. (1987) Identification of glutathione S-transferase Yb1 mRNA as the androgen-repressed mRNA by cDNA cloning and sequence analysis. *J. Biol. Chem.* **262**:11901-11903.

Chang, L-H., Fan, J-Y., Liu, L-F. Tsai, S-P. and Tam, M.F. (1992) Cloning and expression of a chick liver.glutathione S-transferase CL3 subunit with the use of a baculvirus expression system. *Biochem. J.* **281**:545-551.

Chasseaud, L.F. (1979). The role of glutathione and glutathione S-transferases in the metabolism of chemical carcinogens and other electrophilic agents. *Adv.Cancer Res.* **29**:175-274.

Chelly, J., Concordet, J.P., Kaplan, J.C. and Kahn, A. (1989) Illegitimate transcription: transcription of any gene in any cell type. *Proc. Natl. Acad. Sci. USA*. **86**:2617-2621.

Chou, T.-C., Zachar, Z., Bingham, P. (1987) Developmental expression of a regulatory gene is programed at the level of splicing. *EMBO J.* **6**:4095-4104.

Chow, N.-W.I., Whang-Peng, J., Kao-Shan, C.-S., Tam, M.F., Lai, H.-C.J., and Tu, C.-P.D. (1988) Human glutathione S-transferases. The Ha multigene family encodes products of different but overlapping substrate specificities. *J. Biol. Chem.* **263**:12797-12800.

Clark, A.C., Letoa, M.and Wong, S.T. (1977) the purification by affinity chromatography of a glutathione S-transferase from a larvae of *Galleria mellonella*. *Life Sci.* **20**:141-148.

Clark, A.G. (1989) The comparative enzymology of the glutathione S-transferases from non-vertebrate organisms. *Comp. Biochem. Physiol.* **92B**:419-446.

Cochrane, B.J., Hargis, M., de Belligny, P., Holtsberg, F. and Coronella, J. (1992). Evolution of glutathione S-transferases associated with insecticide resistance in *Drosophila*. In: *Chemical Molecular Mechanisms*. Proc. Amer. Chemic. Soc. New York.

Cohen, J.B., Broz, S.D. and Levison, A.D. (1989) Expression of H-ras proto-oncogene is controlled by alternative splicing. *Cell* **58**:461-472.

Collins, A., Keats, B.J., Dracopoli, N., Shields, D.C., and Morton, N.E. (1992) Integration of gene maps: chromosome 1. *Proc. Natl. Acad. Sci. USA.* **89**:4598-602.

Comstock, K.E., Johnson, K.J., Rifenbery, D. and Henner, W.D. (1993). Isolation and analysis of the gene and cDNA for a human Mu class glutathione S-transferase, GSTM4. *J.Biol.Chem.* **268**:16958-16965.

Comstock, K.E., Sanderson, B.J. Claflin, G. and Henner, W.D. (1990) GST1 gene deletion determined by polymerase chain reaction. *Nucl. Acids Res.* **18**:3670-1.

Coschigano,P.W. and Magasanik,B. (1991). The URE2 gene product of Saccharomyces cerevisiae plays an important role in the cellular response to the nitrogen source and has homology to glutathione s-transferases. *Mol.Cell Biol.* **11**:822-832.

Cowan, S.W., Bergfors, T., Jones, T.A., Tibbelin, G., Olin, B., Board, P.G., and Mannervik, B. (1989) Crystallization of GST2, a human class alpha glutathione transferase. *J. Mol. Biol.* **208**:369-370.

Cowell, I.G., Dixon, K.H., Pemble, S.E., Ketterer, B., Taylor, J.B. (1988) The structure of the human glutathione S-transferase p gene. *Biochem. J.* **255**:79-83.

Cygler, M., Schrag, J.D., Sussman, J.L., Harel, M., Silman, I., Gentry, M.K. and Doctor, B.P. (1993) Relationship between sequence conservation and three-dimentional structure in a large family of esterases, lipases and related proteins. *Protein Sci.* **2**:366-382.

Dagert, M. and Ehrlich, S.D. (1977) Prolonged incubation in calcium chloride improves the competence of *Escherichia coli* cells. *Gene* **6**:23-28.

Daniel, V., Sharon, R., and Bensimmon, A. (1989) Regulatory elements controlling the basal and drug-inducible expression of glutathione S-transferase Ya subunit gene. *DNA* **8**:399-408.

Daniel, V., Sharon, R., Tichauer, Y. and Sarid, S. (1987) Mouse glutathione S-transferase Ya subunit: gene structure and sequence. *DNA* **6**:317-324.

Danielson, U.H., and Mannervik, B. (1985) Kinetic independence of the subunits of cytosolic glutathione transferase from the rat. *Biochem. J.* **231**:263-267.

Danielson,U.H., Esterbauer,H. and Mannervik,B. (1987). Structure-activity relationships of 4-hydroxyalkenals in the conjugation catalysed by mammalian glutathione transferases. *Biochem.J.* **247**:707-713.

Davis, D.L. (1987) Paleolithic diet, evolution and carcinogens. *Science* **238**:1633-34.

Davis, D.L. (1989) Natural anticarcinogens, carcinogens, and changing patterns in cancer: Some speculation. *Environ. Res* . **50**:322-340.

Dayhoff, M.O., Schwartz, R.M and Orcutt, B.C. (1978) In: *Atlas of Protein Sequence and Structure.* Dayhoff, M.O. (ed) supplement 3, **5**:345.

DeJong, J.L., Chang, C-M., Whang-Peng, J., Knutsen, T., and Tu, C-P.D. (1988a) The human liver glutathione S-transferase gene superfamily: expression and chromosome mapping of an Hb subunit cDNA. *Nucl. Acids Res.* **16**:8542-8554.

DeJong, J.L., Mohandas, T., and Tu, C-P.D. (1991) The human Hb (Mu) class glutathione S-transferases are encoded by a dispersed gene family. *Biochem. Biophys. Res. Comm.* **180**:15-22.

DeJong, J.L., Morgenstern, R., Jornvall, H., DePierre, J.W., and Tu, C-P.D. (1988b) Gene expression of rat and human microsomal glutathione S-transferases. *J. Biol. Chem.* **263**:8430-8436.

Dekant, W., Lash, L.H., and Anders, M.W. (1988) In: *Glutathione Conjugation in Mechanisms and Biological Significance* Sies, H. and Ketterer, B. (Eds), pp.415-447, Academic Press, London.

Del Boccio, G., Di Ilio, C., Alin, P., Jornvall, H., and Mannervik, B. (1987a) Identification of a novel glutathione transferase in human skin homologous with class alpha glutathione transferase 2-2 in the rat. *Biochem. J.* **244**:21-25.

Del Boccio, G., Di Ilio, C., Miranda, M., Manilla, A., Zarivi, O., Bonfigli, A. and Federici, G. (1987b). Glutathione transferase activity during Bufo bufo development. *Comp.Biochem.Physiol.B.* **86**:749-753.

Di Ilio,C., Del Boccio,G., Aceto,A., Casaccia,R., Mucilli,F.andFederici,G. (1988). Elevation of glutathione transferase activity in human lung tumor. *Carcinogenesis* **9**:335-340.

Diesperger,, H. and Sanderman, H. (1979) Soluble and microsomal glutathione S-transferases in pea seedlings (*Pisum sativum* (L.)) *Plant* **146**:643-648.

Ding, G.J., Lu, A.Y., and Pickett, C.B. (1985) Rat liver glutathione S-transferases. Nucleotide sequence analysis of a Yb1 cDNA clone and prediction of the complete amino acid sequence of the Yb1 subunit. *J. Biol. Chem.* **260**:13268-13271.

Dixon,K.H., Cowell,I.G., Xia,C.L., Pemble,S.E., Ketterer,B.andTaylor,J.B. (1989). Control of expression of the human glutathione S-transferase pi gene differs from its rat orthologue. *Biochem.Biophys.Res.Commun.* **163**:815-822.

Donlon,T.A. (1986) Practical approaches to in situ hybridisation. *Karyogram* **12**:3-10.

Doolittle, R.F. (1981). Similar amino acid sequences: chance or common ancestry?. *Science* **214**:149-159.

Dracopoli, N.C., O'Connell, P., Elsner, T.I., Lalouel, J.M., White, R.L., Buetow, K.H., Nishimura, D.Y., Murray, J.C., Helms, C., Mishra, S.K., and *et al,* (1991) The CEPH consortium linkage map of human chromosome 1. *Genomics.* **9**: 686-700.

Easteal S. (1990) The pattern of molecular evolution and the relative rate of molecular evolution. *Genetics* ; **124:** 165-173

Easteal S. (1992) A mammalian molecular clock? *Bioessays,* **14**:415-419.

Easteal, S. (1988) Rate constancy of globin gene evolution in placental mammals. *Proc. Natl. Acad. Sci. USA.* **85**:7622-7626.

Fahey, R.C. (1977) Biologically important thiol-disulfide reactions and the role of cyst(e)ine in proteins: An evolutionary perspective. *Adv. Exp. Mec. Biol.* **86A**:1-30.

Fahey, R.C. and Sundquist, A.R. (1991) Evolution of glutathione metabolism. *Adv. Enzymol. Relat. Areas. Mol. Biol.* **64**:1-53.

Fahey, R.C., Newton, G.L., Arrick, B., Overdank, Bogart, T. and Aley, S.B. (1984) *Entamoeba histolytica*: a eukaryote without glutathione metabolism. *Science* **224**:70-72.

Farrel, P., Sharp, S. and Defranca, D. (1981) Laboratory Manual, Yale University, Connecticut USA.

Faulder, C.G., Hirrell, P.A., Hume, R. and Strange, R. (1987) Studies of the development of basic, neutral and acidic isoenzymes of glutathione S-transferases in human liver, adrenal, kidney and spleen. *Biochem. J.* **241**:221-228.

Feinberg, A.P. and Vogelstein, B. (1983a) A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **132**:6-13.

Feinberg, A.P. and Vogelstein, B. (1983b) A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Addendum. Anal. Biochem.* **137**:266-267.

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**:783-791.

Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package) version 3.5c Distributed by the author Dept. of Genetics, University of Washington, Seattle.

Felsenstein, J., Sawyer, S. and Kochin, R. (1982). An efficient method for matching nucleic acid sequences. *Nucleic Acids Res.* **10**:133-139.

Felsenstein, K.M. and Goff, S.P. (1988). Expression of the gag-pol fusion protein of Moloney murine leukemia virus without gag protein does not induce virion formation or proteolytic processing. *J.Virol.* **62**:2179-2182.

Feng, D.-F. and Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**:351-360.

Fonknechten, N., Chelly, J., Lepercq, J., Kahn, A., Kaplan, J.C., Kitzis, A., and Chomel, J.C. (1992) CFTR illegitimate transcription in lymphoid cells: quantification and applications to the investigation of pathological transcripts. *Hum. Genet.* **88**: 508-512.

Forrester, L.M., Hayes, J.D., Millis, R., Barnes, D., Harris, A.L., Schlager, J.J., Powis, G. and Wolf, C.R. (1990) Expression of glutathione S-transferases and cytochrome P450 in normal and tumor breast tissue. *Carcinogenisis (Lond.)* **11**:2163-2169.

Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., Wolfe, R.S., Balch, W.E., Tanner, R.S., Magrum, L.J., Zablen, L.B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B.J., Stahl, D.A., Luehrsen, K.R., Chen, K.N. and Woese, C.R. (1980) The phylogeny of prokaryotes. *Science* **209**:457-463.

Friling, R.S., Bensimmon, A., Tichauer, Y., and Daniel, V. (1990) Xenobiotic-inducible expression of murine glutathione S-transferase Ya subunit gene is controlled by an electrophile-responsive element. *Proc. Nat. Acad. Sci. USA.* **87**:6258-6262.

Fryer, A.A., Hume, R., and Strange, R.C. (1986) The development of glutathione S-transferase and glutathione peroxidase activities in human lung. *Biochim. Biophys. Acta* **883**:448-453.

Fryer, A.A., Zhao, L., Alldersea, J., Pearson, W.R. and Strange, R.C. (1993) Use of site-directed mutagenesis of allele-specific PCR primers to identify the GSTM1 A, GSTM1 B, GSTM1 A,B and GSTM1 null polymorphisms at the glutathione S-transferase, GSTM1 locus. *Biochem.J.* **295**:313-315.

Gill, S. S., Ota, K. and Hammock, B.D. (1983) Radiometric assays for mammalian epoxide hydrolases and glutathione S-transferase. *Analyt. Biochem.* **131**:273.

Goosens, M., Dozy, A.M. Embury, S.H. Zachariades, Z., Hadjiminas, M.G., Stamatoyannopoulos, G. and Kan, Y.W. (1980) Triplicated α-globin loci in humans. *Proc. Natl. Acad. Sci. USA.* **77**:518-521.

Gronenborn, B. and Messing, J.N. (1978) Methylation of single-stranded DNA *in vitro* introduduces new restriction endonuclease cleavage sites. *Nature* **272**:375-377.

Grove, G., Zarlengo, R.P., Timmerman, K.P., Li, N-Q., Tam, M.F., and Tu, C-P.D. (1988) Characterization and heterospecific expression of cDNA clones of genes in the maize GSH S-transferase multigene family. *Nucl. Acids Res.* **16**:425-438.

Grunebaum, L., Cazenave, J-P., Camerino, G., Kloepfer, C., Mandel, J-L., Tolstoshev, P., Jaye, M., De La Salle, H. and Lecocq, J-P. (1984) Carrier detection of Hemophilia B by using a restriction site polymorphism associated with the coagulation factor IX gene. *J. Clin. Invest.* **73**:1491-1495.

Guthenberg, C., Alin, P., and Mannervik, B. (1985) Glutathione transferase from rat testis. *Methods Enzymol.* **113**: 507-510.

Guthenberg, C., Warholm, M., Rane, A., and Mannervik, B. (1986) Two distinct forms of glutathione transferase from human foetal liver. Purification and comparison with isoenzymes isolated from adult liver and placenta. *Biochem. J.* **235**:741-745.

Habig, W.H. and Jakoby, W.B. (1981). Glutathione S-transferases (rat and human). *Methods Enzymol.* **77**:218-231.

Habig, W.H., Pabst, M.J. and Jakoby, W.B. (1974b). Glutathione S-transferases: The first enzymatic step in mercapturic acid formation. *J. Biol. Chem.* **249**: 7130-7139.

Habig, W.H., Pabst,M.J., Fleischner,G., Gatmaitan,Z., Arias,I.M. and Jakoby,W.B. (1974a). The identity of glutathione S-transferase B with ligandin, a major binding protein of liver. *Proc. Natl. Acad. Sci. U.S.A.* **71**:3879-3882.

Hapgood, J., Cuthill, S., Denis, M., Poellinger, L., and Gustaffsson, J-A. (1989) Specific protein-DNA interactions at a xenobiotic-responsive element: Copurification of dioxin receptor and DNA-binding activity. *Proc. Nat. Acad. Sci. USA.* **86**:60-64.

Hood,L., Campbell, J.H. and Elgin, S.C. (1975). The organisation, expression and evolution of antibody genes and other multigene families. *Ann. Rev. Genettic* **9**:305-353.

Harada, S., Abei, M., Tanaka, N., Agarwal, D., and Goedde, H.W. (1987) Liver glutathione S-transferase polymorphism in Japanese and its pharmacogenetic importance. *Hum. Genet.* **75**:322-325.

Harden, D.G. and Klinger, H.P. Eds (1985) *"An International System for Human Cytogenetics Nomenclature"*. Klarger, Basel.

Harris, J.M., Meyer, D.J., Coles, B. and Ketterer, B. (1991) A novel glutathione transferase (13-13) isolated from the matrix of rat liver mitochondria having structural similarity to class theta enzymes. *Biochem. J.* **278**:137-141.

Harvey, J.W. and Beutler, E. (1982) Binding of heme by glutathione S-transferase: a possible role for the erythrocyte enzyme. *Blood* **60**:1227-1230.

Hasegawa, M. and Hashimoto, T. (1993) Ribosomal RNA trees misleading? *Nature* **361**:23

Hatayama, I., Satoh, K and Sato, K. (1986) Developmental and hormonal regulation of the major form of hepatic glutathione S-transferase in male mice. *Biochem Biophys. Res. Comm.* **140**:581-588.

Hayashi, S., Watanabe J. and Kawajiri, K. (1992) High susceptibility to lung cancer analysed in terms of combined genotypes of P450IA1 and Mu-class glutathione S-transferase genes. *Jpn. J. Cancer Res.* **83**:866-870.

Hayes, J.D. and Wolf, C.R. (1988) The role of glutathione transferase in drug resistance. In: *Glutathione Conjugation in Mechanisms and Biological Significance* Sies, H. and Ketterer, B. (Eds), pp.315-355, Academic Press, London.

Hayes, J.D., Judah, D.J., Neal, G.E., and Nguyen, T. (1992). Molecular cloning and heterologous expression of a cDNA encoding a mouse glutathione S-transferase Yc subunit possessing high catalytic activity for aflatoxin B1-8,9-epoxide. *Biochem. J.* **285**: 173-180.

Henkle, K.J, Davern, K.M., Wright, M.D., Ramos, A.J. and Mitchell, G.F. (1990) Comparison of the cloned genes of the 26- and 28- kDa glutathione S-transferase of *Schistosoma japonicum* and *Schistosoma masoni*. *Mol. Biochem. Parasitol.* **40**:23-44.

Higgins, D.G. (1993) CLUSTALV, multiple alignment package.

Higgins, D.G. and Sharp, P. M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. *Cabios* **5**:151-153.

Hillis, D.M. and Moritz, C. (1990) *Molecular Systematics*. Sinauer, New York.

Hiratsuka, A., Sebata, N., Kawashima, K., Okuda, H., Ogura, K., Watabe, T., Satoh, K., Hatayama, I., Tsuchida, S., Ishikawa, T., and Sato, K. (1990) A new class of rat glutathione S-transferases Yrs-Yrs inactivating reactive sulphate esters as metabolites of carcinogenic arglmethanols. *J. Biol. Chem.* **265**:11973-11981

Holmes, D.S. and Quigley, M. (1981) A rapid boiling method for the preparation of bacterial plasmids. *Anal. Biochem.* **114**:193-197.

Holmgren,A., Soderberg,B.O., Eklund,H. and Branden,C.I. (1975). Three-dimensional structure of Escherichia coli thioredoxin-S2 to 2.8 A resolution. *Proc.Natl.Acad.Sci.U.S.A.* **72**:2305-2309.

Husby,P., Srai,K.S., Ketterer,B.andRomslo,I. (1981). Effect of ligandin on the efflux of Co-deuteroporphyrin from isolated rat liver mitochondria. *Biochem.Biophys.Res.Commun.* **100**:651-659.

Hussey, A.J., Kerr, L.A., Cronshaw, A.D., Harrison, D.J. and Hayes, J.D. (1991). Variation in the expression of Mu class glutathione S-transferase isoenzymes from human skeletal muscle. Evidence for the existence of heterodimers. *Biochem. J.* **273**, 323-332.

Hussey,A.J. and Hayes,J.D. (1992). Characterization of a human class-Theta glutathione S- transferase with activity towards 1-menaphthyl sulphate. *Biochem.J.* **286**:929-935.

Igarishi, T., Irokawa, N., Ono, S., Ohmori, S., Ueno, K. and Kitagawa, H. (1987) Difference in the effects of phenobarbital and 3-methylcholanthrene on subunit composition of hepatic glutathione S-transferase in male and female rats. *Xenobiotica* **17**:127-137.

Ishigaki, S., Abramovtz, M. and Listowsky, I. (1989) Glutathione S-transferases are major cytosolic thyroid hormone binding proteins. *Arch. Biochem. Biophys.* **273**:265-270.

Islam, M.Q., Platz, A., Szpirer, J., Szpirer, C., Levan, G., and Mannervik, B. (1989) Chromosomal localization of human glutathione transferase genes of classes alpha mu and pi. *Hum. Genet.* **82:**338-342.

Jakoby, W.B. (1978) The glutathione S-transferases: a group of multifunctional detoxification proteins. *Adv. Enzymol.* **46**:383-414.

Jakoby, W.B., and W.H. Habig (1980) Glutathione S-transferases. In: *Enzymatic basis of detoxication.* Jakoby, W.B. (Ed). (**Vol II**) pp63-94, Academic Press, New York.

Jakoby, W.B., Ketterer, B. and Mannervik, B. (1984) Glutathione transferases: nomenclature. *Biochem.Pharmacol.* **33**:2539-2540.

Jakoby, W.B., W.H. Habig,.J.H. Keen, J.N. Kettley, and M.J. Pabst (1976). Glutathione S-transferases: Catalytic aspects. In: *Glutathione: Metabolism and function.* I.M. Arias and W.B. Jakoby. (Eds.). pp189-212, Raven Press, New York.

Ji., X., Zhang, P., Armstrong, R.N. and Gilliland, G.L. (1992) The three dimentional structure of glutathione S-transferase from the mu gene class. Structural analysisof the binary complex of isoenzyme 3-3and glutahtione at 2.2Å resolution. *Biochem.* **27**:10169-10184.

Johnson, J.A., Neal, T.L., Collins, J.H. and Siegel, F.L. (1990). Characterization of methylation of rat liver cytosolic glutathione S-transferases by using reverse-phase h.p.l.c. and chromatofocusing. *Biochem.J.* **270**:483-489.

Kamisaka, K., Habit, W.H., Ketley, J.N., Arias, I.M., and Jakoby, W.B. (1975) Multiple forms of human glutathione S-transferase and their affinity for bilirubin. *Eur. J. Biochem.* **60**:153-161.

Kano, T., Sakai, M. and Muramatsu, M. (1987) Structure and expression of a human class pi glutathione S-transfersase messenger RNA. *Cancer Res.* **47**:5626-5630.

Keen, J.H., and Jakoby, W.B. (1978). Glutathione transferases. Catalysis of nucleophilic reactions of glutathione. *J. Biol. Chem.* **253**: 5654-5657.

Keen, J.H., Habig, W.H., and Jakoby, W.B. (1976). Mechanism for the several activities of the glutathione S- transferases. *J. Biol. Chem.* **251**: 6183-6188.

Keeran, W.S. and Lee, R.F. (1987) The purification and characterization of glutathione S-transferase from the hepatopancreas of blue crab, *Callinectes sapidus*. *Arch. Biochem. Biophys*. **255**:233-243.

Kelly, R.B., Cozzarelli, N.R., Deutscher, M.P., Lehman, I.R., and Kornberg, A. (1970) Enzymatic synthesis of deoxyribonucleic acid. XXXII. Replication of duplex deoxyribonucleic acid by polymerase at a single strand break. *J. Biol. Chem.* **245**: 39-45.

Kempner, E.S., and Miller, J.H. (1972) The molecular biology of Euglena gracilis. VII. Inorganic requirements for a minimal culture medium. *J. Protozool.* **19**: 343-346.

Kenter, A.L. and Birshtein, B.K. (1981). Chi, a promoter of generalized recombination in lambda phage, is present in immunoglobulin genes. *Nature* **293**:402-404.

Kerklaan, P.R., Zoetemelk,C.E. and Mohn,G.R. (1985) Mutagenic activity of various chemicals in Salmonella strain TA100 and glutathione-deficient derivatives. On the role of glutathione in the detoxification or activation of mutagens inside bacterial cells. *Biochem.Pharmacol.* **34**:2151-2156.

Ketterer B., Meyer, D. J. and Clark, A. G. (1988) Soluble glutathione transferase isozymes. *In: Glutathione Conjugation. Mechanisms and Biological Significance*. Sies, H., and Ketterer, B. (Eds), pp 281-314, Academic Press, London.

Ketterer, B., Ross-Mansell, P. and Whitehead, J.K. (1967) The isolation of carcinogen-binding protein from livers of rats given 4-dimethyl-aminoazobenzene. *Biochem. J.* **103**:316-321.

Ketterer, B., Tipping, E., Beale, D. and Meuwissen, J.A.T.P.(1976) Ligandin, glutathione and carcinogen binding. *In:* I.M. Arias and W.B. Jakoby. (Eds.). *Glutathione: Metabolism and function.* pp243-258, Raven Press, New York.

Kimura, M. (1980) A simple method for estimating evolutionary rates for base substitutions through comparitive studies of nucleotide sequences. *J. Mol. Evol.* **16**:111-120.

Kimura, M. (1983) *The Neutral Theory of Evolution.* Cambridge University Press.

Klapheck, S., Latus, c. and Bergmann, L. (1987) *J. Plant Physiol.* **131**:123-131.

Klone, A., Hussnatter, R. and Sies, H. (1992) Cloning sequencing and characterization of the human alpha glutathione S-transferase gene corresponding to the cDNA clone pGTH2. *Biochem. J.* **285**:925-928.

Koch, A.L. (1972). Enzyme evolution. I. The importance of untranslatable intermediates. *Genetics* **72**:297-316.

Kolm, R.H., Sroga, G.E. and Mannervik, B. (1992). Participation of the phenolic hydroxyl group of Tyr-8 in the catalytic mechanism of human glutathione transferase P1-1. *Biochem.J.* **285**:537-540.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. *Nature* **304**:412-417.

Kuzmich, S., Vanderveer, L.A. and Tew, K.D. (1991) Evidence for a glycoconjugate form of glutathione S-transferase pi. *Int. J. Pep. Protein Res.* **37**:565-.

La Roche, S.D. and Leisinger, T. (1990). Sequence analysis and expression of the bacterial dichloromethane dehalogenase structural gene, a member of the glutathione S-transferase supergene family. *J.Bacteriol.* **172**:164-171.

Ladenstein,R., Epp,O., Bartels,K., Jones,A., Huber,R. and Wendel,A. (1979). Structure analysis and molecular model of the selenoenzyme glutathione peroxidase at 2.8 A resolution. *J.Mol.Biol.* **134**:199-218.

Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**:680-685.

Lai, H-C. J., Qian, B., Grove, G. and Tu. C-P.D. (1988) Gene expression of rat glutathione S-transferases. Evidence for gene conversion in the evolution of the Yb multigene family. *J. Biol. Chem.* **263**:11389-11395.

Lai, H.-C.J., Grove, G., and Tu, C-P.D. (1986) Cloning and sequence analysis of a cDNA for a rat liver glutathione S-transferase Yb subunit. *Nucl. Acids Res.* **14**:6101-6114.

Laisney, V., Van Cong, N., Cross, M.S. and Frezel, J. (1984) Human genes for glutathione S-tranferases. *Human Genet.* **68**:221-227.

Lake, J.A. (1991). Tracing origins with molecular sequences: metazoan and eukaryotic beginnings. *Trends.Biochem.Sci.* **16**:46-50.

Lamoureux, G.L. and Frear, D.S. (1979) Pesticide metabolism in higher plants: *In vitro* enzyme studies. *ACS Symp. Ser.* **97**:77-128.

Lau, E.P., Niswander,L., Watson, D. and Fall, R.R. (1980) Detection of microbial glutathione S-transferases. *Chemosphere* **9**:565-569.

Laverack, M.S. and Dando, J. (1979) *Lecture Notes on Invertebrate Zoology*. 2nd ed. Blackwell Scientific public. Oxford.

Lawrence, R.A. and Burk, R.F. (1976). Glutathione peroxidase activity in selenium-deficient rat liver. *Biochem.Biophys.Res.Commun.* **71**:952-958.

Lennox, E. S. (1955) Transduction of linked genetic characters of the host by bacteriophage P1. Virology 1:190-206.

Lesk, A.M. and Clothia, C. (1980) How different amino acid sequences determine similar protein strucures: The structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**:225-270.

Lesk, A.M., and Boswell, D.R. (1992) Does protein structure determine amino acid sequence?. *Bioessays* **14**: 407-410.

Li, W.-H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**:96.

Li, W.-H., Gouy, M., Sharp, P.M., O'Huigin, C., and Yang, Y.-W. (1990) Molecular phylogeny of rodentia, lagomorpha, primates, artiodactyla, and carnivora and molecular clocks. *Proc. Natl. Acad. Sci. USA.* **87**:6703-6707.

Li, W.-H., Wu, C.-I., and Luo, C.-C. (1985) A new method for estimating synonymous and nonsynonymous rates in nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* **2**:150-174.

Listowsky, I., Abramovitz, M., Homma, H. and Niitsu, Y. (1988) Intracellular binding and transport of hormones and xenobiotics by glutathione-S-transferases. *Drug Metab. Rev.* **19**:305-312.

Listowsky, I., Takahashi, Y., Ishigaki, S., Abramovitz, M. and Homma, H. (1990) Extrahepatic glutathione S-transferases: binding and hormonal regulation. In: *Glutathione S-Transferases and Drug Resistance.* Hayes, J.D., Pickett, C.B., and Mantle, T.J. (Eds). Taylor and Francis, London.

Litwack, G., Ketterer, B., and Arias, I.M. (1971) Ligandin: a hepatic protein which binds steroids, bilirubin, carcinogens and a number of organic anions. *Nature* **234**:466-467.

Liu, L-F. and Tam, M.F. (1991) Nucleotide sequence of a class μ glutathione S-transferase from chicken liver. *Biochim. Biophys Acta* **1090**: 343-344.

Liu, S., Zhang, P., Ji, X., Johnson, W.W., Gilliland, G.L. and Armstrong, R.N. (1992). Contribution of tyrosine 6 to the catalytic mechanism of isoenzyme 3-3 of glutathione S-transferase. *J.Biol.Chem.* **267**:4296-4299.

Liu,L.F., Wu,S.H. and Tam,M.F. (1993). Nucleotide sequence of class-alpha glutathione S-transferases from chicken liver. *Biochim.Biophys.Acta* **1216**:332-334.

Lowry, O.H., Rosebrough, N.J., Farr, A.L. and Randall, R.J. (1951) Protein measurement with folin phenol reagent. *J. Biol.Chem.* **193**:265-275.

MacIntyre, R.J. and O'Brien, S.J. (1976) Interacting gene-enzyme systems in Drosophila. *Annu.Rev.Genet.* **10**:281-318.

Maniatis, T. (1991) Mechanisms of alternative pre-mRNA splicing. *Science* **251**:33-34.

Mannervik, B. (1985) The isozymes of glutathione S-transferase. *Adv. Enzymol. Rel. Areas Mol. Biol.* **57**:357-417.

Mannervik, B. (1986) Glutathione and the evolution of enzymes for detoxication of products of oxidative metabolism. *Chemica Scripta 26B*, 281-284

Mannervik, B. and Danielson, U.H. (1988) Glutathione S-transferases - structure and catalytic activity. *CRC Critic. Reviews in Biochem.* **23**:283-337.

Mannervik, B. and Guthenberg, C. (1981) Glutathione transferase (human placenta). *Methods Enzymol.* **77**:231-235.

Mannervik, B., Alin, P., Guthenberg, C., Jensson, H., Tahir, M.K.,Warholm, M., and Jornvall, H. (1985) Identification of three classes of cytosolic glutathione transferase common to several mammalian species: correlation between structural data and enzymatic properties. *Proc. Natl. Acad. Sci. USA.* **82**:7202-7206.

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., Pickett, C.B., Sato, K., Tu, C.-P. D., Widersten, M. and Wolf, C.R. (1992) Nomenclature for human glutathione transferases. *Biochem. J.* **282**:305.

Mannervik, B., Board, P.G., Berhane, K., Bjornestedt, R., Castro, V.M., Danielson, U.H., Hao, X.-Y., Kolm, R., Olin, B., Principato, G.B., Ridderstrom, M., Stenberg, G., and Widersten, M. (1990) Classes of glutathione S-transferases: Structural and catalytic properties of the enzymes. In: *Glutathione S-Transferases and Drug Resistance* Hayes, J.D., Pickett, C.B., and Mantle, T.J. (Eds) pp35-45, Taylor and Francis, London.

Mannervik, B., Gutenberg, C., Jensson, H., Tahir, M.K., Warholm, M., and Ålin, P. (1984a) Species and tissue differences in the occurrence of isozymes of glutathione S-transferase. *In*: Paton, W., Mitchell, J., & Turner, P. (Eds), *Proc. IUPHAR IXth Intern. Congr. Pharmacol.* **3**:225 Macmillan Press, London.

Mannervik, B., Guthenberg, C., Jensson, H., Warholm, M., and Alin, P. (1983) Isozymes of glutathione S-transferases in rat and human tissues. In: *Functions of Glutathione: Biochemical, Physiological, Toxicological, and Clinical Aspects* Larsson, A., Orrenius, S., Holmgren, A., and Mannervik, B. (Eds) pp75-88, Raven Press, New York.

Mannervik, B., Jensson, H., Alin, P., Orning, L. and Hammarstrom, S. (1984b) Transformation of leukotriene A_4 methyl ester to leukotriene C_4 monoethyl ester by cytosolic rat glutathione transferases. *FEBS Lett.* **174**:289-293.

Marcus, C.J., Habig, W.H. and Jakoby, W.B. (1978). Glutathione transferase from human erythrocytes. Nonidentity with the enzymes from liver. *Arch.Biochem.Biophys.* **188**:287-293.

Martin, A.P. and Palumbi, S.R. (1993) Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. U.S.A.* **90**:4087-4091.

Maruyama, H. and Listowsky, I. 1(984) Preferential binding of steroids by anionic forms of rat glutathione S-transferase. *J. Biol. Chem.* **259**:12449-12555.

McKeown, M. (1992). Alternative mRNA splicing. *Annu. Rev. Cell Biol.* **8**, 133-155.

McLellan, L.I. and Hayes, J.D. (1987) Sex-specific constitutive expression of the pre-neoplastic marker glutathione transferase Y_fY_f, in mouse liver. *Biochem. J.* **245**:399-406.

McLellan, L.I., Kerr, L.A., Cronshaw, A.D. and Hayes, J.D. (1991) Regulation of mouse glutathione S-transferases by chemoprotectors. Evidence for the existence of three distinct alpha-class glutathione S-transferase subunits Ya1, Ya2 and Ya3, in mouse liver. *Biochem. J.* **276**:461-469.

Meister, A. (1983) Selective modification of glutathione metabolism. *Science* **220**:472-477.

Messing, J. (1983) New M13 vectors for cloning. *Methods in Enzymology* **101**:28-78.

Messing, J. and Viera, J. (1982) A new pair of M13 vectors for selecting either strand of a double-digest restriction fragment. *Gene* **19**:20-79.

Metzenberg, A.B., Wurzer, G., Huisman, T.H. and Smithies, O. (1991). Homology requirements for unequal crossing over in humans. *Genetics* **128**:143-161.

Meyer, D.J., Christodoulides, L.G., Hong Tan, K., and Ketterer, B. (1984) Isolation, properties and tissue distribution of rat glutathione transferase E. *FEBS Lett.* **173**:327-330

Meyer, D.J., Coles, B., Pemble, S. E., Gilmore, K.S., Fraser, G.M. and Ketterer, B. (1991). Theta, a new class of glutathione transferases purified from rat and man. *Biochem J.* **274**:409-414.

Miller, J.H. (1972) *Experiments in Molecular Genetics*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Miyata, T., Yasunaga, T. and Nishida, T. (1980). Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc.Natl.Acad.Sci.U.S.A.* **77**:7328-7332.

Moore, R.E., Davies, M.S., O'Connell, K.M., Harding, E.I., Wiegand, R.C., and Tiemeier, D.C. (1986) Cloning and expression of a cDNA encoding a maize glutathione S-transferase in *E. coli*. *Nucl. Acids Res.* **14**:7227-7235.

Morgenstern, R. and DePierre, J.W. (1983) Microsomal glutathione S-transferase. Purification in unactivated form and further characterization of the activation process, substrate specificity and amino acid composition. *Eur. J. Biochem. 134*, 591-597

Morgenstern, R. and DePierre, J.W. (1988) Glutathione conjugation; its mechanism and biological significance. In: *Glutathione Conjugation in Mechanisms and Biological Significance* Sies, H. and Ketterer, B. (Eds). Academic Press, London.

Morgenstern, R. Lunqvist, G. Mosialou, E. and Andersson, C., (1990) Membrane-bound glutathione transferase: function and properties. *In*: Hayes, J.D., Pickett, C.B., and Mantle, T.J. (eds) *Glutathione S-Transferases and Drug Resistance* pp35-45. Taylor and Francis, London.

Morgenstern, R., De Pierre, J.W., and Jörnvall, H. (1985) Microsomal glutathione transferase: primary structure. *J. Biol. Chem. 260*, 13976-13983

Morgenstern, R., J. Meeijer, J.W. Depierre, L. Ernster. (1980) Characterisation of rat liver microsomal glutathione S-transferase activity. *Eur. J. Biochem*. **104**:167-172.

Morrow, C.S., Chiu, J. & Cowen, K.H. (1992). Posttranscriptional control of glutathione S-transferase p gene expression in human breast cancer cells *J. Biol. Chem.* **267**:10544-10550.

Morton, M.R., Bayney, R.M. and Pickett, C.B. (1990). Isolation and characterization of the rat glutathione S- transferase Yb1 subunit gene. *Arch.Biochem.Biophys.* **277**:56-60.

Moscow, J.A. and Cowan, K.H. (1988a) Multidrug resistance. *J. Nat. Canc. Inst.* **80**:14-20.

Moscow, J.A., Townsend, A.J., Goldsmith, M.E., Whang-Peng, J., Vickers, P.J., Poisson, R., Legault-Poisson, S., and Myers, C.E. (1988b) Isolation of the human anionic glutathione S-transferase cDNA and the relation of its gene expression to estrogen-receptor content in primary breast cancer. *Proc. Natl. Acad. Sci. USA.* **85**:6518-6522.

Moscow, J.A.,Fairchild,C.R., Madden, M.J., Ransom,D.T., Wieand, H.S., O'Brien, E.E., Poplack, D.G., Cossman, J., Myers, C.E. and Cowen, K.Y. (1989) Expression of anionic glutathione S-transferase and P glycoprotein genes in human tissues and tumours. *Cancer Res.*, **49**:1322-1428.

Mozer, T.J., Tiemeier, D.C. and Jaworski, E.G. (1983). Purification and characterization of corn glutathione S- transferase. *Biochemistry* **22**:1068-1072.

Muramatsu, M., Okuda, A., Kano, T., and Sakai, M. (1987) Structure and regulation of rat glutathione S-transferase P (GST P) gene. In: Mantle, T.J., Pickett, C.B. and Hayes, J.D. (eds) *Glutathione S-transferases and Carcinogenesis*. p111- Taylor and Francis, London.

Myers, E.W., and Miller, W. (1988) Optimal alignments in space. *Cabios* **4**:11-17.

Nebert, D. (1990). Growth signal pathways. *Nature* **347:**709-710.

Norrander, J., Kempt, T., and Messing, J. (1983). Construction of improved M13 vectors using oligonucleotide-directed mutagenesis. *Gene* **26**:101-106.

O'hUigin, C. and Li, W.H. (1992). The molecular clock ticks regularly in muroid rodents and hamsters. *J.Mol.Evol.* **35**:377-384.

Ogura, K., Nishiyama, T., Okada, T., Kajital, J., Narihata, H., Watabe, T., Hiratsuka, A. and Watabe, T. (1991) Molecular cloning and amino acid sequencing of rat liver class theta glutathione S-transferase Yrs- Yrs inactivating reactive sulphate esters of carcinogenic arylmethanols. *Biochem. Biophys. Res. Comm*. **181**:1294-1300.

Ohno, S. (1970) *Evolution by Gene Duplication*. Springer-Verlag, New York.

Okuda, A., Imagawa, M., Sakai, M. & Muramatsu, M. (1990). Functional cooperativity between two TPA responsive elements in undifferentiated F9 embryonic stem cells. *EMBO J.* **9**:1131-1343.

Okuda, A., Sakai, M., and Muramatsu, M. (1987) The structure of the rat glutathione S-transferase P gene and related pseudogenes. *J. Biol. Chem.* **262**:3858-3863.

Ollis, D.L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S.M., Harel, M., Remington, S.J., Silman, I., Schrag, J. and et al, (1992). The alpha/beta hydrolase fold. *Protein.Eng.* **5**:197-211.

Overbaugh, J.M. and Fall, R. (1982) *FEMS Microbiol. Lett.* **13**:371-375.

Overbaugh, J.M., Lau, E.P., Marino, V.A., and Fall, R. (1988) Purification and preliminary characterisation of a monomeric glutathione S-transferase from *Tetrahymena thermophila. Arch. Biochem. Biophys.* **261**:227-234.

Pacifici, G.M., Warholm, M., Guthenberg, C., Mannervik, B., and Rane, A. 1986. Organ distribution of glutathione transferase isoenzymes in the human fetus: differences between liver and extrahepatic tissues. *Biochem. Pharmacol.* **35**: 1616-1619.

Pardue, M.L. (1985) *In situ* hyridization. *In*: Hames, B.D. and Higgins, S.J. (Eds), *Nucleic Acid Hybridisation-A Practical Approach.* pp179-202 IRL Press Oxford, Washington: .

Paulson, K.E., Darnell, J.E., Rushmore, T., and Pickett, C.B. (1990) Analysis of the upstream elements of the xenobiotic compound-inducible and positionally regulated glutathione S-transferase Ya gene. *Mol. Cell. Biol.* **10**:1841-1852.

Pearson, W.R., Reinhart, J., Sisk, S.C., and Anderson, K.S. (1988) Tissue-specific induction of murine glutathione transferase mRNAs by butylated hydroxyanisole. *J. Biol. Chem.* **263**:13324-13332.

Pearson, W.R., Windle, J.J. and Morrow, J.F. (1983) Increased synthesis of glutathione S-transferases in response to anticarcinogenic antioxidants. *J. Biol. Chem.* **258**:2052-2062.

Pearson, W.R., Xu, S.-J., Vorachek, W.R.,and Patterson, D. (1993) The organisation of human class-mu glutathione transferase genes GSTM1- GSTM5 on chromosome 1p13. In: *Structure and Function of Glutathione S-transferases*. Tew, K.D., Pickett, C.B., Mantle, T.J. and Hayes, J.D. (Eds.) CRC Press, Boca Raton, Florida.

Pemble, S.E. and Taylor, J.B. (1992). An evolutionary perspective on glutathione transferases inferred from class-theta glutathione transferase cDNA sequences. *Biochem.J.* **287**:957-963.

Pemble, S.E., Taylor, J.B., and Ketterer, B. (1986) Tissue distribution of rat glutathione transferase subunit 7, a hepatoma marker. *Biochem. J.* **240**:885-889.

Penrose, J.F., Gagnon, L., Goppelt-Struebe, M., Myers, P., Lam, B.K., Jack, R.M., Austen, K.F. and Soberman, R.J. (1992) Purification of human leukotriene C_4 synthase. *Proc. Natl. Acad. Sci. USA* **89**:11603-11606.

Persson, B., Jornvall, H., Alin, P., and Mannervik, B. (1988) Structural classes of glutathione transferase: Distinctions between isozymes and enzymes. *Prot. Seq. Data Anal.* **1**:183-186.

Peterson, G.L. (1977) A simplification of the protein assay method of Lowry et al. which is more generally applicable. *Anal. Biochem.* **83**: 346-356.

Porter, S., and Mintz, B. (1991) Multiple alternatively spliced transcripts of the mouse tyrosinase-encoding gene. *Gene* **97**: 277-282.

Proudfoot, N.J. and Maniatis, T. (1980). The structure of a human alpha-globin pseudogene and its relationship to alpha-globin gene duplication. *Cell* **21**:537-544.

Ramsdell, H.S. and Eaton, D.L. (1990) Mouse liver glutathione S-transferase isoenyme activity toward aflatoxin B1-8,9-epoxide and benzo(a)pyrene-7,8-dihydrodiol-9,10-epoxide. *Toxicol. Appl. Pharmacol.* **105**:216-225.

Redick, J.A., Jakoby, W.B. and Baron, J. (1982) Immunohistochemical localization of glutathione S-transferases in livers of untreated rats. *J. Biol. Chem.* **257**:15200-15203.

Reed, K.C. and Mann, D.A. (1985) Rapid transfer of DNA from agarose gels to nylon membranes. *Nucl. Acids Res.* **13**:7207-7221.

Reinemer, P., Dirr, H., Laderstein, R., Huber, R., Lo Bello, M., Federici, G. and Parker, M.W. (1992) Three-dimentional structure of class pi glutathione S-transferase from human placenta in complex with S-hexylglutathione at 2.8 A resolution. *J. Mol. Biol.* **227**:214-226.

Reinemer, P., Dirr, H.W., Laderstein, R., Schaffer, J., Galley, O. and Huber, R. (1991) The three-dimensional structure of class pi glutathione S-transferase in complex with glutathione sulfonate at 2.3 A resolution. *EMBO J.* **10**:1997-2005.

Rhoads, D.M., Zarlengo, R.P., and Tu, C-P.D. (1987) The basic glutathione S-transferases from human livers are products of separate genes. *Biochem. Biophys. Res. Commun.* **145**:474-481.

Richardson, C.C. (1965) Phosphorylation of nucleic acid by an enzyme from T4 bacteriophage-infected *Escherichia coli. Proc. Natl. Acad. Sci. USA* **54**:158-165.

Roberts,R.G., Bentley,D.R. and Bobrow,M. (1993). Infidelity in the structure of ectopic transcripts: a novel exon in lymphocyte dystrophin transcripts. *Hum.Mutat.* **2**:293-299.

Rohrdanz, E., Nguyen, T. and Pickett, C.B. (1992) Isolation and characterisation of the human glutathione S-transferase A2 subunit gene. *Arch. Biochem. Biophys.* **298**:747-752.

Rosenthal and Janzen,D.H. (1979) *Herbivores: Their Interaction with Secondary Plant Metabolites.* Academic Press New York.

Ross, V.L. and Board, P.G. (1993) Molecular cloning and heterologous expression of an alternatively spliced human Mu class glutathione S-transferase. *Biochem. J.* **294**:373-380.

Rost, B., Snider, R. and Sander, C. (1992) Progress in protein structure prediction? *Trends in Biochem. Sci.* **18**:120-123.

Rothkopf, G.S., Telakowski-Hopkins, C.A., Stotish, R.L. and Pickett, C.B. (1986) Multiplicity of glutathione S-transferase genes in the rat and association with a type 2 Alu repetitive element. *Biochemistry* **25**:993-1002.

Rushmore, T.H. and Pickett, C.B. (1990) Transcriptional regulation of the rat glutathione S-transferase Ya subunit gene. *J. Biol. Chem.* **265**:14648-14653.

Rushmore, T.H., King, R.G., Paulson, K.E., and Pickett, C.B. (1990) Regulation of glutathione S-transferase Ya subunit gene expression: Identification of a unique xenobiotic-responsive element controlling inducible expression by planar aromatic compounds. *Proc. Nat. Acad. Sci. USA. 87*, 3826-3830

Saito, K., Yamazoe, Y., Kamataki, T. and Kato, R., (1983) Activation and detoxication of *N*-hydroxy-Trp-P-2 by glutathione and glutathione transferases, *Carcinogenesis (Lond.)* **4**:1551-1557.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol.Biol.Evol.* **4**:406-425.

Sakai, M., Okuda, A., Nishi, S., and Muramatsu, M. (1990) Regulation of glutathione transferase P (GST-P) gene expression during chemical hepatogenesis. *Prog. Clin. Biol. Res.* **344**:123-138.

Sambrook, J., Frisch, E.F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* 2nd edition. Cold Spring Harbor Laboratory, Cold Spring Harbour, N.Y.

Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H.and Roe, B.A. (1980). Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J. Mol. Biol.* **143**, 161-178.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors *Proc. Nat. Acad. Sci. USA.* **74**:5463-5467.

Satoh, K., Kitahara, A., Soma, Y., Inaba, Y., Hatayama, I. and Sato, K. (1985). Purification, induction, and distribution of placental glutathione transferase: a new marker enzyme for preneoplastic cells in the rat chemical hepatocarcinogenesis. *Proc.Natl.Acad.Sci.U.S.A.* **82**:3964-3968.

Scott, E.M. and Wright, R.C. (1980) Variability of glutathione S-transferase of human erythrocytes. *Am. J. Hum. Genet.* **32**:115-116.

Seidegård, J. & Pero, R.W. (1985). The hereditary transmission of high glutathione transferase activity towards trans-stilbene oxide in human mononuclear leukocytes. *Hum. Genet.* **69**, 66-68.

Seidegård, J., De Pierre, J.W., Birberg, W., Pilotti, A., and Pero, R.W. (1984) Characterization of soluble glutathione transferase activity in resting mononuclear leukocytes from human blood. *Biochem. Pharmacol.* **33**: 3053-3058.

Seidegård, J., Guthenberg, C., Pero, R.W. & Mannervik, B. (1987). the trans-stilbene oxide-active glutathione transferase in human mononuclear leucocytes is identical to hepatic glutathione transferase mu. *Biochem. J.* **246**, 783-785.

Seidegård, J., Pero, R.W., Markowitz, M.M., Roush, G., Miller, D.G. and Beattie, E.J. (1990) Isoenzyme(s) of glutathione transferase (class mu) as a marker for the susceptibility to lung cancer: a follow up study. *Carcinogenesis (Lond.).* **11**:33-36.

Seidegård, J., Pero, R.W., Miller, D.G., and Beattie. (1986) A glutathione transferase in human leukocytes as a marker for the susceptibility to lung cancer. *Carcinogenesis (Lond.)* **7**:751-753.

Seidegård, J., Vorachek, W.R., Pero, R.W., and Pearson, W.R. (1988) Hereditary differences in the expression of the human glutathione transferase active on trans-stilbene oxide are due to a gene deletion. *Proc. Natl. Acad. Sci. USA.* **85**:7293-7297.

Shah, D.M., Hironaka, C.M., Wiegand, R.C., Harding, E.I., Krivi, G.G., and Tiemeier, D.C. (1986) Structural analysis of a maize gene coding for glutathione S-transferase involved in herbicide detoxification. *Plant Mol. Biol.* **6**:203-211.

Shamaan, N.A., Hamidah, R., Jeffries, J., Hashim, A.J. and Wan Ngah, W.Z. (1993). Insecticide toxicity, glutathione transferases and carboxylesterase activities in the larva of the Aedes mosquito. *Comp.Biochem.Physiol.C.* **104**:107-110.

Sharma, R., Gupta, S., Singhal, S.S., Ahmad, H., Haque, A., and Awasthi, Y.C. (1991). Independent segregation of glutathione S-transferase and fatty acid ethyl ester synthase from pancreas and other human tissues. *Biochem. J.* **275**: 507-513.

Shea, T.C., Claflin, G., Comstock, K.E., Sanderson, B.J.S., Burstein, N. A., Keenan, E.J., Mannervik, B. and Henner, W.D. (1990) Glutathione transferase activity and isoenzyme composition in primary human breast cancers. *Cancer Res.* **50**:6848-6853.

Shea, T.C., Kelly, S.L., and Henner, W.D. (1988) Identification of an anionic form of glutathione transferase present in many human tumours and human tumour cell lines. *Cancer Res.* **49**:527-533.

Simons, P.C. and Vander Jagt, D.L. (1980) Bilirubin binding to human liver ligandin (glutathione S-transferase). *J. Biol. Chem.* **255**:4740-4744.

Singh, S.V., Ahmad, H., Kurosky, A. and Awasthi, Y.C. (1988) Purification and characterization of unique glutathione S-transferases from human muscle. *Arch.Biochem.Biophys.* **264**:13-22.

Singh, S.V., Kurosky, A. and Awasthi, Y.C. (1987) Human liver glutathione S-transferase psi. Chemical characterization and secondary-structure comparison with other mammalian glutathione S-transferases. *Biochem.J.* **243**:61-67.

Singhal, S.S., Ahmad, H., Sharma, R., Gupta, S., Haque, A.K. & Awasthi, Y.C. (1991). Purification and characterisation of human muscle glutathione S-transferases: evidence that ζ corresponds to a locus distinct from GST1, GST2, and GST3. *Arch. Biochem. Biophys.* **285**,64-73.

Sinning,I., Kleywegt,G.J., Cowan,S.W., Reinemer,P., Dirr,H.W., Huber,R., Gilliland,G.L., Armstrong,R.N., Ji,X., Board,P.G. and et al, (1993) Structure determination and refinement of human alpha class glutathione transferase A1-1, and a comparison with the Mu and Pi class enzymes. *J.Mol.Biol.* **232**:192-212.

Slater, T.F., (1984) Free radical mechanisms in tissue injury. *Biochem.J.* **222**:1-15.

Smith, C.W., Patton, J.G., and Nadal-Ginard, B. (1989) Alternative splicing in the control of gene expression. *Annu. Rev. Genet.* **23**:527-577.

Smith, D.B., Davern, K.M., Board, P.G., Tiu, W.U., Garcia, E.G., and (1986) Mr 26,000 antigen of schistosoma japonicum recognised by resistant WEHI 129/J mice is a parasite glutathione S-transferase. *Proc. Nat. Acad. Sci. USA* **83**:8703-8707.

Smith, M.T., Evans, C.G., Doane-Setzer, P., Castro, V.M., Tahir, M.K. and Mannervik, B. (1989) Denitrosation of 1, 3-bis(2-chloroethyl)-1-nitrosourea by class Mu glutathione transferases and its role in cellular resistance in rat brain tumour cells. *Cancer Res.* **49**:2621-2625.

Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical taxonomy*. Freeman, San Franscisco.

Söderström. M., Hammarstrom, S. and Mannervik, B. (1988) Leukotriene C synthase in mouse mastocytoma cells. An enzyme distinct from cytosolic and microsomal glutathione transferases. *Biochem. J.* **250**:713-718.

Soma, Y., Satoh, K., and Sato, K. (1986) Purification and subunit-structural and immunological characterization of five glutathione S-transferases in human liver, and the acidic form as a hepatic tumor marker. *Biochim. Biophys. Acta* **869**: 247-258.

Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**: 503-517.

Stachelek, J.L. and Liskay, R.M. (1988). Accuracy of intrachromosomal gene conversion in mouse cells. *Nucleic Acids Res.* **16**:4069-4076.

Staden, R. (1984) Computer methods to aid the determination and analysis of DNA sequences. *Biochem. Soc. Trans.* **12**: 1005-1008.

Stenberg, G., Board, P.G. and Mannervik, B. (1991). Mutation of an evolutionarily conserved tyrosine residue in the active site of a human class Alpha glutathione transferase. *FEBS Lett.* **293**:153-155.

Stenberg, G., Ridderstrom, M., Engstrom, A., Pemble, S.E. and Mannervik, B. (1992). Cloning and heterologous expression of cDNA encoding class alpha rat glutathione transferase 8-8, an enzyme with high catalytic activity towards genotoxic alpha,beta-unsaturated carbonyl compounds. *Biochem.J.* **284**:313-319.

Strange, R.C., Faulder, C.G., Davis, B.A., Hume, R., Brown, J.A.H., Cotton, W., and Hopkinson, D.A. (1984) The human glutathione S-transferases: studies on the distribution and genetic variation of GST-1, GST-2, and GST-3 isozymes. *Ann. Hum. Genet.* **48**:11-20.

Strange, R.C., Fryer, A.A., Hiley, C., Bell, J., Cossar, D., and Hume, R. (1990) Developmental expression of GST in human tissues. In: Hayes, J.D., Pickett, C.B., and Mantle, T.J. (eds) *Glutathione S-Transferases and Drug Resistance* pp35-45, Taylor and Francis, London.

Strange, R.C., Johnson, P.h., Lawton, A., Moult, J.A., Tector, M.J., Tyminski, R.J. and Cotton, W. (1982) Studies on the variability of glutathione S-transferase from human erythrocytes. *Clin. Chim. Acta* **120**:251-60.

Strange, R.C., Matharoo, B., Faulder, G. C., Jones, Cotton, W., Elder, J. B., Deakin, M. (1991) The human glutathione S-transferases: a case-control study of the

Taylor, J.B., Vidal, H., Torpier, G., Meyer, D.J., Roitsch, C., Balloul, J.M., Southan, C., Sondermeyer,P., Pemble, S., Lecocq, J.P., *et al.* (1988). The glutathione transferase activity and tissue distribution of a cloned Mr28K protective antigen of *Schistosoma mansoni*. *EMBO J.* **7**:465-472.

incidence of the GST1 0 phenotype in patients with adenocarcinoma. *Carcinogenisis* **12**:25-28.

Strange,R.C., Hirrell,P.H., Kitley,G.A., Hopkinson,D.A.andCotton,W. (1983). Erythrocyte glutathione S-transferase. Electrophoretic identification of two enzyme forms. *Biochem.J.* **215**:213-216.

Strange,R.C., Johnston,J.D., Coghill,D.R.andHume,R. (1980). A comparison of erythrocyte glutathione S-transferase activity from human foetuses and adults. *Biochem.J.* **188**:475-479.

Suguoka, Y., Kano, T., Okuda, A., Sakai, M., Kitagawa, T. and Muramatsu, M. (1985) Cloning and the nucleotide sequence of rat glutathione S-transferase P cDNA. *Nucl. Acids Res.* **13**:6049-6057

Suzuki, T., and Board, P.G. (1984) The gene for glutathione S-transferase 3 (GST3) is on chromosome 11. *Som. Cell Gen.* **10**:319-320.

Suzuki, T., Coggan, M., Shaw, D.C. and Board, P.G. (1987) Electrophoretic and immunological analysis of human glutathione S-transferase isozymes. *Ann. Hum. Genet.* **51**:95-106.

Suzuki, T., Johnston, P.N. and Board, P.G. (1993) Structure and organisation of the human alpha class glutathione S-transferase genes and related pseudogenes. *Genomics* **18**:680-686.

Suzuki, T., Kovacs, M., and Board, P.G. (1990) Do the major human glutathione S-transferases have fatty acid ethylester synthase activity? *FEBS Lett.* **275**: 58-60.

Suzuki, T., Shaw, D.C. & Board, P.G. (1991). Purification of acidic glutathione S-transferase 6 from human brain. *Biochem. J.* **274**, 405-408.

Tahir, M.K. and Mannervik, B. (1986) Simple inhibition studies for distinction between homodimeric and heterodimeric isoenzymes of glutathione transferase. *J. Biol. Chem.* **261**:1048-1051.

Takahashi, Y. and Nagata, T. (1992). parB: an auxin-regulated gene encoding glutathione S-transferase. *Proc.Natl.Acad.Sci.U.S.A.* **89**:56-59.

Takahashi, Y., Campbell, E.A., Hirata, Y., Takayama, T., and Listowsky, I. (1993) A basis for differentiating among the multiple human Mu- glutathione S-transferases and molecular cloning of brain GSTM5. *J. Biol. Chem.* **268**: 8893-8898.

Tamaki, H., Kumagai,H. and Tochikura. T. (1991) Nucleotide sequence of the yeast glutathione S-transferase cDNA. *Biochim. Biophys. Acta* **1089**:276-279.

Tan, K.H., Meyer, D.J., Gillies, N., and Ketterer, B. (1988) Detoxification of DNA hydroperoxide by glutathione transferases and the purification and characterization of glutathione transferases of the rat liver nucleus. *Biochem. J.* **254**:841-845.

Taniguchi, H. and Pyerin, W. (1989). Glutathione S-transferase is an *in vitro* substrate of Ca++- phospholipid-dependent protein kinase (protein kinase C). *Biochem.Biophys.Res.Commun.* **162**:903-907.

Taylor, J.B., Oliver, J., Pemble, S.E., and Ketterer, B. (1990) Structural relationships of human, class mu glutathione S-transferase genes: evidence for gene conversion and for a sub-division of the class. In *Glutathione S-Transferases and Drug Resistance* Hayes, J.D., Pickett, C.B., and Mantle, T.J. (Eds), pp35-45, Taylor and Francis, London.

Taylor, J.B., Oliver, J., Sherrington, R. and Pemble, S.E. (1991) Structure of human glutathione S-transferase class Mu genes. *Biochem. J.* **274**:587-593.

Telakowski-Hopkins, C.A., King, R.G. and Pickett, C.B. (1988) Glutathione S-transferase Ya subunit gene: Identification of regulatory elements required for basal level and inducible expression. *Proc. Natl. Acad. Sci. USA.* **85**:1000-1005.

Telakowski-Hopkins, C.A., Rothkopf, G.S., and Pickett, C.B. (1986) Structural analysis of a rat liver glutathione S-transferase Ya gene. *Proc. Natl. Acad. Sci. USA* **83**:9393-9397.

Tew, K.D., Schisselbauer, J.C. Clapper, M.L. and Kuzmich, S. (1990). IN: Hayes, J.D., Pickett, C.B., and Mantle, T.J. (eds) *Glutathione S-Transferases and Drug Resistance* pp309-318. Taylor and Francis, London.

Tipping, E., and Ketterer, B. (1978) The role of intracellular proteins in the transport and metabolism of lipophilic compounds. In: Blauer, G. and Sund, H (eds) *Transport by Proteins*. p369, Walter de Gruyter & Co., Berlin.

Tipping, E., and Ketterer, B. (1981) The role of slouble binding proteins on lipophilic transport and metabolism in hepatocytes. *Biochem. J.* **195**:441-452.

Tomarev, S.I. and Zinovieva, R.D. (1988) Squid major lens polypeptides are homologous to glutathione S-transferase subunits. *Nature* **336**:86-88.

Tomarev, S.I., Zinovieva, R.D. and Piatigorsky, J. (1991) Crystallins of the octopus lens. Recruitment from detoxification enzymes. *J.Biol.Chem.* **266**:24226-24231.

Tomarev, S.I., Zinovieva, R.D. and Piatigorsky, J. (1992) Characterization of squid crystallin genes. Comparison with mammalian glutathione S-transferase genes. *J.Biol.Chem.* **267**:8604-8612.

Toung, Y.-P.S., Hsieh, T-S., and Tu, C-P.D. (1990) *Drosophila* glutathione S-transferase 1-1 shares a region of sequence homology with the maize glutathione S-transferase III. *Proc. Natl. Acad. Sci. USA* **87**:31-35.

Towbin, H., Staehelin, T. and Gordon, J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc.Natl.Acad.Sci.U.S.A.* **76**:4350-4354.

Tsuchida, S. and Sato, K. (1992). Glutathione transferases and cancer. *CRC Crit. Rev. Biochem. Mol. Biol.* **27**,337-384.

Tsuchida, S. Izumi, T. Shimizu, T., Ishikawa, T., Hatayama, I., Satoh, K.and Sato, K. (1987) Purification of a new acidic glutathione S-transferase, GST-Yn1-Yn1, with a high leucotriene C_4 synthase activity from rat brain. *Eur. J. Biochem.* **170**:159-.

Tsuchida, S. Maki, T. and Sato, K. (1990) Purification and characterization of glutathione transferases with an activity towards nitroglycerin from human aorta and heart. Multiplicity of the human class mu forms. *J. Biol. Chem.* **265**:7150-7157.

Tu, C.-P.D. and Qian, B. (1986) Human liver glutathione S-transferases: Complete primary sequence of an Ha subunit cDNA. *Biochem. Biophys. Res. Commun.* **141**:229-237.

van Bladeren, P.J. (1988) Formation of toxic metabolites from drugs and other xenobiotics by glutathione conjugation. TIPS. **9**:295-299.

van Poppel, G., de Vogel, N. van Bladeren,P.J. and Kok, F.J. (1992) Increased cytogenetic damage in smokers deficient in glutathione S-transferase isozyme m. *Cacinogenisis (Lond.)* **13**:303-305.

Vander Jagt, D.L., Hunsaker, L.A., Garcia, K.B. & Royer, R.E. (1985) Isolation and characterisation of the multiple glutathione S-transferases from human liver. Evidence for unique heme-binding site. *J. Biol. Chem.* **260**:11603-11610.

Vorachek, W.R., Pearson, W.R. and Rule, G. S. (1991) Cloning, expression and characterisation of a class-mu glutathione transferase from human muscle, the product of the *GST4* locus. *Proc. Natl. Acad. Sci.* **88**:4443-4447.

Wang, A.L. and Tew, K.D. (1985) Increased glutathione S-transferase activity in a cell line with acquired resistance to nitrogen mustards. *Canc. Treat. Rep.* **69**:677-82.

Wang, A.L., Teicher, B.A., Shea, T.C., Holden, S.A., and Henner, W.D. (1989) Cross-resistance and glutathione S-transferase-p levels among four melanoma cell lines selected for alkylating agent resistance. *Canc. Res.* **49**:6185-92.

Warholm, M., Guthenberg, C., and Mannervik, B. (1983) Molecular and catalytic properties of glutathione transferase mu from human liver: an enzyme efficiently conjugating epoxides. *Biochem. J. 22*, 3610-3617 or Biochemistry, **22**, 3610-3617.

Wattenberg, L.W., (1978). Inhibitors of chemical carcinogenesis. *Adv. Can. Res.* **26**:197-226.

Warholm, M., Guthenberg, C., Mannervik, B., Pacifici, G.M., and Rane, A. (1981). Glutathione S-transferases in human fetal liver. *Acta Chem. Scand. Ser. B., B35,* 225-227

Whittaker, R.H. and Margulis, L. (1978) Protist classification and the kingdoms of organisms. *Biosystems* **10**:3-18.

Widersten, M., Pearson, W.R., Engstrom, A. and Mannervik, B. (1991) Heterologous expression of the allelic variant mu-class glutathione transferases mu and psi. *Biochem.J.* **276**:519-524.

Wiencke, J.K., Kelsey, K.T., Lamela, R.A and Toscano Jr., W.A., (1990) Human glutathione S-transferase deficiency as a marker of susceptibility to epoxide-induced cytogenetic damage. *Cancer Res.* **50**:1585-1590.

Wilbur, W.J. and Lipman, D.J. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Nat. Acad. Sci. USA.* **80**:726-730.

Woese, C.R. (1987). Bacterial evolution. *Microbiol.Rev.* **51**:221-271.

Wolf, C.R., Wareing, C.J., Black, S.M. and Hayes, J.D. (1990). Glutathione S-transferases and resistance to alkylating agents. *In*: Hayes, J.D., Pickett, C.B., and Mantle, T.J. (Eds), *Glutathione S-Transferases and Drug Resistance* pp295-308. Taylor & Francis, London.

Wolfe,K.H., Sharp,P.M.andLi,W.H. (1989) Mutation rates differ among regions of the mammalian genome. *Nature* **337**:283-285.

Wolkoff, A.W., Goresky, C.A., Sellin, J. Gatmaitan, Z., and Arias, I.M. (1979) Role of ligandin in transfer of bilirubin from plasma into liver. *Am. J. Physiol.* **236**:E638-648.

Wool, D., Noiman, S., Manheim, O. and Cohen, E. (1982). Malathion resistance in Tribolium strains and their hybrids: inheritance patterns and possible enzymatic mechanisms (Coleoptera, Tenebrionidae). *Biochem.Genet.* **20**:621-636.

Wu, C.-I. and Li, W.-H. (1985) Evidence for higher rates of nucleotide substitution in rodents. *Proc. Natl. Acad. Sci. USA. 82,* 1741-1745

Yamamoto, K.R., Alberts, B.M., Benzinger, R., Lawhorne, L. and Treiber, G. (1970) Rapid bacteriophage sedimentation in the presence of polyethylene glycol and its application to large-scale virus purification. *Virology* **40**:734-.

Yanisch-Perron, C., Viera, J. and Messing, J. (1985) Improved M13 phage cloning vectors and host strains:nucleotide sequences of the M13 mp18 and pUC19 vectors. *Gene* **33**:103-119.

Yawetz, A. and Agosin, M., (1981) Purification of the glutathione S-transferase of *Trypanosoma cruzi. Comp. Biochem Physiol* **68B**:237-243.

Yenchitsomanus, P., Summers, K.M., Board, P.G., Bhatia, K.K., Jones, G.L., Johnston, K. and Nurse, G.T. (1986). Alpha-thalassemia in Papua New Guinea. *Hum.Genet.* **74**:432-437.

Yeung, T.C. and Gidari, A.S. (1980). Purification and properties of a chicken liver glutathione S- transferase. *Arch.Biochem.Biophys.* **205**:404-411.

Young,R.A. and Davis,R.W. (1983) Yeast RNA polymerase II genes: Isolation with antibody probes. *Science* **222**:778-.

Zachar, Z., Chou, T.-C., Bingham, P. (1987) Evidence that a regulatory gene auto regulates splicing of its transcript. *EMBO J.* **6**:4095-4104.

Zhang, P.H., and Armstrong, R.N. (1990) Construction, expression, and preliminary characterization of chimeric class mu glutathione S-transferases with altered catalytic properties. *Biopolymers.* **29**: 159-169.

Zhong, S., Wolf, C.R., and Spurr, N.K. (1992) Chromosomal assignment and linkage analysis of the human glutathione S-transferase mu gene (GSTM1) using intron specific polymerase chain reaction. *Hum. Genet.* **90**: 435-439.