IDENTIFICATION OF NOVEL GLUTATHIONE TRANSFERASES AND POLYMORPHIC VARIANTS BY EXPRESSED SEQUENCE TAG DATABASE ANALYSIS


John Curtin School of Medical Research, The Australian National University, Canberra, Australian Capital Territory, Australia (P.G.B., G.C., L.J., N.T., A.C.B.); and Department of Pharmacology and Physiology, University of Rochester Medical Center, Rochester, New York (H.-F.T., M.W.A.)

This paper is available online at http://dmd.aspetjournals.org

ABSTRACT:

The human expressed sequence tag (EST) database can be searched by different sequence alignment strategies to identify new members of gene families and allelic variants. To illustrate the value of database analysis for gene discovery, we have focused on the glutathione S-transferase (GST) superfamily, an approach that has led to the identification of the Zeta class. The Zeta class GSTs catalyze the glutathione-dependent biotransformation of α-haloacids and the isomerization of maleylacetoacetic acid to fumarylacetoacetic acid, an essential step in the catabolism of tyrosine.

Allelic variants of the GST Z1 and GST A2 genes have also been identified by EST database analysis. One GST Z1 variant (GST Z1A) has significantly higher activity with dichloroacetic acid as a substrate than other GST Z1 isoforms. This variant may be important in the clinical treatment of lactic acidosis where dichloroacetic acid is prescribed. Our experience with the application of EST database searching methods suggests that it may be productively applied to other gene families of pharmacogenetic interest.

The human expressed sequence tag (EST)4 database contains in excess of 1.6 million partial cDNA sequences. These sequences have been obtained from over 200 cDNA libraries prepared from most tissues. The EST database therefore contains cDNA sequences from a large proportion of the genes expressed in human tissues. Since the cDNA libraries have been prepared from tissues obtained from different individuals, the database also contains polymorphic variants of many genes. Thus, the application of bioinformatic searching strategies to the vast amount of information contained within the EST database can reveal new members of gene families and polymorphic variants of previously discovered genes. Although this approach could be used for many gene families, it is particularly valuable for pharmacogenetic studies wherein it is important to identify variants that may have unusual capacities for the metabolism of xenobiotics and therapeutic agents.

In several experiments carried out to illustrate the value of EST database analysis, we have focused our attention on the glutathione transferase (GST) gene family (Board et al., 1997; Blackburn et al., 2000). The GSTs are a large family of phase II enzymes that conjugate glutathione to a wide range of generally hydrophobic and electrophilic compounds including many carcinogens, therapeutic drugs, and the products of oxidative metabolism. Genetic factors that modulate GST expression and function can be clinically important. Overexpression of certain GSTs in tumors has been shown to contribute to drug resistance, and genetically determined deficiencies of the GST M1 or GST T1 genes have been shown to be risk factors for several cancers (for review see Chenevix-Trench et al., 1995; Hayes and Pulford, 1995). In addition, homozygosity for the GST M1 null allele is a significant positive prognostic indicator for successful chemotherapy and long-term survival in children with acute lymphoblastic leukemia (Hall et al., 1994). In recent studies, polymorphic variants of GST P1 that influence substrate specificity have been shown to be risk factors for Parkinson’s disease in subjects exposed to pesticides (Zimniak et al., 1994; Menegon et al., 1998).

Our analysis of the EST database has utilized differing strategies to identify new members of the GST super family and allelic variants of particular genes. Although there are limitations and pitfalls with this approach, it has successfully identified the Zeta class of GSTs and allelic variants that significantly influence the function of GST Z1–1. Here we review the application of these techniques to the identification of the Zeta class and the detection of allelic variants in the Alpha and Zeta class GSTs.

Materials and Methods

The BLAST Programs Used in These Studies (Altschul et al., 1998) are available online from the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov) and can be used to efficiently search sequence databases by aligning a query sequence with all entries in the database.

Identification of New GST Classes. A variant of the BLAST program, tblastn, can be used to search translations of nucleotide databases with protein sequence. Because of the degeneracy of the genetic code, searching for homologs across wide evolutionary distances has greater sensitivity if protein sequence is used as the query sequence.

To determine whether GST-like proteins identified in lower species have counterparts in humans, we used the tblastn program to search the human EST database with sequences from several plants and insects. The default parameters for the BLAST search were retained, and the similar ESTs thus identified were studied in further detail to determine whether they represented novel or previously identified GSTs. Subsequently, where there was evidence of multiple clones encoding a previously undescribed sequence, a representative EST clone was sequenced completely.

4 Abbreviations used are: GST, glutathione transferase; EST, expressed sequence tag; MAAI, maleylacetoacetate isomerase.

Send reprint requests to: P. G. Board, Molecular Genetics Group, John Curtin School of Medical Research, P.O. Box 334, Australian National University, Canberra ACT 2601, Australia. E-mail: Philip.Board@anu.edu.au
Human Zeta  
Mouse Zeta  
Trout Zeta  
C. elegans Zeta  
D. melanogaster Zeta  
E. nidulans Zeta  
Carnation Zeta  
Kneat Zeta  
A. thaliana Zeta  
Rice Zeta  
Cotton Zeta  
Human Theta  
Human Pi  
Human Mu  
Human Alpha  

Identification of Allelic Variants. To search the EST database for allelic variants of a particular gene, the complete cDNA sequence was used as the query sequence in the blastn program using the default settings. The output of the program used the “flat query-anchored with identities” format, which allows rapid visual scanning of the aligned nucleotides for variation.

Results and Discussion

Identification of the Zeta Class GSTs. A BLAST search of the human EST database with the amino terminal sequence of a GST-like protein identified in carnation (Dianthus caryophyllus) petals (Meyer et al., 1991) revealed a number of similar clones from several cDNA libraries derived from a range of tissues including brain, breast, fibroblast, heart, liver, melanocyte, placenta, skeletal muscle, and pancreas (Board et al., 1997). A representative EST clone (N31040) was completely sequenced and was found to contain an open reading frame encoding 216 residues with a deduced molecular size of 24166 Da. The size of the encoded peptide falls well within the range of the previously described cytosolic GST subunits (Board et al., 1997).

To determine whether this human clone represented a new member of the GST gene family, the amino acid sequence was aligned with representative sequences from previously described GST classes in humans and other species (Board et al., 1997). Phylogenetic analysis based on that alignment clearly demonstrated that the human GST clone and the carnation sequences were closely related to each other, and more distantly related to previously described members of the GST super family. As a result of these studies, it was concluded that the human GST Z1 gene has 8 introns, the C. elegans gene has 5 and the carnation gene has 10 (Itzhaki and Woodson, 1993; Blackburn et al., 1998). Presumably, the long period since the divergence of these species has allowed reorganization of the gene, while the primary structure has been conserved.

In situ hybridization studies have been undertaken to localize the human GST Z1 gene. Its position at 1q42.3 is clearly distinct from the Alpha class genes at 6p12, the Mu class genes at 1p13, the Pi class gene at 11q13, and the Theta class genes on 22q11.2 (Blackburn et al., 1998). Thus the structure and localization of the Zeta class GST genes adds considerable weight to the conclusion that the Zeta class is a distinct new member of the GST super family. This search strategy has also been used to identify new members of the Alpha class (Board, 1998) and could be readily applied to the identification of new members of other gene families.

Functional Characterization of GST Z1–1. Although database analysis has led to the discovery of the Zeta class, this strategy provides few specific indications of the protein’s function. The detection of GST Z1 cDNA clones in libraries from a wide range of tissues and the detection of homologs, over such a wide evolutionary range, suggested that the Zeta class GSTs may catalyze the metabolism of a common metabolite of a product or a significant component in the environment (Board et al., 1997). In initial studies with recombinant GST Z1–1 expressed in Escherichia coli, a range of substrates utilized by other members of the GST family were tested. With the exception of glutathione peroxidase activity with t-butyl hydroperoxide and cumene hydroperoxide, there was little detectable activity. Recent studies have shown that GST Z1–1 can catalyze the glutathione-dependent oxygenation of dichloroacetic acid to glyoxylic acid (Tong et al., 1998a). Subsequently, a number of other α-halokanoic acids have been shown to be substrates (Tong et al., 1998b). Dichloroacetic acid has been shown to be carcinogetic in male B6C3 F1 mice and male Fischer 334 rats (Herren-Friend et al., 1987; Bull et al., 1990; DeAngelo et al., 1996; Pereira, 1996). Humans can be exposed to dichloroacetic acid in drinking water, as it is one of the most common disinfection by-products found in chlorinated water supplies (Uden and Miller, 1983). Despite its potential carcinogenicity, dichloroacetic acid is used clinically in the treatment of congenital lactic-acidosis because of its ability to stimulate mitochondrial pyruvate dehydrogenase (Stacpoole et al., 1997). Fernandez-Canon and Penalva (1998) cloned and sequenced a cDNA encoding human maleylacetoacetate isomerase (MAAI) and found that it was identical to GST Z1. MAAI catalyzes the glutathione dependent isomerization of maleylacetoacetate to fumarylacetoacetate. This reaction is an essential step in the catabolism of phenylalanine and tyrosine. Deficiencies of other steps in this pathway cause alcaptonuria, phenylketonuria, and tyrosinemia (Fernandez-Canon and Penalva, 1998). This essential metabolic role for GST Z1 provides an explanation for its conservation and expression in such a wide range of species.

Identification of Allelic Variation by Database Analysis. Allelic variants that alter the substrate specificity, reaction kinetics, or stability of an enzyme can be of particular clinical significance and result in increased or decreased drug clearance, drug toxicity, and susceptibility to environmental carcinogens or toxins. Identification of such variants has often been dependent on the detection of an associated phenotype or the chance sequencing of a variant cDNA. Thus, many variants may exist that have not been discovered and characterized.

Because of the number of sequences and the number of individuals...
the presence of Arg at position 42. A recent study suggests that dichloroacetic acid inactivates Zeta class GSTs (Cornett et al., 1999), and our data indicate that the higher activity of GST Z1A-1A correlates with the presence of Arg at position 42. A recent study suggests that dichloroacetic acid inactivates Zeta class GSTs (Cornett et al., 1999), and our data indicate that the higher activity of GST Z1A-1A may be due to increased resistance to substrate-mediated inactivation (Tzeng et al., 2000).

Table 1 shows that GST Z1A-1A appears to have significantly higher activity with dichloroacetic acid than the other isoforms. This correlation has facilitated the determination of their enzymatic activity.

**Table 1**

<table>
<thead>
<tr>
<th>Allele</th>
<th>Haplotype</th>
<th>Specific Activity (nmol/min/mg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GST Z1 * A</td>
<td>Lys 32, Arg 42, Thr 82</td>
<td>1610 ± 20</td>
</tr>
<tr>
<td>GST Z1 * B</td>
<td>Lys 32, Gly 42, Thr 82</td>
<td>450 ± 20</td>
</tr>
<tr>
<td>GST Z1 * C</td>
<td>Glu 32, Gly 42, Thr 82</td>
<td>450 ± 20</td>
</tr>
<tr>
<td>GST Z1 * D</td>
<td>Glu 32, Gly 42, Met 82</td>
<td>301 ± 25</td>
</tr>
</tbody>
</table>

Alpha Class Variants. We have also screened the EST database for allelic variants in the Alpha class GSTs. There are three members of the human Alpha class for which full cDNA sequences are known (GST A1, GST A2, and GST A4). So far, two variants in GST A2 have been confirmed. These polymorphisms result in Thr112Ser and Glu210Ala substitutions that reach polymorphic frequencies in the Australian European populations.

**Limitations Associated with EST Database Analysis.** Our strategy for the analysis of the EST database for allelic variants has used readily available BLAST programs and simple rules for the elimination of false positives. While this approach has obvious advantages, such as simplicity and demonstrable productivity, there are a number of disadvantages and limitations that should be recognized.

1. The cDNA that is being studied must be present in libraries from a number of different tissues to ensure that the clones searched are derived from as many individuals as possible. The more clones identified in different libraries, the greater the sensitivity of the search. Rare variants will be missed by this procedure, as they will probably be attributed to sequencing errors.

2. There is little information available on the ethnic origin of the tissues used to construct the cDNA libraries represented in the database. Some variants that are frequent in one ethnic group may be rare or absent in another.

3. As the ESTs are only sequenced from the 5' and 3' ends, there are relatively few sequences covering the central regions of large cDNAs, thus lowering the probability of detecting variation in these areas.

Recent studies in other laboratories have also recognized the potential of EST database analysis for the identification of polymorphisms (Buetow et al., 1999; Forsberg et al., 1999). Buetow et al. (1999) have generated a series of computer programs termed the SNPpipeline to specifically carry out this analysis. In preliminary studies using these programs and our own BLAST-based strategies, we found that while some polymorphisms were identified by both procedures, some were only detected by one procedure. Thus, we conclude that an eclectic approach is possibly the most reliable.

**References**


DeAngelo AB, Daniel FB, Most BM and Olson GR (1996) The carcinogenicity of dichloroacetic acid in the male Fischer 344 rat. Toxicology 136:207–221.


