The Glutathione S-Transferase D Genes

A DIVERGENTLY ORGANIZED, INTRONLESS GENE FAMILY IN DROSOPHILA MELANOGASTER*

Yann-Pyng S. Toung§, Tao-shih Hsieh¶, and Chen-Pei D. Tu††

From the †Department of Molecular and Cell Biology, The Pennsylvania State University, University Park, Pennsylvania 16802 and the ‡Department of Biochemistry, Duke University School of Medicine, Durham, North Carolina 27709

(Received for publication, September 30, 1992)

We have characterized a cluster of glutathione S-transferase genes located at 87B on the Drosophila polytene chromosome near the heat shock genes, hsp70. These genes, designated gst Ds in the glutathione S-transferase gene superfamily, are closely linked within a ~60-kilobase DNA segment. The gene family has a minimum of eight intronless genes organized in divergent orientations. Two of the genes are probably GST pseudogenes in that their open reading frames are shorter than functional GSTs, and no RNAs from them have been detected thus far. The amino acid sequence identity among the functional genes ranges from 53 to 75% in pairwise comparisons. The intergenic regions are much more AT rich (63-73%) than the coding regions (41-52%), consistent with being promoter/regulatory sequences in Drosophila melanogaster. The mRNA size for each gene suggests that these genes are probably expressed individually from separate promoters. This is the first documentation of definitive physical linkage of a functional glutathione S-transferase multigene family. The genes are divergently organized, and a gradation of sequence similarity exists among the encoded GST isozymes. The patterns of sequence similarity in pairwise comparisons of the family members suggest that gene conversion may have played a role in the evolution of this GST multigene family. We propose that the Drosophila gst D genes provide a unique system for studying GST gene regulation, in vivo physiological functions, and evolution of substrate specificities with a global perspective. The gst D genes in other organisms should be intronless and can be isolated directly from genomic DNAs for functional analyses at the gene and protein levels.

* This work was supported by The Pennsylvania State University Agricultural Experiment Station, Biomedical Research Support Grant 2 S07 RR07082 to The Pennsylvania State University, and by United States Public Health Service Grant GM29906 (to T.-s. H.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

†† Present address: Dept. of Biological Sciences, Stanford University, Stanford, CA 94305.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) M97702.

§ Present address: Dept. of Biological Sciences, Stanford University, Stanford, CA 94305.

¶ To whom correspondence should be addressed; 401 Paul M. Althouse Laboratory, The Pennsylvania State University, University Park, PA 16802. Tel.: 814-863-2096; Fax: 814-865-7024.

The abbreviations used are: GSTs, glutathione S-transferases; kb, kilobase(s); bp, base pair(s); DR, direct repeat; IGS, intergenic sequence.

ubiquitous in nature. They are a family of multifunctional proteins essential in xenobiotic biotransformation, drug metabolism, and protection against peroxidative damage (for recent reviews, see 1–4). To accomplish these diverse physiological functions most organisms have the genetic capacity to encode multiple GST isozymes. For example, the human GSTs are encoded by at least four classes of genes, which constitute a gene superfamily (6–11). Different gene families are located on different chromosomes. One of the gene families, the gst B genes, is a dispersed gene family residing on at least three human chromosomes (11, 12). Multiple GSTs exist in various plants and are responsible for the detoxification of a number of herbicides, including atrazine (13–17). Changes in GST protein expression have been implicated in the development of resistance to alkylating agents in cancer chemotherapy (18, 19) and in the resistance of select pesticides in certain insects (for reviews, see Ref. 2). Drosophila melanogaster has at least two classes of GSTs which are immunologically distinct from each other (20). Our laboratory has isolated and characterized a GST cDNA for the Drosophila GST subunit 1, designated pGTDm1 (21). This gene is a member of a multigene family that is designated the gst D genes in the gst gene superfamily (22, 23). The Drosophila is an attractive system for studying the in vivo functions of GSTs because of the simplicity of the Drosophila genome and the relative ease of germ line transformation. The wealth of well characterized genetic mutations in the literature and the fine resolution of the polytene chromosomes are major advantages over the mammalian system in elucidating the physiological functions of GST at the molecular genetic level. The high multiplicity of the GST D isozymes provides a system for understanding the functional evolution of GST substrate specificities. In this paper we intend to provide the structural basis for elucidating the molecular mechanisms of gst gene expression and evolution of substrate specificities.

EXPERIMENTAL PROCEDURES

Materials—General chemicals and antibiotics were purchased from Sigma. Restriction endonucleases and T4 DNA ligase were products of New England Biolabs (Beverly, MA) or Boehringer Mannheim. A random-primed DNA labeling kit was purchased from Boehringer Mannheim. Exonuclease-deficient T7 DNA polymerase was kindly provided by Dr. Kenneth A. Johnson of The Pennsylvania State University. The dGTP labeling mix kit for sequencing was obtained from U. S. Biochemical Corp. The [α-32P]dCTP (specific activity 3,200 Ci/mmol) and α-32P-dATP (specific activity >600 Ci/mmol) were products of Amersham Corp. Pall Biodyne nylon membranes and nitrocellulose sheets were obtained from ICN and Schleicher & Schuell, respectively. The isolation of the GST genomic clone GTDm101 in λEMBL4 vector has been described previously (21). The genomic fragment came from a partial Sau3A1 digestion product inserted at the BamHI site of λEMBL4 vector. Adult flies (Oregon R) for RNA isolation were kindly provided by Dr. David Gilmour of The Pennsylvania State University.
Methods for Nucleic Acid Analysis—Drosophila genomic DNA was isolated from the P2 strain of Oregon R as described (21). Total Drosophila RNA from adult flies was prepared according to the method of Thummel et al. (24) with minor modifications. Flies were homogenized in at least a 10-fold excess of 6 M guanidinium HCl, 0.1 M sodium citrate, 10 mM sodium EDTA, and 0.5% SDS. The DNA was pelleted by centrifugation, and the supernatant was layered over a cushion of 5 M guanidinium HCl, 10 mM EDTA in an ultracentrifuge tube. The RNA was pelleted by centrifugation in a SW 27 rotor at 24,000 rpm for 1 h at 20 °C. The supernatant fraction was carefully removed by aspiration, and the clear RNA pellet was resuspended in 1 ml of 0.3 M sodium acetate (pH 5.5) by heating at 65 °C for 15 min. DNA fragments from AGTDm101 were subcloned into plasmid pUC18 or pUC19 and/or phagemid (pBluescript II KS−/− or pBluescript II SK−/−, Stratagene, San Diego, CA) vectors as described (25). Northern and genomic Southern hybridizations were carried out on nylon membranes and nitrocellulose membranes. Some of the probes are restriction fragments of the genomic clone that lie 3′ to the end of each DmGST coding region but before the beginning of a neighboring gene (please refer to Fig. 1 for their locations).

DNA Sequencing—DNA sequence analysis was performed on double-stranded DNA templates using the dideoxynucleotide chain termination method and α-32P-dATP (26–28). All subclones were sequenced on both strands more than twice.

DNA Data Analysis—Sequence data were assembled and analyzed using the DNA Inspector IIe program from the Textco company (West Lebanon, NH) for amino acid composition analysis and base composition analysis. Sequence homology searches in the GenBank (5,27) and UEMBL 27-68 data base were performed using IntelliGenetics Fast DB software package provided through The Pennsylvania State University Biotechnology Institute. Multiple sequence alignment was carried out by GENALIGN of IntelliGenetics suite programs (solution parameters: amino research length, 2; deletion weight, 5.0; length factor, 0; matching weight, 1.0; nucleic research length, 4; spread factor, 50). Several alignments necessitated the introduction of gaps at various locations to increase the identity in the overall protein sequences. Phylogenetic analysis was performed using the Phylogenetic Analysis Using Parsimony (PAUP) program version 2.0 (9), provided by Dr. David L. Swofford of Illinois Natural History Survey via The Pennsylvania State University Biotechnology Institute.

Pulsed Field Gel Electrophoresis—Intact Drosophila DNAs were prepared from overnight embryos in agarose inserts (100–1 ml volume) and stored in buffer (10 mM Tris-HCl (pH 7.8), 1 mM EDTA, 50 mM NaCl, 0.5% SDS) (22). The agarose inserts were soaked overnight in 10 ml buffer without SDS with several changes to remove residual SDS. Each insert was cut into three approximately equal pieces, and each piece was digested with 50 units of restriction endonucleases in the presence of 20 pg of bovine serum albumin for 6.5 h at 37 °C. The restriction digests were resolved by adding 1 ml of 20 mM EDTA and 0.5% SDS to the agarose for 20 min. The agarose inserts were then placed into agarose gel (1%) slots with a bent Pasteur pipette and sealed with low melting agarose. Electrophoresis was carried out in 0.5 × TBE at 240 V with a 10-s switch time on a CHEF apparatus (29). The cooling circulator was set at 4 °C. The duration of electrophoresis was 15.5 h. Ligated λDNA and EcoRI-digested λDNA were used as size markers. The gel was photographed after staining with ethidium bromide and processed for Southern hybridization according to Sambrook et al. (25).

RESULTS

Organization of the Drosophila gst D Genes in AGTDm101—A partial restriction map of the AGTDm101 DNA was generated from the single and double restriction digests of phage DNA with BamHI, EcoRI, HindIII and resolution of the fragments on 8% agarose gel (data not shown). Fragments homologous to the pGTDm1 cDNA probe were identified by Southern hybridization and subsequently cloned in pUC18/19 vectors for fine restriction mapping and nucleotide sequence analysis (25–28). These cloned inserts were characterized by several restriction digestions before a series of further subcloning and sequencing experiments. DNA sequence analysis was accomplished on double-stranded templates according to the sequencing strategy in Fig. 1. The sequence was determined on both strands, and all the restriction sites used for subcloning were crossed. The complete sequence of the AGTDm101 genomic DNA consists of 14,707 nucleotides (GenBank accession number M87702).

Analysis of AGTDm112 sequences revealed a total of eight Drosophila GST genes and pseudogenes, including the genomic equivalent of the pGTDm1 cDNA (21). These gst genes were designated gstD1 (encoding DmGST subunit 1), gstD21, gstD22, gstD23, gstD24, gstD25, gstD26, and gstD27. They were identified on the basis of DNA sequence similarity to the pGTDm1 cDNA. A dot-matrix analysis of AGTDm101 was carried out relative to the pGTDm1 cDNA. The result is shown in Fig. 2. The nomenclature reflected the divergent organization of gd1 relative to the rest of the family members. Their relative locations in AGTDm101 were shown in Fig. 1. Such a divergent organization has been observed previously in a genomic clone for the rat Yγ genes.2 The gstD1 gene is transcribed in an orientation opposite to that of the other seven genes; it is separated from gstD21 by 3,149 nucleotides of intergenic DNA. Additional GST genes in AGTDm101 are arranged in tandem, along the orientation of gd21 with less than 1 kb of intergenic DNA between genes. Each of these GST sequences was contiguous in reading frames relative to the codons of the cDNA sequence in pGTDm1 (21). Sequences surrounding the ATG codon of six GST D genes (gstD1, D21, D22, D23, D24, D25, D27) conform to Kozak’s consensus sequence (PuNNATGG) in the translation initiation (30). Each of the gstD genes has an open reading frame beginning with a methionine codon and extends longer than 170 amino acids. The deduced amino acid sequences from these eight gst genes/pseudogenes are designated DmGST1 (209 amino acids), DmGST21 (215 amino acids), DmGST22 (199 amino acids), DmGST23 (215 amino acids), DmGST24 (216 amino acids), DmGST25 (215 amino acids), DmGST26 (171 amino acids), and DmGST27 (212 amino acids). They are aligned for maximal identity and presented in Fig. 3.

The DmGST22 (open reading frame of gstD22) does not have a purine nucleotide at the −3 position for optimal translation initiation (30). The DmGST22 may not be functional as a GST even if it is synthesized. It has a truncation in the amino-terminal region which contains a tyrosine residue highly conserved among the GSTs (5,8). Therefore, the gstD22 is probably a pseudogene. The DmGST26 (open reading frame of gstD26) is preceded by the Kozak’s consensus (PnNATGG) in the translation initiation (30). Each of the gstD genes has an open reading frame beginning with a methionine codon and extends longer than 170 amino acids. The deduced amino acid sequences from these eight gst genes/pseudogenes are designated DmGST1 (209 amino acids), DmGST21 (215 amino acids), DmGST22 (199 amino acids), DmGST23 (215 amino acids), DmGST24 (216 amino acids), DmGST25 (215 amino acids), DmGST26 (171 amino acids), and DmGST27 (212 amino acids). They are aligned for maximal identity and presented in Fig. 3.

Phylogenetic Analysis of the DmGST D Gene Family—Pairwise comparisons among members of the deduced gst D gene products revealed substantial similarity (53–75%) (Table I). The amino acid compositions of the putative DmGST subunits are very similar, except for the open reading frames of the presumed pseudogenes gstD22 and gstD26. Many of the amino acid differences are in the number of basic residues such as lysine (especially between DmGST24 and DmGST25/27) and acidic/polar residues such as aspartic acid and asparagine (especially between DmGST1 and DmGST25/27). Most notably, the 4 tryptophan residues are found to be strictly conserved in number among all GST D isozymes. The position of these 4 tryptophan residues within each subunit is also conserved except for the DmGST25 subunit, where 1 trypto-

FIG. 1. Schematic representation and sequencing strategy of \( \lambda \)GTDM101 DNA. Schematic representation of partial restriction map and sequencing strategy of \( \lambda \)GTDM101. Only restriction sites used for subcloning and sequencing are shown. The enzyme abbreviations used are: A, AluI; B, BamHI; Bg, BglII; E, EcoRI; Ha, HaeIII; Hc, HincII; H, HindIII; K, KpnI; P, PstI; Pu, PvuII; R, RsaI; S3, Sau3A1; Sc, ScaI; St, StuI; T, TaqI; X, XbaI; and Xm, XmaI. Bold arrows indicate relative position and direction of transcription of each \( \text{gstD} \) gene. Arrows on the top and bottom portions of the figure represent the direction and extent of each sequence determination. The open reading frames of the \( \text{gstD} \) genes are shown by the solid black bars, whereas the 3\(^\prime\) noncoding probe for each specific \( \text{gstD} \) gene is represented by different designed box with proportional length. The complete nucleotide sequence (14,707 nucleotides) can be retrieved from the GenBank via accession number M97702.

FIG. 2. Dot-matrix analysis of \( \lambda \)GTDM101 (14,707 nucleotides) versus \( p \)GTDM1 (770 nucleotides). The complementary strand of \( p \)GTDM1 was used for the comparison with the insert in \( \lambda \)GTDM101. The genomic sequence of \( p \)GTDM1 did not show up because it is divergently organized relative to the other members of the \( \text{gstD} \) gene family. The search element length is 12 nucleotides, and the maximum number of mismatches allowed is 2.

To explore the evolutionary relationship among members of this family with housefly \( M_d \text{GST1} \) as an outgroup (31), the phylogenetic analysis of the DNA sequences was performed by using the PAUP program. The shortest tree is shown (Fig. 4) together with the phylogenetic distance between gene members (Table II). \( \text{gstD21} \) and \( \text{gstD24} \), with 83\% identity at the DNA sequence level, are related to each other by the shortest phylogenetic distance (0.171, Table II). \( \text{gstD1} \) is distantly related to the other members, appropriate for its divergent orientation.

The nucleotide sequences in the coding regions of \( D_m \text{GSTs} \) showed substantial numbers of identical sequence blocks in pairwise comparisons as shown in Fig. 5. For example, there is a stretch of DNA sequence (nucleotide positions 133–197) found to be identical in both \( D_m \text{GST21} \) and \( D_m \text{GST24} \), resulting in a stretch of 22 identical amino acids. Different blocks of identical sequences exist between different pairs of \( \text{GST} \) sequences. Another feature commonly observed in functional \( Drosophila \) and human genes is a strong preference (over 60\%) for G and C at the third positions of the codons in all \( D_m \text{GSTs} \). It has been speculated that specific codon usage may play a role in translational regulation through limitation of tRNA availability (32) or secondary structure at the mRNA level (33).

Linkage of the \( Drosophila \) \( \text{gstD} \) Genes—The genomic Southern hybridization patterns with two GST DNA probes (\( p \)GTDM1, and \( \lambda \)GTDM101) are shown in Fig. 6. Panels \( A \) and \( B \) represent digestion patterns with seven restriction enzymes on regular agarose gel electrophoresis and hybridized with \( \text{gstD1} \) (\( p \)GTDM1 cDNA insert, panel \( A \)) and \( \text{gstD25} \) (a \( \sim 900\)-bp EcoRI-RsaI fragment, panel \( B \)), respectively. Panels \( C \) and \( D \) represent the patterns of a second set of restriction endonuclease digestions separated by CHEF and hybridized with the same \( \text{gstD1} \) (panel \( C \)) and \( \text{gstD25} \) (panel \( D \)) probes as mentioned earlier. The hybridization patterns in panels \( A \) and \( B \) reveal that \( \lambda \)GTDM101 contains most of the genomic fragments from Southern hybridization except for the EcoRI fragments at 5.8 and 2.5 kb. The sequence divergence between the two probes is reflected in the intensity difference in some of the band signals. Several bands (e.g. the 2.4-kb \( \text{PstI} \) fragment in lane 6 and the 2.55-kb \( \text{SacI} \) fragment in lane 7) in
**FIG. 3.** Sequence comparison among *Drosophila* GST D family members. Vertical dashes between sequences indicate identical amino acids (capital letter). Gaps to maximize sequence identity are indicated by blank spaces. A consensus sequence is shown at the bottom of the figure with those amino acids conserved among all family members capitalized. Dashes in consensus sequence represent highly divergent positions. The two putative pseudogene products are also included for comparison.

Panel A are weakened to the extent of distinction in panel B and vice versa (e.g. the 0.94-kb *PstI* fragment in lane 6). Similar observations are obvious in panels C and D. For example, additional bands of 103 and 18 kb were revealed by the *gstD25* probe in the *SmaI*-digested DNA in lane 6. The presence of multiple GST genes in *Drosophila* is obvious from the complex hybridization patterns and from the fact that members of this gene family are intronless and consequently very short (Fig. 1, Ref. 23). The CHEF blot hybridization patterns revealed that this *gst* gene family may be closely linked within a maximal distance of ~60 kb. The two DNA probes revealed different hybridization patterns (panels C and D) in *SspI* (lanes 4) and *SmaI* (lanes 5) digestions. The smallest single band signal with both DNA probes is at the ~60-kb *MluI* fragment (lanes 3). Because there is no *MluI* site in XGTDmlOl, it is most probable that the *gstD* multigene family is located on a ~60-kb *MluI* fragment. Clones overlapping with XGTDmlOl have been isolated, and partial sequence...
Analysis of genomic DNA from Drosophila embryos was performed using the 3' region of gstD25 (230 bp, PvuII-EcoRI) as probe (Fig. 7). The gstD25 was chosen because this gene is the most divergent fromgstD1 (pGTDM1). The genomic DNA was restricted by each of the four restriction enzymes, BamHI, EcoRI, HindIII, and PstI. The major fragments observed in each enzyme-digested sample was 3.9 kb (BamHI), 4.0 kb (EcoRI), 3.5 kb (HindIII), and 8.3 kb (PstI), respectively. In addition, several minor (weaker) signals were also revealed in this hybridization. They are BamHI digest: 16, 7.8, and 0.85 kb; EcoRI digest: 5.8 and 2.5 kb; HindIII digest: 5.5, 4.2, 2.0, and 1.0 kb; PstI digest: 5.1 kb. Whether these signals represent the presence of additional related genes or pseudogenes or just the presence of a repetitive sequence element in the 230-bp hybridization probe remains to be determined.

Expression of GST D Gene Family Members in Adult Flies—To identify the specific GST transcripts in the RNA preparations, the unique 3' regions (Fig. 1) from each DmGST D gene in MDGSTm101 were used as specific probes in RNA blot hybridization (25). The gstD1 mRNAs were the most abundant species among transcripts of this family as shown in Fig. 8. Compared with gstD1 expression, gstD21, gstD23, and gstD25 had a low but detectable expression level in adult flies. Two distinct mRNA species were detected by the gstD25-specific probe. After stripping off the gstD25 probe, the filter was hybridized with gstD1 probe, and only one band showed up (Fig. 8, lane 5). There are four polyadenylation signals (either AATAAAA or ATTAAAA) in the 3'-noncoding region at 60, 227, 407, and 473 nucleotides downstream from the stop codon of gstD25. The choice of alternative polyadenylation sites in gstD25 expression may account for the observation of two distinct transcripts in the Northern hybridization results (Fig. 8). It is interesting to note that the ATTAAA sequence is just one nucleotide upstream of the TAA termination codon in gstD23, a second polyadenylation signal ATAAA is 544 nucleotides downstream of TAA. In gstD24 one polyadenylation signal ATAAA overlaps with the termination codon (underlined) whereas a second signal is 503 nucleotides away. The messages specific for gstD22, gstD24, gstD26, and gstD27 have not been detected in adult flies.

Analysis of Intergenic Sequences—The intergenic sequence (IGS) refers to the nucleotide sequences between two adjacent genes. The IGS between gstD1 and gstD21, designated IGS1-21, is presumably the promoter region for the two divergently oriented genes. The other intergenic sequences, designated IGS21-22, IGS22-23, etc., were regions containing the 3'-untranslated sequence of the preceding gene and the 5' untranslated sequence of the next downstream gene. The nucleotide sequences of these regions are much more highly AT rich than the coding sequences (Fig. 9). The extents of AT richness are 63.4% (IGS1-21, 3,419 bp), 67.6% (IGS21-22, 401 bp), 68.2% (IGS22-23, 425 bp), 67.4% (IGS23-24, 966 bp), 73.3% (IGS24-25, 759 bp), 73.1% (IGS25-26, 675 bp), and 63.6% (IGS26-27, 976 bp), respectively. No significant open reading frame (>100 amino acids) could be identified, in either orientation, in these intergenic regions. In contrast, the coding regions have compositions of 40.8% AT (gstD1), 46% (gstD21), 46.5% (gstD27), 46.6% (gstD24), 51.2% (gstD23), and 51.8% (gstD25).

Detailed examination of these IGSs revealed additional interesting features. Several consensus sequences for transcription factors were found in IGS1-21, a putative promoter region for gstD1 and gstD21. Two binding sites of activator protein-1, consensus sequence (C/G)TGAC(C/G)TC(A/A); 34, 35) are present in the 5'-flanking region of gstD21 relative to the ATG codon (−376 to −369, 5'GTGACTCA3'). Two 12-O-tetradecanoylphorbol-13-acetate responsive-like elements were found in the mRNA.

### Table I

<table>
<thead>
<tr>
<th>Pairwise comparison</th>
<th>Amino acid sequence homology</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 us. 21</td>
<td>70</td>
</tr>
<tr>
<td>1 us. 23</td>
<td>68</td>
</tr>
<tr>
<td>1 us. 24</td>
<td>68</td>
</tr>
<tr>
<td>1 us. 25</td>
<td>58</td>
</tr>
<tr>
<td>1 us. 27</td>
<td>67</td>
</tr>
<tr>
<td>21 us. 23</td>
<td>75</td>
</tr>
<tr>
<td>21 us. 24</td>
<td>60</td>
</tr>
<tr>
<td>21 us. 25</td>
<td>65</td>
</tr>
<tr>
<td>22 us. 24</td>
<td>72</td>
</tr>
<tr>
<td>23 us. 25</td>
<td>58</td>
</tr>
<tr>
<td>23 us. 27</td>
<td>62</td>
</tr>
<tr>
<td>24 us. 25</td>
<td>59</td>
</tr>
<tr>
<td>24 us. 27</td>
<td>64</td>
</tr>
<tr>
<td>25 us. 27</td>
<td>53</td>
</tr>
<tr>
<td>1 us. 22</td>
<td>62</td>
</tr>
<tr>
<td>1 us. 26</td>
<td>74</td>
</tr>
</tbody>
</table>

*DmGST22 and DmGST26 are probably pseudogenes.

### Table II

<table>
<thead>
<tr>
<th>Pairwise phylogenetic distances between gstD genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Above diagonal, mean distances; below diagonal, absolute distances.</td>
</tr>
<tr>
<td><strong>gstD1</strong></td>
</tr>
<tr>
<td>0.273</td>
</tr>
<tr>
<td>0.339</td>
</tr>
<tr>
<td>0.164</td>
</tr>
<tr>
<td>0.161</td>
</tr>
<tr>
<td>0.347</td>
</tr>
<tr>
<td>0.222</td>
</tr>
<tr>
<td>0.347</td>
</tr>
<tr>
<td>0.351</td>
</tr>
<tr>
<td>0.155</td>
</tr>
</tbody>
</table>

Analysis has revealed additional gst genes.3

Related Sequences Elsewhere in the Drosophila Genome—Analysis of genomic DNA from Drosophila embryos was performed using the 3' region of gstD25 (230 bp, PvuII-EcoRI) as probe (Fig. 7). The gstD25 was chosen because this gene is...
strand of gstD1 at positions about 0.6 kb (−674 to −667, TGATACAG) and 2.2 kb (−2173 to −2166, TGATTTCAG) upstream of the ATG initiation codon, respectively. This finding, if proven in vivo, would provide gstD1 gene a specific signal for responding to xenobiotic compounds like 12-O-tetradecanoylphorbol-13-acetate. A heat shock control element CNNGAANNNTTCNNG sequence is present at 1.9 kb (−1902 to −1889, CAGGAATATTCAAG) upstream of ATG codon of gstD1. Furthermore, the presence of direct repeat sequences is one of the interesting features in IGS23–24. An extensive direct repeat sequence of 106 and 109 bp with four nucleotides in between and a few mismatches within the direct repeats occurs at −387 upstream of the ATG of gstD24. This region of direct repeat sequences actually consist of five pairs of direct repeats (DR1, 23 nucleotides; DR2, 18 nucleotides; DR3, 27 nucleotides; DR4, 12 nucleotides; DR5: 20 nucleotides). Another sequence ATTTTG is repeated five times within the intergenic region IGS26–27. Whether these sequence features have any biological significance remains to be determined.

There are also basic RNA polymerase II transcription signals in the intergenic regions. A perfect match to the cap site consensus sequence ATCA(G/T)T(C/T) (36 and references therein) exists at −53 to −47 upstream of the ATG of gstD23 and at −47 to −41 upstream of the ATG of gstD24. Further upstream of the putative mRNA starts, at position −29 of both gstD23 and gstD24, a Goldberg-Hogness box (TATA) was found with a perfect match with the consensus sequence TATAATA (37). Still further upstream of the TATA sequences, a consensus CAAT box was found at positions −81 and −71 upstream of the putative transcription initiation sites of gstD23 and gstD24, respectively.

Other gst D genes in AGTDm101 also have their respective consensus TATA box or TATA-like sequences within 120 nucleotides of each initiation codon. The CAAT box sequence, which is a recognition site for the transcription factors CTF, is also present in the 5′-flanking region of each gene (38). No GC box (Sp1 binding site) was found in any of these intergenic sequences (39). The polyadenylation signals (either AATAAA or ATTAAA) were found within 550 bp after the termination codon in all of the Drosophila gst D genes.

**DISCUSSION**

Structure and Function Relationships—Other than the D. melanogaster gstD gene family, the organization of GST genes and their evolution in insects have not been studied at the molecular level. Insect GSTs are of interest not only because of their potential role in insecticide resistance but also because of the enormous insect species diversity. Such diversity provides a rich source of GST isozymes which in turn provides a large variety of sequence variation for studying the structure/function relationship from an evolutionary perspective.

**FIG. 5.** Block identity of the nucleotide sequence in the coding regions of the deduced DmGSTs. The black filled rectangles are regions of identical nucleotide sequences. They were identified by the DNA Inspector IIe program.
Among a limited number of insects examined for GST activities, only a subset of GST isozymes in these insects is associated with insecticide resistance (40-42). The multiplicity of insect GSTs would predict that they should have many physiological functions other than pesticide detoxification. The multiplicity of GST genes in Drosophila suggests that DmGSTs are encoded by multiple genes and may have diverse functions; most of them are yet to be characterized.

The involvement of specific amino acid residues in substrate recognition and catalysis for GSTs is an active area of research pursuit. A conserved tyrosine residue near the amino-terminal region of mammalian GST's ($\alpha$, $\mu$, $\tau$) has been proven by site-directed mutagenesis to be essential for catalytic activity (5, 43-46). The role specific amino acid residue(s) or domain(s) play in substrate binding and ligand binding (e.g., heme, bilirubin) can be resolved by using molecular genetic approaches based on x-ray structural analysis and functional expression in Escherichia coli. The absence of post-translational modification of GSTs and their cytosolic localization have been ideal for high level expression in E. coli. The fact that GST D genes are intronless provides another advantage in studying these GST isozymes. Genomic DNA can be manipulated directly for heterospecific expression in E. coli to allow purification of GST proteins for biochemical studies. Circumvention of the cDNA (mRNA) level makes the study of molecular enzymology of rare GST's much more practical and should provide biochemical insights of GST's unattainable otherwise.

The high AT content in the 5'-noncoding region of the Drosophila gstD genes is in direct contrast to the base composition of the GST coding sequences. Many consensus sequences for DNA binding proteins are localized in the intergenic regions of gstD genes in addition to many direct and inverted repeats (data not shown). In the 5'-flanking region of gstD1, no match to the consensus cap site motif was obvious, and attempts to determine precisely the transcription initiation site(s) for the S1 digestion to succeed (47). On the other hand, such an uncertain result has prevented ruling out post-transcriptional RNA processing as a mechanism for generating the various mRNA species. A consensus CAAT motif was observed at 222 nucleotides upstream of the ATG codon of gstD1, but only one TATA-like sequence was found in its vicinity. More experiments are needed to locate the specific transcription initiation site(s) for gstD1.

The AT richness is a feature previously associated with rapid mRNA turnover caused by an mRNA processing pathway (48). Instability of mRNA would be responsible for the rapid termination of such a transient gene expression signal. It remains to be determined, however, whether the AT rich 3'-noncoding region plays any role in DmGST mRNA stability. Multiple polyadenylation signals, which are quite common in both animal and plant genes, are found among all gstD
The probe used was in the 3'-noncoding region of the filter was washed at room temperature three times in Denhardt's solution at staining.

Arrows indicated two signals revealed probes for coding DNA fragments of tively (Fig. 1). Numbers at left indicate size markers (λ-HindIII and pBR322:AluI) in kb unit. Hybridization was carried out in 5x SSC, 5 × Denhardt’s solution 50 mm sodium pyrophosphate, 0.1% SDS, 100 μg/ml salmon testis DNA, 40% (v/v) formamide at 40 °C for 40 h. The filter was washed at room temperature three times in 2 × SSC and 0.1% SDS for 90 min and then washed in 0.1% SDS, 0.1 × SSC at 55 °C for 2 h.

The choice of alternative polyadenylation signal AATAAA overlaps with the termination codon (TAA) whereas a second signal is 303 nucleotides away. The fact that not all of the gast genes are expressed in adults suggests differential regulation. The differential expression of each member gene may be controlled by different mechanisms using different combinations of cis-acting and trans-acting factors. As a consequence of such differential regulation, each DmGST may play a specific role at different developmental stages, in different tissues or cells, or under chemical (e.g. drugs, pesticides) or physical (e.g. heat) stresses. Results presented in this communication provide a structural and conceptual basis to elucidate the mechanism of tissue-specific expression and developmental regulation of this divergently organized gene family.

The alignment of the gast gene sequences provides interesting insights into the evolutionary and structural relationship among members. The range of sequence similarity among members (53–75%, Table I) is much broader than that observed in the mammalian GST gene families (e.g. ~70% in the Y1 genes; 11, 49). No trends for block gene duplication or amplification are obvious from the evolution tree analysis. The sequence data fit the conventional model of gene duplication followed by divergence for the evolution of the gast gene family. However, the analysis presented in Fig. 5 would suggest that gene conversion may have played a role in the evolution of the gast gene family. There are regions of sequence identity interspersed with regions of divergence between pairs of different gast genes. Furthermore, the regions of identity vary with different pairwise comparisons. These patterns are supportive of an earlier proposal on the role that gene conversion plays in generating the overlapping but different substrate specificities for the GST multigene family (49).

Although GST 1-1 is the most abundant isozyme in adult flies, the intronless D. melanogaster gast genes not expressed in adult (e.g. gastD24, gastD27) are unlikely to be processed pseudogenes for several reasons. First, by direct expression of genomic DNA in E. coli, gastD27 has been demonstrated to encode a functional GST. Second, these gast genes do not contain any significant length of poly(A) tracts at the 3' end of the gene, nor is it flanked by short direct repeats (5–15 nucleotides), both of which are the hallmarks of processed pseudogenes (50). Third, they have Kozak's consensus sequence, and they lack other obvious defects for translation. A more thorough investigation with gene fusion techniques should reveal their tissue origin of expression at certain developmental stages. The gastD22 and gastD26 are presumed to be gast pseudogenes because their deduced polypeptides are shorter than other functional GSTs. However, it is intriguing that reading frames of such significant size (199 amino acids and 171 amino acids) are maintained in both cases. It is possible that they may serve other functions unrelated to xenobiotic biotransformation. However, no RNAs have been detected in adults from these regions so far. Intronless gast genes are also found in prokaryotic organisms, such as the one encoding dichloromethane dehalogenase from Methylobacterium sp. (51).

The intronless feature of gast genes may be maintained for a selective advantage under stress conditions such as heat shock (52). Most of the heat shock genes in Drosophila are intronless and can be expressed when RNA splicing is impaired under heat shock stress. On the other hand, the splicing
of primary transcripts of hsp83 gene (which contains introns) in Drosophila is known to be inhibited under heat shock (53), similar to the processing of other split genes. Consequently, the intronless feature of D. melanogaster gstD genes may allow translation, bypassing the block in RNA splicing under heat shock stress. Their cytogenetic location at 87B, which is flanked by the two hsp70 gene clusters at 87A7 and 87C1, further supports this view. This possibility, together with the preliminary findings that GST activity measured by 1-chloro-2,4-dinitrobenzene conjugation is not reduced under heat shock in Drosophila cultured cells, further implies that GSTs may serve important functions such as alleviation of oxidative stresses under heat shock conditions.

Tissue specificity of GST activities may be related to the particular detoxification needs of each tissue, which in turn would dictate the consequence of detoxification or bioactivation. We have recently found that gstD1, gstD21, gstD23, and gstD27 are expressed at the third larval stage. Also, drug-induction experiments revealed that gstD21 mRNA was dramatically increased, whereas mRNAs from gstD1, gstD23, and gstD27 showed only a slight increase under the same conditions. Whether gstD24 is transcriptionally silent in the adult fly or its transcripts are unstable remains to be determined by gene fusion experiments. In this respect, the presence of several potential mRNA destabilization signals (5'-AUUA-3') in the 3'-untranslated region of gstD24 becomes important for further investigation (54).

GST activities may be prominently affected by developmental events in other insects. In the sheep blowfly, Lucilia cuprina (Wiedemann), GST conjugation activity measured in crude extracts with 1-chloro-2,4-dinitrobenzene and 1,2-dichloro-4-nitrobenzene as substrates reached a peak during the pupal stage, falling to about 15% of the peak during most of the adult stage (55). Similarly in the cabbage fly D. abbreviatus, GST enzyme levels are maximal during larval and pupal stages but decreased to half of these peak values in the imago (56). Tissue-specific expression of GST has been well documented in mammalian systems (57). In nonvertebrates, organs with functions analogous to that of the vertebrate liver, such as the insect fat body (55, 58) or hepatopancreas of molluscs or crustacea (59), are also found to be high in GST activities.

Analysis of secondary structures of DmGST according to the algorithm of Chou and Fasman (60, 61) revealed a potential helix-turn-helix motif near their carboxyl-terminal regions (amino acids 185–199 in DmGST1, or 184–198 in DmGST21, 23, 24, 25, 27). This structure behavior feature is unique among the GST isozymes, because similar analysis with α, μ, π microsomal GSTs failed to show such features. The helix-turn-helix motif is known to bind DNA in a sequence-specific manner (62, 63). Whether DmGSTs have any nucleic acid binding function remains to be demonstrated. The high multiplicity and special features of GSTs clearly imply that insecticide metabolism is not the only function of GST isozymes. It has been found repeatedly that not all GST isozymes are associated with known insecticide resistance (40–42). Many other classes of compounds have been shown to be good substrates for GSTs (64, 65). Recent reports (31, 66) on the housefly GST cDNA sequences indicate that GST gene families may be highly conserved among insects. The molecular characterization of the Drosophila gdst gene family demonstrates that the cDNA level can be bypassed for biochemical studies in many of the intronless insect GST genes.

Acknowledgments.—We thank Hong Tang for sharing unpublished results, M. Frank Mallette for reading the manuscript, and Eileen McConnell for typing it.

REFERENCES
GSH S-Transferase D Gene Family