Characterization of the Human Class Mu Glutathione S-Transferase Gene Cluster and the GSTM1 Deletion*

(Received for publication, September 19, 1997, and in revised form, December 1, 1997)

Shi-rie Xu‡, Ying-ping Wang§, Bruce Roe§, and William R. Pearson‡§

From the §Department of Biochemistry, University of Virginia, Charlottesville, Virginia 22908 and the ¶Department of Chemistry, University of Oklahoma, Norman, Oklahoma 73019

A partial physical map has been constructed of the human class Mu glutathione S-transferase genes on chromosome 1p13.3. The glutathione S-transferase genes in this cluster are spaced about 20 kilobase pairs (kb) apart, and arranged as 5′-GSTM4–GSTM2–GSTM1–GSTM5-3′. This map has been used to localize the end points of the polymorphic GSTM1 deletion. The left repeated region is 5 kb downstream from the 3′-end of the GSTM2 gene and 5 kb upstream from the beginning of the GSTM1 gene; the right repeated region is 5 kb downstream from the 3′-end of the GSTM1 and 10 kb upstream from the 5′-end of the GSTM5 gene. The GSTM1-0 deletion produces a novel 7.4-kb HindIII fragment with the loss of 10.3- and 11.4-kb HindIII fragments. The same novel fragment was seen in 13 unrelated individuals (20 null alleles), suggesting that most GSTM1-0 deletions involve recombinations between the same two regions. We have cloned and sequenced the deletion junction that is produced at the GSTM1-null locus; the 5′- and 3′-flanking regions are more than 99% identical to each other and to the deletion junction sequence over 2.3 kb. Because of the high sequence identity between the left repeat, right repeat, and deletion junction regions, the crossing over cannot be localized within the 2.3-kb region. The 2.3-kb repeated region contains a reverse class IV Alu repetitive element near one end of the repeat.

The glutathione S-transferases (EC 2.5.1.18) are a superfamily of catalytic and binding proteins that detoxify chemical carcinogens (1). Based primarily on protein sequence similarity, the soluble glutathione S-transferases have been divided into four protein families: classes Alpha, Mu, Pi, and Theta. Class Alpha and Mu families typically contain four or more members in mammalian genomes. In humans, a large fraction of the population carries polymorphic deletions for the class Mu GSTM1 gene (about 50% of the population is homozygous GSTM1-null; Ref. 2) and the class Theta GSTT1 gene (40% of the population is homozygous GSTT1-null; Ref. 3).

Significant associations between the GSTM1-0 deletion and increased cancer incidence in case/control studies have been reported for lung (4–6), bladder (7, 8), and skin cancer (9–11). Significant associations between the GSTM1-0 deletion and lung cancer among caucasians and African Americans in southern California suggests that, in these populations at least, the association of the GSTM1-0 deletion and cancer is not strong. London et al. (6) found a significant association between lung cancer and the GSTM1 deletion only in cancer patients with a history of smoking but who had smoked less than 40 pack-years (e.g. 2 packs/day for 20 years) (odds ratio = 1.77, 95% CI = 1.11–2.82). After regression to remove age, sex, and smoking history effects, no significant association was found for total lung cancer cases or for cases where the patient had smoked more than 40 pack-years. These authors reviewed the literature on lung cancer and the GSTM1 deletion and argue that while an association may be present, it is not strong for Caucasian and African American populations. However, a meta-analysis by McWilliams et al. (12) comes to the opposite conclusion. They examined 12 case/control studies with a total of 1593 cases and 2135 controls and concluded that GSTM1-0 is a moderate risk factor with an odds ratio of 1.41 (95% CI = 1.23–1.61), accounting for about 17% of lung cancer cases (12).

Although there is a substantial literature on epidemiological associations between the GSTM1-0 deletion and cancer, very little is known about the mechanism of the deletion. We do not know the size of the deletion or whether other expressed genes are lost. Likewise, the mechanism of the GSTM1-0 deletion is unknown. In this paper, we present a physical map of four of the five class Mu genes in the GSTM1 cluster on chromosome 1 and show that the GSTM1-0 deletion apparently results from homologous unequal crossing over between two highly identical regions that flank the GSTM1 gene, resulting in a 15-kb deletion that contains the entire GSTM1 gene. The GSTM1 gene and 5′- and 3′-flanking regions are excised relatively precisely, leaving the flanking GSTM4, GSTM2, GSTM5, and GSTM3 genes intact. There is no change in at least 5 kb 3′ to the 5′-flanking GSTM2 gene or in at least 10 kb 5′- to the 3′-flanking GSTM5 gene. In addition, the same deletion has occurred in all of the deleted alleles examined. Identification of the GSTM1-0 recombination region provides a hybridization probe that can be used to distinguish GSTM1+-/- heterozygotes from GSTM1-/+ homozygotes.

**EXPERIMENTAL PROCEDURES**

Reagents—[α-32P]dATP, [α-32P]dCTP and [γ-32P]ATP were obtained from ICN Biochemicals. Restriction enzymes were obtained either from Life Technologies, Inc. or New England Biolabs. T4 DNA ligase and A-HindIII molecular size markers were obtained from Life Technologies. T4 polynucleotide kinase and DNA size markers for pulse field gel electrophoresis (Mid-range PFG Marker II) were purchased from New England Biolabs. DNA Tag polymerase and Klenow fragment were obtained from Promega. The Sequenase version 2.0 kit was obtained from U.S.C. Section 1734 solely to indicate this fact.

*This work was supported by American Cancer Society Grant CN-27D (to W. R. P.) and National Human Genome Research Institute Grant R01 HG00313 (to B.A.R.). 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 1754 solely to indicate this fact.

†To whom correspondence should be addressed: Dept. of Biochemistry, Box 440 Jordan Hall, University of Virginia, Charlottesville, VA 22908. Tel.: 804-924-2818; Fax: 804-924-5069; E-mail: wrp@virginia.edu.

‡The abbreviations used are: CI, confidence interval; kb, kilobase pair(s); PCR, polymerase chain reaction; YAC, yeast artificial chromosomes; GST, glutathione S-transferase.
Cloning of a Deletion Junction Fragment—A GeneAmp XL PCR kit (Perkin-Elmer) was used to amplify a 7.4-kb HindIII junction fragment with the following primers: primer 1, 5'-CTCGAGCCCTCTCTCTGTGGT-3'; and primer 2, 5'-GATGTCGACGATCCACACTAAGAT-3'. Primer 1 anneals to the 3'-end of GSTM2, primer 2 anneals to the 5'-end of GSTM5. Long range PCR reactions were performed in a 100-μl solution of 1× buffer (supplied with the PCR kit), a 200 μM concentration of each dNTP, 40 pmol of each primer, 1.5 mM of MgOAc₂, 1 μg of DNA template and 4 units of enzyme. A "hotstart" protocol with AmpliWax™ was used to increase the specificity of the reaction, which was cycled in order for 1 min at 94 °C, 16 cycles of 30 s at 94 °C, 30 s at 54 °C, and 72 °C for 1 min. PCR products were phenol-extracted, cut with HindIII, and separated on a low melting agarose gel. The 7.4-kb fragment was recovered and purified. The 7.4-kb fragment was then ligated to pGEM-TZI vector and used to transform SURE competent cells. Plasmid DNA was purified with a QIAGEN Plasmid Maxi kit; an extra ethanol precipitation step was used to reduce the salt concentration for DNA sequencing.

DNA Sequencing and Sequence Assembly—Subcloned fragments were either sequenced manually using a Taq Cycle Sequencing kit (Amersham) or were sequenced on an ABI PRISM 377 sequencer with the ABI PRISM Dye Terminator Cycle Sequencing Ready Reaction kit at the University of Virginia Sequencing Center. Sequence data were further manipulated at the Human Genome Center at University of Oklahoma. At the University of Oklahoma, cosmid DNAs were isolated free from E. coli host contamination by the cleared lysate, diatomaceous earth-based procedure describe earlier (17) and sequenced to a level of 4.5-fold redundancy via the previously described, double-stranded, shotgun-based approach (18) using the ABI PRISM fluorescence-labeled terminators and either forward or reverse universal primers. Sequencing vector regions were removed, and the resulting data were assembled into contiguous fragments initially using the TED and XGAP programs (19) and more recently using the Phred, Phrap, and Consed programs. The individual contigs were joined into a final, unique sequence using custom, synthetic primers and Taq DNA polymerase cycle sequencing with fluorescent terminators. Each base was sequenced at least three times. Repeat sequence elements were identified using the RepeatMasker program.2

RESULTS

The Structure of the Human Class Mu Glutathione S-Transferase Gene Cluster—In an earlier paper (14), we reported the identification of three yeast artificial chromosome (YAC) clones that contain human class Mu GST genes. Locus-specific PCR primers were used to show that GSTM-YAC1 contains GSTM1, GSTM3, GSTM5; GSTM-YAC2 contains all five members of the class Mu family; and GSTM-YAC3 contains GSTM2 and GSTM4. The observation that all five class Mu GST genes are contained within the GSTM-YAC2 clone indicated that all five human class Mu GST genes are located on a single chromosome within the 600-kb insert. Fluorescent in situ hybridization was used to map GSTM-YAC2, and thus the human class Mu GST gene cluster, to chromosome 1p13.3 (14).

As a first step toward mapping the end points of the GSTM1 deletion in humans, we compared the organization of the class Mu glutathione S-transferase genes in GSTM-YAC1, -YAC2, and -YAC3 to their organization in human genomic DNA. Because two of the three YAC clones (GSTM-YAC1 and GSTM-YAC3) contain only a portion of the GSTM2 gene cluster, they can be used to determine the order and orientation of the genes. We first compared the sizes of HindIII and EcoRI restriction fragments containing class Mu glutathione S-transferase genes that cross-hybridized with the GSTM1 cDNA clone GTH411 (22) in GSTM-YAC1, -YAC2, and -YAC3 to restriction fragments in human DNAs. Fig. 1A shows the pattern of hybridization seen with HindIII digests of the YAC and human DNAs. Locus-specific PCR primers were used to show that GSTM-YAC1 contains GSTM1, GSTM3, and GSTM5; GSTM-YAC2 contains all five members of the class Mu family; and GSTM-YAC3 contains GSTM2 and GSTM4. The observation that all five class Mu GST genes are contained within the GSTM-YAC2 clone indicated that all five human class Mu GST genes are located on a single chromosome within the 600-kb insert. Fluorescent in situ hybridization was used to map GSTM-YAC2, and thus the human class Mu GST gene cluster, to chromosome 1p13.3 (14).

DNA Sequencing and Sequence Assembly—Subcloned fragments were either sequenced manually using a Taq Cycle Sequencing kit (Amersham) or were sequenced on an ABI PRISM 377 sequencer with the ABI PRISM Dye Terminator Cycle Sequencing Ready Reaction kit at the University of Virginia Sequencing Center. Sequence data were further manipulated at the Human Genome Center at University of Oklahoma. At the University of Oklahoma, cosmid DNAs were isolated free from E. coli host contamination by the cleared lysate, diatomaceous earth-based procedure describe earlier (17) and sequenced to a level of 4.5-fold redundancy via the previously described, double-stranded, shotgun-based approach (18) using the ABI PRISM fluorescence-labeled terminators and either forward or reverse universal primers. Sequencing vector regions were removed, and the resulting data were assembled into contiguous fragments initially using the TED and XGAP programs (19) and more recently using the Phred, Phrap, and Consed programs.2 The individual contigs were joined into a final, unique sequence using custom, synthetic primers and Taq DNA polymerase cycle sequencing with fluorescent terminators. Each base was sequenced at least three times. Repeat sequence elements were identified using the RepeatMasker program.2

RESULTS

The Structure of the Human Class Mu Glutathione S-Transferase Gene Cluster—In an earlier paper (14), we reported the identification of three yeast artificial chromosome (YAC) clones that contain human class Mu GST genes. Locus-specific PCR primers were used to show that GSTM-YAC1 contains GSTM1, GSTM3, and GSTM5; GSTM-YAC2 contains all five members of the class Mu family; and GSTM-YAC3 contains GSTM2 and GSTM4. The observation that all five class Mu GST genes are contained within the GSTM-YAC2 clone indicated that all five human class Mu GST genes are located on a single chromosome within the 600-kb insert. Fluorescent in situ hybridization was used to map GSTM-YAC2, and thus the human class Mu GST gene cluster, to chromosome 1p13.3 (14).

As a first step toward mapping the end points of the GSTM1 deletion in humans, we compared the organization of the class Mu glutathione S-transferase genes in GSTM-YAC1, -YAC2, and -YAC3 to their organization in human genomic DNA. Because two of the three YAC clones (GSTM-YAC1 and GSTM-YAC3) contain only a portion of the GSTM2 gene cluster, they can be used to determine the order and orientation of the genes. We first compared the sizes of HindIII and EcoRI restriction fragments containing class Mu glutathione S-transferase genes that cross-hybridized with the GSTM1 cDNA clone GTH411 (22) in GSTM-YAC1, -YAC2, and -YAC3 to restriction fragments in human DNAs. Fig. 1A shows the pattern of hybridization seen with HindIII digests of the YAC and human DNAs. Locus-specific PCR primers were used to show that GSTM-YAC1 contains GSTM1, GSTM3, and GSTM5; GSTM-YAC2 contains all five members of the class Mu family; and GSTM-YAC3 contains GSTM2 and GSTM4. The observation that all five class Mu GST genes are contained within the GSTM-YAC2 clone indicated that all five human class Mu GST genes are located on a single chromosome within the 600-kb insert. Fluorescent in situ hybridization was used to map GSTM-YAC2, and thus the human class Mu GST gene cluster, to chromosome 1p13.3 (14).

As a first step toward mapping the end points of the GSTM1 deletion in humans, we compared the organization of the class Mu glutathione S-transferase genes in GSTM-YAC1, -YAC2, and -YAC3 to their organization in human genomic DNA. Because two of the three YAC clones (GSTM-YAC1 and GSTM-YAC3) contain only a portion of the GSTM2 gene cluster, they can be used to determine the order and orientation of the genes. We first compared the sizes of HindIII and EcoRI restriction fragments containing class Mu glutathione S-transferase genes that cross-hybridized with the GSTM1 cDNA clone GTH411 (22) in GSTM-YAC1, -YAC2, and -YAC3 to restriction fragments in human DNAs. Fig. 1A shows the pattern of hybridization seen with HindIII digests of the YAC and human DNAs. Locus-specific PCR primers were used to show that GSTM-YAC1 contains GSTM1, GSTM3, and GSTM5; GSTM-YAC2 contains all five members of the class Mu family; and GSTM-YAC3 contains GSTM2 and GSTM4. The observation that all five class Mu GST genes are contained within the GSTM-YAC2 clone indicated that all five human class Mu GST genes are located on a single chromosome within the 600-kb insert. Fluorescent in situ hybridization was used to map GSTM-YAC2, and thus the human class Mu GST gene cluster, to chromosome 1p13.3 (14).

As a first step toward mapping the end points of the GSTM1 deletion in humans, we compared the organization of the class Mu glutathione S-transferase genes in GSTM-YAC1, -YAC2, and -YAC3 to their organization in human genomic DNA. Because two of the three YAC clones (GSTM-YAC1 and GSTM-YAC3) contain only a portion of the GSTM2 gene cluster, they can be used to determine the order and orientation of the genes. We first compared the sizes of HindIII and EcoRI restriction fragments containing class Mu glutathione S-transferase genes that cross-hybridized with the GSTM1 cDNA clone GTH411 (22) in GSTM-YAC1, -YAC2, and -YAC3 to restriction fragments in human DNAs. Fig. 1A shows the pattern of hybridization seen with HindIII digests of the YAC and human DNAs. Locus-specific PCR primers were used to show that GSTM-YAC1 contains GSTM1, GSTM3, and GSTM5; GSTM-YAC2 contains all five members of the class Mu family; and GSTM-YAC3 contains GSTM2 and GSTM4. The observation that all five class Mu GST genes are contained within the GSTM-YAC2 clone indicated that all five human class Mu GST genes are located on a single chromosome within the 600-kb insert. Fluorescent in situ hybridization was used to map GSTM-YAC2, and thus the human class Mu GST gene cluster, to chromosome 1p13.3 (14).
The GTH411 cDNA probe cross-hybridizes to GSTM1 gene (14). The truncation of GSTM1 in GSTM-YAC1 by the pYAC4 vector implies that GSTM1 is at one end of GSTM-YAC1 insert and thus it must be situated in the middle of the entire class Mu GST gene cluster, separating GSTM3 and GSTM5 from GSTM2 and GSTM4. Moreover, the 11.4-kb HindIII band is derived from the 5’-end of GSTM1 (Fig. 2B). Thus, GSTM1 is oriented with its 5’-end toward GSTM2 and GSTM4.

By comparing the restriction enzyme digestion pattern in GSTM-YAC3 with that of GSTM-YAC2, we inferred the order of GSTM4 and GSTM2 with respect to the remaining members of the cluster. The GSTM4 gene contains a rare cut SmaI site in intron 2 (25). To estimate the distance between GSTM4 and GSTM2, YAC DNA sequences were digested with SmaI and hybridized with GTH411 probe (data not shown). Only a strong 18-kb band and a very faint 1.5-kb band appeared in GSTM-YAC3 DNA, suggesting that the majority of both GSTM4 and GSTM2 are contained in a 18-kb DNA segment. The cDNA sequence of GSTM2 (26) contains a rare cut NruI site, as does the published portion of its genomic sequence (27). GSTM-YAC3 contains an 8.0-kb NruI fragment rather than the full-length 10.5-kb fragment found in GSTM-YAC2 (data not shown). Hybridization with probe for the pYAC4 vector showed that the aberrant 8.0-kb band is due to the truncation of the original 10.5-kb band by the vector-insert junction. Since GSTM4 and GSTM2 are about 18 kb apart, the NruI digestion result suggested that the GSTM2 is located next to the vector-insert junction in GSTM-YAC3, since the opposite order would result in a normal GSTM2-containing NruI band. The 5.0-kb abnormal GSTM2-containing HindIII fragment in GSTM-YAC3 was subsequently shown to hybridize only with the 3′-half of the GTH411 cDNA probe (Fig. 2C). Together, the data suggest that GSTM2 is adjacent to GSTM1, since it is truncated at the insert-vector junction; furthermore, the observation that the aberrant 4.9-kb HindIII band contains only the 3′-portion of GSTM1 allele and one homozygous for the GSTM1-0 deletion are shown. The GTH411 cDNA probe cross-hybridizes to GSTM4, GSTM2, GSTM5, as well as GSTM1, but GTH411 does not cross-hybridize efficiently to GSTM3 under the conditions used (data not shown). The pattern of hybridization in the GSTM-YAC2 lane is identical to the pattern in the GSTM1+/− individual, with the exception of a 1.5-kb HindIII band that has been assigned to chromosome 3 (14). Hybridization to the GSTM1+/− sample differs at three locations; the 11.4-kb (this band was reported as 12.5 kb in earlier publications (14, 23), but our current map and sequence data have refined the size to 11.4 kb) and 2.2-kb bands associated with GSTM1 are absent in this sample, and an additional 7.6-kb band is present. This latter band is the result of a HindIII restriction fragment length polymorphism at the 5’-end of the GSTM5 gene (14).

In contrast to GSTM-YAC2, which contains all five class Mu glutathione S-transferase genes, both GSTM-YAC1 and GSTM-YAC3 contain restriction fragments that are not found in human genomic DNA. Hybridization with pBR322, which cross-hybridizes with the YAC vector sequence (Fig. 1B), shows that the novel 9.2-kb HindIII band in GSTM-YAC1 and the novel 4.9-kb HindIII band in GSTM-YAC3, which are not present in GSTM-YAC2 or human genomic DNA, contain both human (Fig. 1A) and YAC vector DNA (Fig. 1B). Thus, those two bands reflect a truncation of the human genomic HindIII fragment at the insert-vector junction.

The absence of other abnormal bands demonstrates that the organization of the class Mu GST gene cluster was unaltered during the YAC cloning process. This was confirmed by probing an EcoRI digestion of YAC and human DNA samples. The YAC library from which our three clones were isolated was constructed from a partial EcoRI digestion and ligated into the EcoRI site of the pYAC4 vector; no novel bands are expected. All of the EcoRI bands in the three class Mu YAC clones match human genomic DNA EcoRI fragments from the individual with the GSTM1+/− genotype (Fig. 1C). The absence of the 7.4-kb EcoRI band (this band was referred to as 8 kb in Refs. 14 and 24) in the DNA from an individual with the GSTM1+/− genotype results from the GSTM1 deletion. The extra 15-kb EcoRI band shown only in human genomic DNA may be from the class Mu GST sequence on chromosome 3.

To determine the general organization of the gene cluster, we exploited the observation that both GSTM-YAC1 and GSTM-YAC3 have a class Mu GST gene flanked immediately by a pYAC4 sequence as indicated by the aberrant 9.2-kb (GSTM-YAC1) and 4.9-kb (GSTM-YAC3) HindIII fragments. Since GSTM-YAC1 lacks the GSTM2 and GSTM4 genes (14), and the GTH411 probe used in Fig. 1 does not cross-hybridize with GSTM3, the 9.2-, 6.5-, and 2.2-kb HindIII fragments in GSTM-YAC1 must be derived from GSTM1 and GSTM5. The 2.2-kb HindIII band in Fig. 1A contains a portion of the GSTM1 gene as mentioned above; the 6.5-kb band is from the GSTM5 gene (14). The 9.2-kb band is thus derived from the 11.4-kb HindIII band that also contains GSTM1. The truncation of GSTM1 in GSTM-YAC1 by the pYAC4 vector implies that GSTM1 is at one end of GSTM-YAC1 insert, and thus it must be situated in the middle of the entire class Mu GST gene cluster, separating GSTM3 and GSTM5 from GSTM2 and GSTM4. Moreover, the 11.4-kb HindIII band is derived from the 5’-end of GSTM1 (Fig. 2B). Thus, GSTM1 is oriented with its 5’-end toward GSTM2 and GSTM4.

By comparing the restriction enzyme digestion pattern in GSTM-YAC3 with that of GSTM-YAC2, we inferred the order of GSTM4 and GSTM2 with respect to the remaining members of the cluster. The GSTM4 gene contains a rare cut SmaI site in intron 2 (25). To estimate the distance between GSTM4 and GSTM2, YAC DNA sequences were digested with SmaI and hybridized with GTH411 probe (data not shown). Only a strong 18-kb band and a very faint 1.5-kb band appeared in GSTM-YAC3 DNA, suggesting that the majority of both GSTM4 and GSTM2 are contained in a 18-kb DNA segment. The cDNA sequence of GSTM2 (26) contains a rare cut NruI site, as does the published portion of its genomic sequence (27). GSTM-YAC3 contains an 8.0-kb NruI fragment rather than the full-length 10.5-kb fragment found in GSTM-YAC2 (data not shown). Hybridization with probe for the pYAC4 vector showed that the aberrant 8.0-kb band is due to the truncation of the original 10.5-kb band by the vector-insert junction. Since GSTM4 and GSTM2 are about 18 kb apart, the NruI digestion result suggested that the GSTM2 is located next to the vector-insert junction in GSTM-YAC3, since the opposite order would result in a normal GSTM2-containing NruI band. The 5.0-kb abnormal GSTM2-containing HindIII fragment in GSTM-YAC3 was subsequently shown to hybridize only with the 3′-half of the GTH411 cDNA probe (Fig. 2C). Together, the data suggest that GSTM2 is adjacent to GSTM1, since it is truncated at the insert-vector junction; furthermore, the observation that the aberrant 4.9-kb HindIII band contains only the 3′-portion of GSTM1.
GSTM2 indicates that GSTM2 is oriented with its 3′-end toward the insert-vector junction. Thus, GSTM2 is oriented with its 3′-end toward GSTM1. This interpretation was confirmed by sequencing a PCR-generated junction fragment spanning from the 3′-end of GSTM2 to the pYAC4 vector sequence in GSTM-YAC3 (data not shown). These data show that the orientation of the gene cluster in GSTM-YAC2 is **GSTM4–5′-GSTM2–3′-GSTM1–3′-GSTM5**.

To limit our search for the region deleted in the GSTM1-0 allele, we assigned each HindIII band in Fig. 1 to a specific GSTM gene. Three strategies were used in the assignment: 1) comparison of GSTM-YAC1 and GSTM-YAC3 (above); 2) polymorphisms in the GSTM1 and GSTM5 genes; and 3) 5′- and 3′-specific hybridization probes to determine the orientation of the class Mu genes (Fig. 2).

Hybridization patterns of the HindIII and EcoRI digestions of human genomic DNA with a GSTM1 cDNA probe show polymorphisms in only two of the four class Mu genes that hybridize with this probe. The GSTM1 deletion removes 11.4- and 2.2-kb HindIII fragments and a 7.4-kb EcoRI fragment. In addition, a HindIII restriction fragment length polymorphism is found near GSTM5 (14). To assign each fragment in the HindIII digestion, the 5′-half and 3′-half of the GSTM1-cDNA probe were used to probe HindIII digests of GSTM-YAC1–3 (Fig. 2). The 5′- and 3′-subclones divide GTH411 at a BglII site at residue 593 of the 1117-nucleotide insert, which is 10 nucleotides into exon 8 of the GSTM1 gene.

Of the six HindIII bands that hybridize with the complete GTH411 probe, H2 (11.4 kb) and H6 (2.2 kb) can be assigned to GSTM1 based on polymorphism, and H3 (6.5 kb) can be assigned to GSTM5 because it is present in GSTM-YAC1 and -YAC2 but not in GSTM-YAC3. Assignment of GSTM2 and GSTM4 is more complex, but H5 (5.2 kb in GSTM-YAC2, 4.9 kb in GSTM-YAC3) can be assigned to GSTM2 because GSTM2 is nearest to the GSTM-YAC3 vector junction. H1 (13.6 kb) can be assigned to GSTM4 and GSTM2 (see below); by process of elimination, H4 (5.3 kb) is assigned to GSTM5. The complete assignment of each band shows that there are no other unidentified class Mu GST genes in the GSTM-YAC2 that share enough sequence similarity with GSTM1 to be detected by the GTH411 probe.

Hybridizations with separate 5′- and 3′-portions of GSTM1 confirm that the 11.4-kb H2 fragment contains the 5′-end of GSTM1, while the 2.2-kb H6 fragment contains the 3′-end of GSTM1. Since the 13.6-kb H1 fragment hybridizes with both parts of GSTM1 and the 3′-end of GSTM2 is found on H5, H1 must contain both the 3′-end of GSTM4 gene and 5′-end of the GSTM2 gene (Fig. 2B), suggesting that the 3′-end of GSTM4 is oriented toward GSTM2. The 13.6-kb GSTM2-M4 intergenic distance is consistent with the estimate obtained from the Smal digestion. The distance between GSTM2 and GSTM1 was estimated using the rare cut SacII enzyme, which has a cutting site in the exon 1 of both the GSTM2 and GSTM1 cDNA sequences. GSTM-YAC2 contains a unique 20-kb band, which is not present in either GSTM-YAC1 or GSTM-YAC3 DNA (data not shown). The absence of the 20-kb band in GSTM-YAC1 and GSTM-YAC3 suggests this fragment is from GSTM2-GSTM1 intergenic region, since only GSTM-YAC2 contains this entire region.

The general organization of class Mu GST gene cluster can be inferred from these data (Fig. 2D). The class Mu GST gene cluster is organized as 5′-GSTM4–GSTM2–GSTM1–GSTM5–3′, with all four genes organized in the same tail-to-head orientation. (The orientation of the GSTM5 was determined by sequencing the end of the cosmID cDNA insert as described below.) GSTM4 and GSTM2 are separated by about 15 kb; GSTM2 and GSTM1 are separated by about 20 kb.

This assignment of class Mu genes to specific restriction fragments shows that neither the 3′-end of the GSTM2 gene (5.2-kb fragment H5) nor the 5′-end of the GSTM5 gene (6.5-kb fragment H3) are rearranged in the GSTM1-0 deletion. Thus, the GSTM1 deletion end points are in the intragenic regions between GSTM2 and GSTM1 and between GSTM1 and GSTM5. This excludes one simple model for the GSTM1 dele-
tion, that the GSTM1 deletion results from an unequal homologous recombination event between exon 8 of the GSTM2 gene and exon 8 of the GSTM1 gene. These two exons are more than 99% identical over 583 nucleotides (26) and a recombination between these two regions would produce a chimeric GSTM2/M1 gene whose mRNA and protein products would be indistinguishable from those of an unrecombined GSTM2 gene. The assignment of the 5.2-kb H5 fragment to GSTM2 and the 2.2-kb H6 fragment to the 3′-end of GSTM1 excludes this possibility, since a recombination within exon 8 of the two genes would cause both a deletion of the H5 fragment (which might be difficult to detect, since there is a similar size 5.3-kb H4 from GSTM4) and conservation of the 2.2-kb H6 fragment, which after the deletion would be placed at the end of the recombinant GSTM2 gene. Since the only HindIII restriction fragment polymorphisms associated with the GSTM1 deletion are the loss of H2 (11.4 kb) and H6 (2.2 kb), we conclude that the deletion does not involve a recombination with either flanking class Mu glutathione S-transferase gene.

A Physical Map of the Class Mu Glutathione S-Transferase Cluster—To examine the deletion end points in the intragenic regions flanking GSTM1, we constructed a cosmids library from the GSTM-YAC2 clone. The cosmim library contains about 3 × 10^6 independent colonies with an average insert size of about 35 kb, equivalent to about 700-fold coverage of a haploid yeast genome. The library was screened initially with a probe hybridizing with the 3′-sequence immediately flanking both GSTM2 and GSTM1. Positive clones were confirmed by hybridization of the GTH4111 cDNA clone to HindIII and EcoRI digests. Two overlapping clones, cGTM1 (38 kb) and cGTM12 (36 kb), were used for the subsequent mapping of the gene cluster.

EcoRI and HindIII sites in cGTM1 and cGTM12 were mapped using partial digestion and indirect end labeling with T3 and T7 oligonucleotides that hybridize adjacent to the pWE15 cloning site (data not shown). The accuracy of the map (Fig. 3) in the region flanking GSTM1 was confirmed by subcloning, mapping, and sequencing the ends of several of the HindIII and EcoRI restriction fragments in the GSTM2–GSTM1 and GSTM1–GSTM5 intergenic regions during the analysis of the GSTM1 deletion break points. Direct sequencing of cGTM12 showed that one end of the cGSTM12 insert contains sequences from GSTM5 exon 5, which confirms that GSTM5 has the same gene orientation as GSTM4, M2, and M1.

Localization of the Break Points of the GSTM1 Deletion—The physical map of the GSTM gene cluster and cosmid clones cGTM1 and cGTM12 provided the reagents necessary to map the ends of the GSTM1 deletion. To localize the break points of the GSTM1 deletion, the sequences that normally flank GSTM1 were examined in DNA from a GSTM1−/− individual using hybridization probes from sequences flanking the GSTM1 (Fig. 4). To locate the 5′-boundary of GSTM1 deletion, a 9-kb fragment (P1; Fig. 4E) from cosmid cGTM1 covering the 5′-immediate flanking sequence of GSTM1 was used as a probe (Fig. 4A). This fragment contains repetitive DNA sequence elements, so the hybridization was done in the presence of human COT-1-blocking DNA. Comparison of the hybridization patterns of the EcoRI fragments from the GSTM1+/−/− individuals and from the GSTM1−/− individual indicates that the GSTM1−/− individual lacks 5′-flanking 3.5- and 1.8-kb EcoRI fragments, although the 1.8-kb fragment is difficult to see in the reproducion. This result suggests that the 5′-break point is either within the 1.8-kb fragment or farther upstream. However, since the 5.2-kb HindIII band containing the 3′-end of GSTM2 is intact in the GSTM1−/− individual (Fig. 1A, Fig. 2), the 5′-break point was localized to a 3.4-kb region consisting of a 1.6-kb HindIII–EcoRI fragment and the 1.8-kb EcoRI fragment.

To examine the 3′-break point of GSTM1 deletion, a 5.2-kb EcoRI fragment was subcloned from cosmid cGTM12 and used as a probe (P2; Fig. 4). The result (Fig. 4B) shows no alteration of the 5.2-kb EcoRI fragment in DNA from the GSTM1−/− individual. Thus, the 3′-break point appears to be located 5′ to the 5.2-kb EcoRI fragment. The hybridization pattern also shows that P2 cross-hybridizes with the 1.8- and 3.5-kb EcoRI fragments on the 5′-side of the GSTM1 gene. This was further confirmed when the cross-hybridization was shown directly with cGTM1 DNA (data not shown).

Probe P1 cross-hybridizes with 5.0- and 5.2-kb EcoRI fragments on the 3′-side of GSTM1 and the 5.0-kb EcoRI fragment on the 5′-side of GSTM1, making it difficult to interpret the data in Fig. 4. To simplify the interpretation, we subcloned and purified a 1.6-kb HindIII–EcoRI fragment (P3; Fig. 4) from cGTM1. In addition to hybridizing to the expected 5.0-kb EcoRI fragment (data not shown) and the 1.6-kb EcoRI/HindIII in cGTM1 and GSTM-YAC2 (Fig. 4C), P3 cross-hybridizes with the 4.5-kb HindIII–EcoRI fragment on the 3′-side of the GSTM1 gene in cGTM12 (data not shown). The 1.6-kb HindIII–EcoRI fragment is unaltered in the GSTM1−/− genotype (Fig. 4C), which is within the 3.4-kb (1.8 plus 1.6 kb) zone containing the 5′-break point of the deletion inferred from the results in Figs. 2A and 4A. P3 does not cross-hybridize with the 4.5-kb HindIII–EcoRI fragment on the 3′-side of the GSTM1 gene in the DNA from a GSTM1−/− genotype (Fig. 4C), suggesting that either the 3′-break point is in the 4.5-kb HindIII–EcoRI fragment or 3′ to the fragment.

The experiments in Fig. 4, A–C, do not show a novel deletion junction fragment that is present in human DNA from either the GSTM1+/− or GSTM1−/− individuals and not in GSTM-YAC2. These observations, together with the sequence similarity implied by cross-hybridization between the 1.6-kb (P3) frag-

![Fig. 3. A physical map of the class Mu glutathione S-transferase gene cluster.](image-url)
GSTM1 Gene Deletion

FIG. 4. Localization of the GSTM1 deletion break points. Restriction fragment probes flanking the GSTM1 gene were used to identify differences in cross-hybridizing restriction fragments from GSTM1+/+ and GSTM1−/− individuals. Hybridizations in panels A and B were done in the presence of human repetitive COT-1 DNA to reduce repetitive sequence background. A, hybridization of P1, a 9-kb HindIII fragment from one end of the cGTM1 insert, with EcoRI digests of GSTM-YAC2 and human genomic DNAs. B, hybridization of P2, a 5.2-kb EcoRI fragment from clone cGTM12, with EcoRI digests of GSTM-YAC2 and human genomic DNAs. The 4.5-kb band in the GSTM-YAC2 and GSTM1+/− lanes corresponds to a 4.5-kb HindIII/EcoRI double digest fragment near the 3′-end of GSTM1. C, hybridization of P3, a 1.6-kb EcoRI/HindIII fragment from P1, with an EcoRI/HindIII digest of GSTM-YAC2 and human genomic DNAs. The 4.5-kb band in the GSTM-YAC2 and GSTM1+/− lanes corresponds to a 4.5-kb HindIII/EcoRI double digest fragment near the 3′-end of GSTM1. D, hybridization of P3 with a HindIII digest of GSTM-YAC2 and human genomic DNA. E, a restriction map of the GSTM2–GSTM1–GSTM5 region showing the locations of the P1, P2, and P3 probes.

ment from the 5′-side of GSTM1 and the 4.5-kb HindIII–EcoRI fragment on the 3′-side of GSTM1 and between the 1.8- and the 3.5-kb EcoRI fragments on the 5′-side and the 5.2-kb EcoRI fragment on the 3′-side of GSTM1 strongly suggest that the deletion could have arisen from an unequal homologous recombination in the regions depicted in Fig. 5B. The 5′-break point is located within a 3.4-kb region whose boundary is defined on the left by a HindIII site and the right by an EcoRI site. If the recombination is caused by homologous unequal crossing over, the 3′-break point should be located in the corresponding region on the 3′-side of the GSTM1 gene. A recombination 5′ to the HindIII site at the left boundary of the left junction region (the smaller shaded box between M2 and M1 in Fig. 5B) would eliminate the 1.6-kb fragment that is seen in Fig. 4C, since a corresponding HindIII site is not present on the 3′-side of GSTM1. If the 5′-break point were closer to the GSTM1 than the right EcoRI site at the right boundary of the left repeat/junction region, a homologous recombination would not remove the 1.8- and 3.5-kb EcoRI fragments that are absent (Fig. 4A) in a GSTM1−/− genome.

The recombination diagrammed in Fig. 5B predicts the production of a novel 7.4-kb HindIII fragment from the GSTM1−/− allele that should hybridize with the 1.6-kb P3 probe. This band can be seen in Figs. 4D and 5A, which display the human genomic HindIII fragments that hybridize with the 1.6-kb P3 probe. Fig. 4D shows that human DNA samples contain a 7.4-kb HindIII fragment that is not seen in GSTM-YAC2. In individuals homozygous for the GSTM1-0 deletion, the appearance of the 7.4-kb band is associated with a loss of two HindIII bands at 10.3 and 11.4 kb. (The 13.6-kb HindIII bands in Fig. 4D and 5A are due to cross-hybridization between P3 and sequences 3′ to GSTM4 on fragment H1.) Heterozygous GSTM1+/− individuals display the 10.3- and 11.4-kb bands, which are seen in GSTM-YAC2, and also the novel 7.4-kb band. The intensity of the 7.4-kb band is about twice as strong in the GSTM1−/− individual as in the GSTM1+/− individual (Fig. 4D).

To examine further the homogeneity of the deletion junction in the population, additional DNA samples were studied. Fig. 5A shows that all but one of the DNA samples have the 7.4-kb HindIII junction fragment. One of the GSTM1+ individuals lacks the 7.4-kb band, suggesting that her genotype is GSTM1+/+. This confirms that the 7.4-kb band reflects a recombination on chromosome 1 and not a cross-hybridizing fragment that is missing from GSTM-YAC2 but is found elsewhere in the genome. Thus, the 1.6-kb HindIII–EcoRI fragment probe can distinguish GSTM1+/+ genotype from a GSTM1−/− genotype and should be useful in case/control studies of GSTM1 genotypes and cancer.

The Sequence of the GSTM1 Deletion Junction Region—To confirm the homologous recombination model and to locate the deletion end points, we sequenced the left and right junction segments predicted by the restriction mapping to be involved in the homologous recombination. The sequence of the left junction region (Fig. 6) was determined from subclones of cGTM1 and thus cannot contain sequences from the right recombination region. Sequences for the right recombination region were obtained from subclones of the 11.4-kb HindIII fragment from the 3′-flanking region of cGTM12 and thus cannot be contaminated with sequence from the 5′-flanking region of GSTM1. To locate the deletion break points, we also cloned and sequenced a portion of 7.4-kb HindIII deletion junction fragment from the GSTM1−/− individual shown in Fig. 4.

The sequences of the left, right, and deletion junction regions are shown in Fig. 6. The sequence shown extends from the position of the HindIII site on the left (GSTM2) side of the left junction region to the position of the EcoRI site on the right (GSTM1) side of the left junction region. These two restriction sites are diagnostic for the left and right junction regions. Both sites are present in the left junction region, and both are absent from the right junction region. However, the left junction diagnostic HindIII site is found in the deletion junction fragment, while the diagnostic EcoRI site is not. Thus, the left end of the deletion junction matches the left junction region, while the right end of the deletion junction matches the right junction region, suggesting that recombination took place to the right of the HindIII site and to the left of the EcoRI site in the left and right junction regions. The patterns of conservation between the left junction and the deletion junction sequence on the left and the right junction and the deletion junction on the right is summarized graphically in Fig. 7.

The left and right junction (unrecombined) regions share 98.3% identity over the 3355 nucleotides between the HindIII and EcoRI sites at the ends of the sequences in Fig. 6. The left
Fig. 5. A novel restriction fragment produced by the GSTM1 deletion/recombination. A. HindIII digests of human genomic DNA (15 μg) hybridized with the 1.6-kb P3 probe from Fig. 4. The GSTM1 genotype of the individual (+/+, +/-, or -/-) is indicated. GSTM1 +/- genotypes were determined by HindIII digestion and hybridization with the GSTM1 cDNA probe GTH411 (14). GSTM1 +/- and GSTM1 -/- genotypes were distinguished by the presence of a novel 7.4-kb HindIII band that is found only in individuals carrying the GSTM1 - (null) allele. B. The proposed homologous recombination that results in the deletion of the GSTM1 gene and flanking DNA between the left and right repeated regions. The light gray boxes between M2 and M1 and between M1 and M5 indicate the locations of the shaded recombination junction regions in Figs. 6 and 7. Sizes of the EcoRI/HindIII fragments surrounding the deletion junctions are shown. The open triangles flanking the right repeat junction region indicate HindIII (left, △) and EcoRI (right, ▽) sites that flank the left repeat junction but are absent from the right repeat junction. The position of the EcoRI site that is absent from the recombined junction fragment is also marked (▽) on the lower map of the recombined GSTM1-0 allele.

The very high level of identity within this 2.3-kb region and the absence of a consistent pattern of differences between the left or right junction sequence and the deletion junction sequence makes it impossible to determine precisely where within this region the recombination between the left and right junction regions occurred. The boundaries of the 2.3-kb recombination region are defined by the reduced identity between the left or right junction region and the deletion junction sequence. This can be more clearly seen in Fig. 7A and B, where the average percentage of identity in a 20-nucleotide window is plotted for comparisons of the left and right junction sequences with the deletion junction sequence. Fig. 7C shows a more schematic summary of the alignment data: circles indicate differences between the left junction region and the deletion junction sequence, while filled squares indicate differences between the right junction region and the deletion junction sequence.

The left boundary of the recombination region is easily identified because of the large number of differences between the right and deletion junction region to the left of nucleotide 660. The right boundary of the cross-over zone is more subtle, since sequences are very similar in that region between the left and right repeats. We assigned the right junction boundary by comparing nucleotide composition at four diagnostic positions where the nucleotides show a difference between those in the left repeat and those in right repeat. The left repeat has C, T, C, and T, at positions 2957, 2977, 3110, and 3353, respectively, while the right repeat has A, G, T, and A at those positions. In the deletion junction fragment, the pattern is A, G, T, and A, matching the right repeat completely and suggesting that the right boundary of the cross-over zone is probably located 5’ to nucleotide position 2957.

Beyond the right end of the sequences shown in Figs. 6 and 7, the similarity between the left and right junction regions drops off to about 90% sequence identity for about 1.2 kb and then to less than 60% identity about 2.5 kb beyond the EcoRI site at the end of Figs. 6 and 7 (data not shown). Thus, there is a longer left and right junction repeated region that extends for about 4.2 kb. The left end of this repeated junction region begins with the reduced similarity to the left of nucleotide 660 in Figs. 6 and 7. Our estimate of the right boundary of the recombination region is based on a consistent pattern of identity at four nucleotide positions; if the identity between the junction fragment sequence and the right junction sequence is fortuitous, the recombination region might include this entire 4.2-kb duplicated region. The localization of the homologous recombination cross-over site in the indicated 2.3-kb region agrees completely with the result obtained from the deletion mapping analysis based on Southern blotting, since the left and right 2.3-kb segments are contained totally within the homologous recombination regions defined by Southern analysis (Fig. 4).

The 2.3-kb recombination subregion in the 4.2-kb repeat contains a reverse oriented Alu element (nucleotides 2885–2584) and a reverse MER20 DNA transposon (nucleotides 1557–1757; Fig. 6, Ref. 28). In addition, a forward MLT2B2 retroviral long terminal repeat is located just upstream from the recombination region (nucleotides 92–541). The Alu contains a poly(A) tail (nucleotides 2613–2584) and is abutted by short direct repeats at its two ends (TTACCTAA), with a small truncation at the 5’-end of the Alu element. The 2.3-kb cross-over zone is not particularly AT-rich (A = 1 = T = 56.5%). The Southern blot in Figs. 4D and 5A suggest that, except for the Alu repeated sequence element, the 2.3-kb junction region is present only twice in the genome on the 5’- and 3’-intergenic sequence flanking the GSTM1 gene.

Computer analysis did not reveal any overrepresentation of either inverted repeats or direct repeats, which are often found near deletion break points, in the 2.3-kb region. In addition, the region does not encode any sequences with significant similarity to protein sequence data bases. With the exception of the three repeated elements, the 2.3-kb region does not share significant similarity with other DNA sequences in GenBank™.

Discussion

We have determined a physical map of the GSTM1, GSTM2, GSTM4, and GSTM5 class Mu glutathione S-transferase genes
on human chromosome 1p13.3 and have identified the left and right junction regions involved in the unequal crossing over event that produces the GSTM1-0 deletion. The human class Mu GST gene cluster contains two almost identical 4.2-kb regions that flank the GSTM1 gene; the GSTM1-0 deletion is caused by a homologous recombination involving the left and right 4.2-kb repeats. However, extensive sequence identity between the left and right repeats conceals the exact break points within a 2.3-kb region (Figs. 6 and 7). Sequences of deletion junctions from additional individuals may narrow the zone. Most GSTM1 deletions appear to be caused by the same homologous recombination, since all 20 null allele chromosomes (3 in Fig. 4 and 17 in Fig. 5) that we examined have the same 7.4-kb HindIII junction fragment.

Gene deletion by homologous unequal crossing over is now well documented (29–31), including a deletion of a member of the cytochrome P450 detoxification gene family (32). Striking features of the GSTM1 deletion are its high frequency in the population and the apparent homogeneity of the recombination region. These features may reflect an ancient acquisition of the population and the apparent homogeneity of the recombination process in the formation of the human class Mu GST gene cluster, which probably occurred more than 20 million years ago (23). If the 4.2-kb repeats arose at that time, then their high sequence identity is not the result of a very recent sequence duplication. The regions that flank the left and right repeat share about 92% sequence identity, as opposed to >99% identity within the two repeats (Fig. 7). The 8% sequence divergence in the regions flanking the repeats (i.e. 4% divergence of each region from the duplicated ancestral sequence) suggests that the surrounding region was duplicated about 30 million years ago (4%/0.15%, assuming a divergence drift rate of 0.15% per million years for noncoding regions; Ref. 33). Thus, the class Mu gene cluster may have been duplicated or rearranged between the divergence of Old World monkeys and apes about 25 million years ago and New World monkeys and apes about 40 million years ago.

We searched the GenBank DNA and expressed sequence tag data bases and various protein data bases for possible coding sequences from the 4.2-kb repeated region that might account for its high sequence conservation. In addition to the repetitive elements present in the 4.2-kb repeat, a search of the

---

**Fig. 7. The structure of the deletion junction region.** The sequence data from Fig. 6 are summarized graphically. Nucleotide 1 on this graph corresponds to the HindIII site marked with an asterisk in Figs. 3 and 4 and the beginning of the sequences in Fig. 6. The shaded region in this figure corresponds to the shaded region in Fig. 6. Plots in A and B were produced with the GCG PLOTsimilarity program from the alignment in Fig. 6. The graph shows the fraction identical over a 20-nucleotide window; the smallest change in the graph corresponds to one nucleotide substitution in 20 residues. A, alignment of the left (GSTM2–GSTM1) nondeleted repeat region and the deletion junction region. B, alignment between the right (GSTM1–GSTM5) nondeleted repeat region and the deletion junction region. C, a schematic that indicates each of the nucleotide differences between the left region and the deletion junction region (○) or between the right region and the deletion junction region (□). The positions of the plot boundaries with respect to the GSTM2 and GSTM1 genes are shown with vertical arrows. The HindIII and EcoRI restriction sites that bound the left repeat region are absent from the right repeat region and are denoted by the open triangles (○, □) in Fig. 5B.

Our data suggest that a >8-kb region from the 3′-end of GSTM2 to the 3′-end of the left 4.2-kb repeat is 90–99% identical to the corresponding region flanking the 3′-end of GSTM1. This extensive sequence conservation suggests that the two 4.2-kb repeats probably arose with the original gene duplication process in the formation of the human class Mu GST gene cluster, which probably occurred more than 20 million years ago (23). If the 4.2-kb repeats arose at that time, then their high sequence identity is not the result of a very recent sequence duplication. The regions that flank the left and right repeat share about 92% sequence identity, as opposed to >99% identity within the two repeats (Fig. 7). The 8% sequence divergence in the regions flanking the repeats (i.e. 4% divergence of each region from the duplicated ancestral sequence) suggests that the surrounding region was duplicated about 30 million years ago (4%/0.15%, assuming a divergence drift rate of 0.15% per million years for noncoding regions; Ref. 33). Thus, the class Mu gene cluster may have been duplicated or rearranged between the divergence of Old World monkeys and apes about 25 million years ago and New World monkeys and apes about 40 million years ago.

Gene deletion by homologous unequal crossing over is now well documented (29–31), including a deletion of a member of the cytochrome P450 detoxification gene family (32). Striking features of the GSTM1 deletion are its high frequency in the population and the apparent homogeneity of the recombination region. These features may reflect an ancient acquisition of the population and the apparent homogeneity of the recombination process in the formation of the human class Mu GST gene cluster, which probably occurred more than 20 million years ago (23). If the 4.2-kb repeats arose at that time, then their high sequence identity is not the result of a very recent sequence duplication. The regions that flank the left and right repeat share about 92% sequence identity, as opposed to >99% identity within the two repeats (Fig. 7). The 8% sequence divergence in the regions flanking the repeats (i.e. 4% divergence of each region from the duplicated ancestral sequence) suggests that the surrounding region was duplicated about 30 million years ago (4%/0.15%, assuming a divergence drift rate of 0.15% per million years for noncoding regions; Ref. 33). Thus, the class Mu gene cluster may have been duplicated or rearranged between the divergence of Old World monkeys and apes about 25 million years ago and New World monkeys and apes about 40 million years ago.

We searched the GenBank DNA and expressed sequence tag data bases and various protein data bases for possible coding sequences from the 4.2-kb repeated region that might account for its high sequence conservation. In addition to the repetitive elements present in the 4.2-kb repeat, a search of the

---

**Fig. 6. Left and right junction regions and the GSTM1 deletion junction region.** The sequences are presented in the same orientation as the transcription direction of the GSTM1 gene. The first sequence (left) is from the 4.2-kb repeat in the GSTM2–GSTM1 intergenic region from a nondeleted GSTM1 allele. The third sequence (right) is from the 4.2-kb repeat in the GSTM1–GSTM5 intergenic region; the second sequence (junc) is the deletion junction fragment. The portion of the junction fragment that is equally similar to the left and right repeated regions is shaded. A reverse orientated Alu repeat is indicated. The Alu element contains a poly(A) tail as shown by the presence of a poly(T) tail in the reverse orientation presented in the figure. The Alu element is flanked by 6-base pair direct repeats at its two ends. A HindIII site (marked with an asterisk in Figs. 3 and 4) and two EcoRI sites in the left repeat are marked above the sequence in the figure. The HindIII site and the second EcoRI site are the ones used to define the boundaries of the 5′-break point in the Southern blot analysis in Fig. 4. The beginning of the HindIII site was designated as nucleotide number 1 for each sequence in the figure.
among all three Alu elements, suggesting that a gene conversion event homogenized these Alu sequences as little as 5 million years ago. Multiple gene conversions may have predated the most recent gene conversion event; we cannot determine whether the 10 shared differences were the result of one conversion or accrued from multiple gene conversions. Consistent with the Alu divergence data, the total sequence divergence of less than 1% between the left and right repeats suggests that the last gene conversion in the 4.2-kb region occurred probably no more than about 3 million years ago (1%/20.15%).

Given the high frequency of the GSTM1 deletion in human populations and the evidence that gene conversion event(s) have occurred in the 4.2-kb repeat regions, we speculate that this region is a hot spot for homologous recombination. A striking feature of the two 4.2-kb repeats is their near total identity over such a long segment. Homologous recombination frequency has been shown by numerous studies to be related to the length of homology involved. Studies done in mammalian cells using a plasmid-plasmid recombination system (41, 42), as well as recombination between chromosomally inserted plasmids and native chromosome genes (20, 41), have demonstrated that 200 base pairs of uninterrupted identity is required for efficient recombination in mammalian cells, with mismatches reducing recombination efficiency. The recombining regions flanking the GSTM1 gene are considerably longer.

The GSTM1 deletion is very common in human populations, with about 50% of the individuals having the GSTM1−/− genotype, and 45% having the GSTM1+/− genotype, although the actual numbers vary somewhat in different ethnic groups (2, 21). Although our study so far has only examined 13 caucasian DNA samples, we believe the GSTM1 deletions in people from other ethnic backgrounds are also caused by the same homologous recombination. If the left and right 4.2-kb repeats are a recombination hot spot, then the high frequency of the GSTM1 deletion may reflect multiple independent occurrences of the same homologous recombination. Alternatively, if the GSTM1-0 deletion is not strongly selected against, its high frequency may be the result of an ancient ancestral deletion. Examination of the class Mu GST cluster in other primates may clarify the evolutionary history of this gene cluster and the GSTM1 deletion.

Acknowledgment—We thank Gina Calabrese for excellent technical assistance.

REFERENCES
Characterization of the Human Class Mu Glutathione S-Transferase Gene Cluster and the GSTM1 Deletion
Shi-jie Xu, Ying-ping Wang, Bruce Roe and William R. Pearson

doi: 10.1074/jbc.273.6.3517

Access the most updated version of this article at http://www.jbc.org/content/273/6/3517

Alerts:
• When this article is cited
• When a correction for this article is posted

Click here to choose from all of JBC’s e-mail alerts

This article cites 40 references, 22 of which can be accessed free at http://www.jbc.org/content/273/6/3517.full.html#ref-list-1