Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 851–855, February 1988
Genetics

A "housekeeping" gene on the X chromosome encodes a protein similar to ubiquitin

(CpG "islands"/human genes/protein domains/evolution)

DANIELA TONIOLO*, MARIA PERSICO, AND MYRIAM ALCALAY

International Institute of Genetics and Biophysics, Consiglio Nazionale delle Richerche, 80125 Naples, Italy

Communicated by Ernest Beutler, September 25, 1987

ABSTRACT **An X chromosome gene located 40 kilobases downstream from the *G6PD* gene, at Xq28, was isolated and sequenced. This gene, which we named *GdX*, spans about 3.5 kilobases of genomic DNA. *GdX* is a single-copy gene, is conserved in evolution, and has the features of a "housekeeping" gene. At its 5′ end, a cluster of CpG dinucleotides is methylated on the inactive X chromosome and unmethylated on the active X chromosome. The *GdX* gene can code for a 157 amino acid protein, GdX. Residues 1–74 of GdX show 43% identity to ubiquitin, a highly conserved 76 amino acid protein. The COOH-terminal moiety of GdX is characterized in its central part (residues 110–128) by a sequence homologous to the COOH-terminal hormonogenic site of thyroglobulin. The structural organization of the GdX protein suggests the existence of a family of genes, in addition to the ubiquitin gene, that could play specific roles in key cellular processes, possibly through protein–protein recognition.**

We have cloned 120 kilobases (kb) of DNA of the long arm of the human X chromosome flanking the glucose-6-phosphate dehydrogenase gene (*G6PD*) at Xq28 (1). In this region of the X chromosome we have characterized three clusters of CpG dinucleotides, which are differentially methylated on the active or inactive X chromosome (ref. 2; D.T., G. Martini, B. R. Migeon, and R. Dono, unpublished data). One of the CpG clusters is located at the 5′ end of *G6PD* (D.T. *et al.*, unpublished data). Since many CpG-rich "islands" have now been found in association with genes, generally "housekeeping" genes (3), we wished to determine whether the two CpG clusters 3′ from *G6PD* were at the 5′ end of neighboring genes.

More than 100 genes have been assigned to the human X chromosome. From the study of sex-linked diseases and the use of somatic cell hybrids, several linkage groups have been identified (4). One is at the end of the long arm of the chromosome and includes *G6PD* as well as the loci for hemophilia A and color blindness. However, only a small proportion of the X chromosome has been studied at the molecular level, and defective gene products have not been identified for all sex-linked diseases. The isolation and sequencing of previously unknown human genes of the X chromosome could provide insight into the normal gene and chromosome structure and function and into the nature of molecular defects leading to inherited disorders as well as common pathologies.

By hybridization of single-copy DNA probes to RNA gel blots and to cDNA libraries, we have identified two transcribed regions downstream from each of the CpG clusters. In this paper we describe the structural features of one of the two genes, *GdX*, which has been sequenced at the cDNA and genomic level.[†]

In the cDNA sequence, we identified an open reading frame (ORF) encoding a protein (GdX) of 157 amino acids. The NH_2-terminal moiety, residues 1–74, of this hypothetical protein is 43% identical to ubiquitin, a highly conserved 76 amino acid protein, unchanged in organisms as diverse as humans and plants (5, 6). The ubiquitin-like moiety of the GdX protein is followed by an 80 amino acid sequence. A similar organization has been described for some of the ubiquitin cDNAs in humans (7), *Dictyostelium* (8), and yeast (9). The COOH-terminal portion of ubiquitin, however, has no homology with the GdX COOH-terminal moiety.

MATERIALS AND METHODS

Isolation and Characterization of cDNA and Genomic Clones. The isolation of cDNA clone pGd6405 (10) and of genomic clones λGdT7, λGd11, λGd5A, and λGd3C has been described (1). The remaining GdX cDNA clones were isolated, using cDNA or genomic subclones as hybridization probes, from cDNA libraries prepared from simian virus 40 (SV40)-transformed human fibroblasts (11) or from human teratocarcinoma cells (a gift from J. Skowronski, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY). Small-scale plasmid or phage preparations were analyzed by restriction mapping according to standard procedures (12). The genomic and cDNA clones from λ libraries were subcloned in the plasmids pEMBL8 or pUC18 for further analysis.

DNA Sequencing. The nucleotide sequence of the cDNA clone pGd6405 has been published (13). The sequences of the genomic clones and of cDNA clones isolated from the SV40-transformed human fibroblast library were determined by the procedure of Maxam and Gilbert (14). The sequences of the clones from the teratocarcinoma cDNA library were determined by the dideoxy method modified to allow direct sequencing from the plasmid pUC18 (15). The sequence in Fig. 3 was obtained by sequencing both strands from genomic clones and more than one cDNA clone from different libraries.

DNA Blot Hybridization. DNA (10–20 μg) from tissues or from cultured cells, prepared as described (2), was digested with the restriction enzymes indicated, electrophoresed in 1% agarose gel, transferred to nylon filters (Zetabind, AMF-CUNO, Meriden, CT), and hybridized with probe for 18–20 hr at 65°C, as described (2). Filters were washed at 65°C in 2× SSC and 0.2× SSC (stringent washing). After hybridization of human probes with DNA from other species, filters were washed at 60°C in 2× SSC and 1× SSC. (SSC is 0.15 M NaCl/15 mM sodium citrate, pH 7.)

Abbreviations: ORF, open reading frame; SV40, simian virus 40.

*Present address: Consiglio Nazionale delle Richerche, Instituto di Genetica Biochimica ed Evoluzionistica, 27100 Pavia, Italy.

[†]The sequence reported in this paper is being deposited in the EMBL/GenBank data base (Bolt, Beranek, and Newman Laboratories, Cambridge, MA, and Eur. Mol. Biol. Lab., Heidelberg) (accession no. J03589).

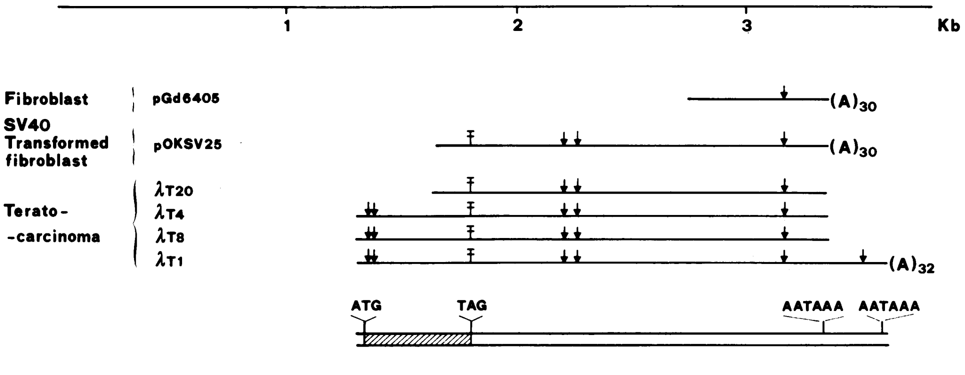


FIG. 1. Restriction map of cDNA clones. Clone pGd6405, isolated from a human fibroblast cDNA library (9), was used to screen the SV40-transformed-fibroblast library of Okayama and Berg (10). A genomic subclone, pGd1.4 (2), was used to screen the teratocarcinoma cDNA library in λgt10 (a gift of J. Skowronski). Restriction sites: ⫪, *Eco*RI; ↓, *Pst* I. Below the restriction maps is a representation of the GdX mRNA; hatching corresponds to the ORF. Potential polyadenylylation signals (AATAAA) are indicated.

RNA Blot Hybridization. HeLa cell total RNA (Fig. 4, lanes 2 and 12) was prepared in our laboratory by the guanidine hydrochloride method (12). Total RNA from HL-60 cells (16) was isolated by cell lysis in 4 M guanidine thiocyanate and sedimentation through 5.7 M CsCl (17). RNA from HeLa (Fig. 4, lane 6) and choriocarcinoma (JEG) cells (18) was a gift from J. Chou (National Institutes of Health, Bethesda, MD). RNA from human NTera2/D1 teratocarcinoma cells (19) and from PA-1, Ca-Ma, and COS-1 cells was a gift from J. Skowronski (20). Human fibroblast RNA was a gift from A. Simeone (International Institute of Genetics and Biophysics, Naples), and normal human thyroid RNA was a gift from A. Fusco (II Medical School, Naples). All the RNAs were total cellular RNAs except the RNA from NTera2/D1 teratocarcinoma cells, which was poly(A)$^+$. Samples (5 μg) were electrophoresed in a 1.5% agarose gel containing formaldehyde (12). The RNA was transferred to nylon filters and hybridized with the probes indicated, using the same hybridization conditions described for DNA blots.

RESULTS

Isolation of the *GdX* Gene and Mapping of Exons.The cDNA and genomic clones shown in Figs. 1 and 2 were obtained by screening several cDNA and genomic libraries. The first cDNA clone (pGd6405) was obtained from a human fibroblast cDNA library during the screening for G6PD cDNA (10). With the insert of the cDNA clone pGd6405, genomic λ clones λGdT7, λGd11, λGd5A, and λGd3C were obtained (1). A repetitive-sequence-free genomic fragment (pGd1.4; ref. 2) was subsequently used to screen SV40-transformed fibroblast and teratocarcinoma cDNA libraries and to obtain cDNA clones corresponding to full-length cDNA (Fig. 1). Further extension of the genomic region cloned has shown that *G6PD* and the genomic region corresponding to pGd6405 are only 40 kb apart on the X chromosome (1). The elucidation of the structure of *G6PD* (1, 17) has shown that the genomic region corresponding to pGd6405 is a transcribed region different from *G6PD*, and it was named *GdX*. In addition, an analysis of this region has revealed a third gene (*P3*) located 5′ to *GdX* (Fig. 2). The structural analysis of *P3* will be published elsewhere.

Hybridization of the ^{32}P-labeled cDNA clones to restriction digests of DNA from phages λGd5A and λGd3C showed that the *GdX* gene spans only about 3.5 kb of genomic DNA. The precise location of exons and introns and of exon/intron boundaries was defined by sequencing the cDNA and the genomic DNA (Fig. 3). The sequence of 3584 nucleotides was obtained by both chemical and enzymatic sequencing methods from the various cDNA and genomic clones shown in Figs. 1 and 2.

GdX is divided into four exons. The first three exons are small [83, 106, and 208 base pairs (bp), respectively]; the fourth exon is 1926 bp long. Two canonical polyadenylylation signals (AATAAA) can be recognized at the 3′ end; they are followed by a 30-residue poly(A) tail in some of the cDNAs, 16–17 bp downstream from the polyadenylylation signals (Fig. 1). The sequences at the 5′ and 3′ intron boundaries are in agreement with the consensus sequence for exon/intron boundaries of eukaryotic genes (22).

DNA and RNA Blot Hybridization. Hybridization of ^{32}P-labeled cDNA or genomic clones to blots of restriction endonuclease-digested genomic DNA showed that the *GdX* gene is unique in the human genome; only the bands expected from restriction analysis of the genomic clones were observed (data not shown).

In RNA blots, the *GdX* probes hybridized to two RNA species of 2.2 and 2.4 kb (Fig. 4*A*). The two polyadenylylation sites appear to be used with the same frequency and their relative abundance does not vary in a number of different cell lines and cell types we have tested. Moreover,

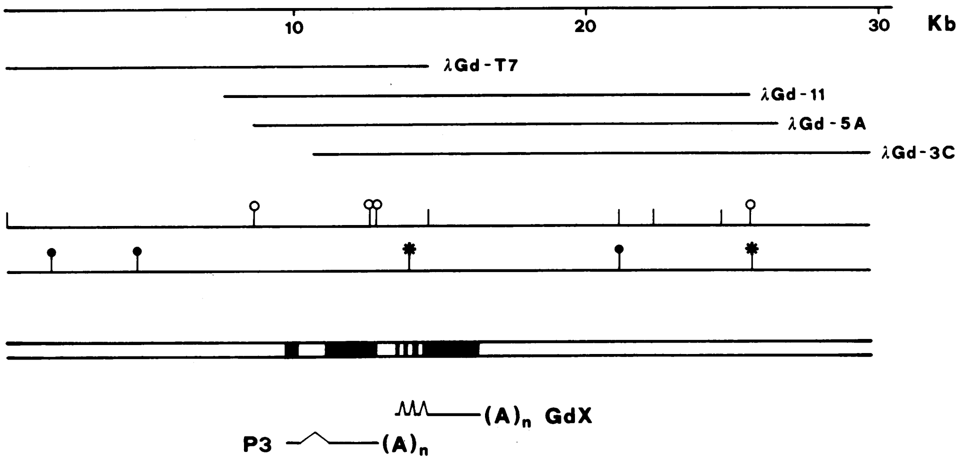


FIG. 2. Restriction map of the *GdX* genomic region and genomic clones. Below the λ genomic clones, isolated as described (1), is a restriction map of the genomic region. Restriction sites: ♀, *Hin*dIII; |, *Eco*RI; †, *Bam*HI; ✝, *Kpn* I. The exons corresponding to the *P3* and *GdX* genes are represented as blacked-in regions below the restriction map.

```
1    cccgggtgtccccactctgggagaggtgggaggatgagggcagaggcgtagcccccctgccccctctcaagctcagtggagcgtttcgcagccagaccagaaaggaggaggtaggaagcc
121  accagacagggagcgggggcgggccgtctgtgggtatgcagggccaggcgccgccagcagtcctagcgcgtggggtgggccgagcagtccgcgggggcgagcgtaccttctggccctccc
241  ttctccatgtcccagctcccagatgacgccaacttggcccgactggttcctgctcccaactcccccgcgtcgttgctttgggaggggtccgccggccagggcggtcgatgcgcgcgtgct
361  aggcgggaccccggggggctcctcccgggcgcaattgggccgggggcggggccggggcgcggacggggcaggagggcggggccgggaagcgcgggcggcgggcgcgcccctctcgctgc
481  ttccggcggcggcgggcggttccagcgcgcgcgcccggggcggcggcgcgcggcggggggtggttggggtGCGCGCCGGCCCGAGTGGACGCCGTCGCGACCGCCATGCAGCTGACGGTG
                                                                                                              M  Q  L  T  V

601  AAGGCGCTGCAGGGCCGCGAGTGCAGCCTGCAGgtagggtccccgggccgggccgccgggtccggcgcgctctctcgcccgctcctggcagggggaccggcagggccgagggcgggcacg
     K  A  L  Q  G  R  E  C  S  L  Q

721  gggaggccgggccgagggccgggagccgggagagctgccctcgggggcggacgcgtgggcttcgccgtctgggggccgcgggctcaacccttgcctcccaccctccgcgcggccagGTGC
                                                                                                                         V

841  CAGAGGACGAGCTGGTGTCCACGCTGAAGCAGCTGGTCTCCGAGAAGCTGAACGTCCCAGTGCGCCAGCAGCGGCTGCTGTTCAAGGGCAAGGCCCTGGCAGgtacccagggagaggaga
      P  E  D  E  L  V  S  T  L  K  Q  L  V  S  E  K  L  N  V  P  V  R  Q  Q  R  L  L  F  K  G  K  A  L  A

961  cgcccggggagccccgcaggaagcgggaccggggtgcgcgggccgaggccagcagccgtgtgtcgggtcggggtgcgggcgccaagcgccacatgaccccaagggaggcggccgcatgcc

1081 tcacagccggatgccgcagcgctccgctccccgcccgaggcggcctcgcgccccgggccacccccacggcggtggaggaggaggaacccatcgatctgtctttggcagATGGGAAACGAC
                                                                                                                  D  G  K  R

1201 TCTCGGATTATAGCATCGGGCCCAACTCCAAGCTCAACCTAGTGGTCAAACCCCTGGAGAAGGTGCTACTAGAAGAAGGCGAGGCCCAGAGGCTGGCCGACTCCCCACCCCCGCAGGTCT
      L  S  D  Y  S  I  G  P  N  S  K  L  N  L  V  V  K  P  L  E  K  V  L  L  E  E  G  E  A  Q  R  L  A  D  S  P  P  P  Q  V

1321 GGCAGCTGATCTCCAAAGTCTTGGCCCGCCACTTCAGTGCGGCAGATGCCAGCAGGGTCCTGGAACAGCTACAGAGGgtgagaagggtaaccctgggcatcctcttcaggcccagttgct
      W  Q  L  I  S  K  V  L  A  R  H  F  S  A  A  D  A  S  R  V  L  E  Q  L  Q  R

1441 cccagcccctacctctgttgcagtgtatgcccccccccactggcagcccctgaatcttcctgcccttctctccccagGATTACGAGAGGTCCCTGAGTCGCCTGACGCTGGACGACATC
                                                                                  D  Y  E  R  S  L  S  R  L  T  L  D  D  I

1561 GAACGGTTGGCCAGCCGCTTCCTGCACCCTGAAGTGACTGAGACAATGGAGAAGGGCTTCTCCAAATAGAATTCTCGGAGCATGGGGAGGTGCCCAACGCCAGGCTACCGCTGCATGTCG
      E  R  L  A  S  R  F  L  H  P  E  V  T  E  T  M  E  K  G  F  S  K

1681 CACTAAGTGTGTTCTCCTGTTGCAGTTGGGCTCATCATCGTCATAGCTGGCATGTACCTGGCTCTGGCCAGGTGCTAGGCACTCCTCACAGCTTGACCTGGGTTTGCTTCCACACCCTCA
1801 GAAGAGGAGAGCCGGCACAGCAGAGGCACCGGCGATGAAGTGGCACAGCCCCAGTCGGAATCCAGCGGCTTCTGAAAGTGCCTTGGTGTGAGAAGGAGGAAGGGGCCGCTTGGAGAGCTG
1921 GGCTCTCGCATGTCATGTTTAGCCCACTGAGAATATACCCTCAGGTGACTCCTCCCGATCCTGAAAGGGAAAGCAGCTGTCACCTGACTTCTGGCGGGTCCAACATAGCCCTCAGCTTAC
2041 CTCTGCAGGAGGCCGAGTGACAGCCAGCCCTGGAACCCCCCGACCCCCGCAGCTGCTGCAGCCAGTGTGCCAGTGTGCGTTCGACAAATGGAAAAGCAGATTGGGGCCAGGGAGTCAGCA
2161 AGGCACCCCAGCTCTGTGATGTGACTTTGGGCGAGTCTCAGTGCCACTTGGTCCCCAGCCATGGACTCTATACAGTGAGGGCCACTCACCGACCTGAGTCAGTGTCCTCTGTCTCACAGG
2281 GCCCCTCTCCCTTCTGTTCAGTAAACAAACTGAAGCCAAAAATGAAGCCGTGGCCAAGCGTGACCCAGAGATGGGGTGTCCTGGCCCTTAGTGACACAGCTCCTCTCTGGGGCACCCTAT
2401 CTGCTTGTCTCGTCCAGGAAGCGTTAGGGCGATGGGGACTCGGCAAGGCAGTGTCAGAGCTGGGACTGGGCGTAGGCCTTTGTCCTCATCCCCAGCACTGGCCTCCCTGTGGGAAGATGA
2521 ACATATCCCAGCCACCTGTGTACAGGGGCTCACTTTGTGTGCTCCTTGTTGCCTGGAGAAGAACCTTGGGGTGCCAGGGTGGGGGCAGAAGCATGGGCTGGGTTCCGGTTCATCCTCCTC
2641 CACCCTGCCGTGTGTGTGGGCACAAGAGGACATCTAACCACCTGCTCCTTGGAGGAGGCCCCCAGGGGTGGTAGAGGCTGGAAGGAAGCCACATCAGGAGGACGCCACTCCGGCCCTTCA
2761 CCCTTGCCAAGTGAGCTGCTCACAGTGTGGTCAGGGCTGCGCGTGCTGGAGGCCCTCCTGCCTGGGCCTTGTGGGGCAAATATTGGGTCCCCAGGCTGGAAAGATGGACAGAGGCCCAAT
2881 GGGTGAAGGCTTTGAAGAGCACACAGAAGCCCCTGGCCCCCCACGAGAGCTGGAGAGCCATGTATATGGCTTCAAAGCCACCTACGGCAGGGACACACTCGTGAGCATGTGTGGCCTGCA
3001 GTTCAGGTGATACATTTACCAGTGTTCTTGTTTGTGTGGTGCCAGGAAATTGATTTTGGAAAAAGTGAAATAACATTAAAGGTGAATGTGAGGCTTCTACTTTTATCCAAAAGGAGCTAT
3121 ATTAGCTAGGCTGTTTCTGATATCCAATCATTGGTTTAACAATAAAGGCAATTTGTTTAATCAGTTAACGGAAATTTCTTGGCTTATGAAATGAAAAGTCCAGTGGTATTGGCATTGGCA
3241 GCAGGTGAGCAATTTCACCCAGTGTCTTCTGCCTCCCTCTGCGTTGGTATCTGCTACATCCCAGGCCACCACCTCCGAGGATGAAAAGATGGCTGCCTGCAGCTTCCACGGAATCCCTCC
3361 CTCACTGCCAGAGCAGCATCTTCTCTGTGTACCACCTCTGCTGTCTCAGATGCCCCTCAGCAAATAAACACTCTTCTCGTTGGTCagaactggattgtgcgtccaatcattttggctgggg
3481 tagggggtaattctccgtcagggctggctccacttggagctaggcatggggaccacttttcactggacacacatcggctatgcaatgggacagcaaggactact
```

FIG. 3. Nucleotide sequence of the *GdX* gene. The cDNA sequence is in uppercase letters; introns and flanking regions are in lowercase letters. ATG and TAG codons are boxed. The CCGCC sequence, at the initiation of translation (21), is indicated by dots over the sequence. Binding sites for transcription factor Sp1 are underlined. The polyadenylation signals are indicated by a double underlining and the positions of the poly(A) tail in cDNAs are indicated by arrowheads. The amino acids encoded by the cDNA are shown (standard one-letter symbols) below the nucleotide sequence.

the *GdX* gene is expressed at low levels in all cell types analyzed (Fig. 4), behaving as a housekeeping gene.

***GdX* Is Conserved.** A full-size cDNA probe, as well as a probe corresponding to the coding region of *GdX* (see below), was hybridized to blots of restriction enzyme-digested DNA from several animal species [monkey, mouse, rat, ox, horse, chicken, *Xenopus laevis, Tachidromus sex lineatus* (a reptile), *Drosophila melanogaster*] and the yeasts *Saccharomyces cerevisiae* and *Histoplasma capsulatum.* Fig. 5 shows the results for some of the species tested; the *GdX* probes hybridize to one or a few DNA bands at rather stringent hybridization washing conditions (1× SSC at 60°C), suggesting that the *GdX* sequence has been conserved during evolution.

Sequence Analysis of the cDNA. The nucleotide sequence of the GdX cDNA was compared with the GenBank database[‡]; no homologies were found with any published nucleotide sequence. A search for ORFs in the GdX cDNA showed one ORF starting from the ATG at position 36 of the cDNA and ending with the TGA stop codon at position 506. The starting ATG is the first in the cDNA sequence. The sequence CCGCC, preceding the ATG, is in good agreement with the eukaryotic consensus sequence for translation initiation sites (21). Upstream to the first ATG, the 5′ noncoding region is very short (35 nucleotides) and G+C-rich, as is the preceding genomic sequence (75–80% G+C). Since the cDNAs isolated correspond to the total length of the mRNA, we presume that this region corresponds, within a few nucleotides, to the whole 5′ noncoding region of the *GdX* gene. Preliminary nuclease S1 mapping data are in agreement with this concept. Downstream from the stop codon are several additional stop codons in all three reading frames.

The ORF of 471 nucleotides codes for a 157 amino acid protein (Fig. 3). Searching the National Biomedical Research Foundation (NBRF) protein bank[§] for homologies with the GdX protein showed 43% identity with ubiquitin (62% if conservative amino acid substitutions are considered as identities) (Fig. 6*A*). The homology starts with the first methionine and ends at amino acid 74 of both proteins. The two COOH-terminal amino acids of ubiquitin (Gly-Gly), through which ubiquitin binds to protein (6, 25), are not conserved in the GdX protein.

The remaining 87 amino acids do not show any homology to ubiquitin precursor sequences (7–9) found as COOH-terminal additions to ubiquitins encoded by cDNAs from humans, *Dictyostelium*, and yeast.

[‡] EMBL/GenBank Genetic Sequence Database (1987): GenBank (Bolt, Beranek, and Newman, Cambridge, MA), Tape Release 48.

[§] Protein Identification Resource (1987) Protein Sequence Database (Natl. Biomed. Res. Found., Washington, DC), Release 11.

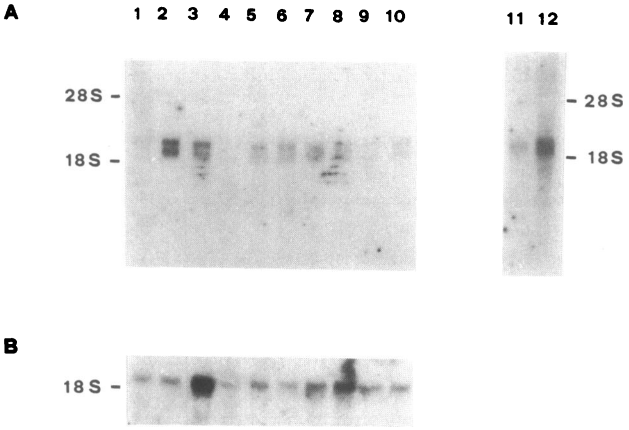


FIG. 4. RNA blot hybridization. (*A*) Hybridization of the *GdX* coding sequence to RNA from various sources. The hybridization probe was an *Eco*RI fragment of the λT4 cDNA (Fig. 1) from the *Eco*RI site of the cloning linkers to the *Eco*RI site in the cDNA. RNA preparations were from human fibroblasts (lane 1), HeLa cells (lanes 2, 6, and 12), NTera2/D1 (undifferentiated human teratocarcinoma cells; lane 3), HL-60 (undifferentiated human myeloid cells; lane 4), human T lymphocytes (lane 5), Ca-Ma (human mammary carcinoma cells; lane 7), PA-1 (human neuroblastoma cells; lane 8), COS-1 (monkey fibroblasts transformed with replication-defective SV40; lane 9), JEG (human choriocarcinoma cells; lane 10), normal human thyroid (lane 11). Markers show positions of 28S and 18S rRNA. (*B*) Hybridization of a chicken β-actin cDNA probe (23) to the same filter shown in *A*.

An additional homology has been found by searching the NBRF bank: amino acids 110–123, a stretch of 14 amino acids in the central part of the COOH-terminal domain, shows 62% identity with the thyroglobulin hormonogenic sequence (26, 27) (Fig. 6*B*). The homology between *GdX* and the thyroglobulin gene, at the nucleotide level, is very low; however, a splice junction is found in exactly the same position in this region in both genes (Fig. 6*B*, indicated by an arrow) (28).

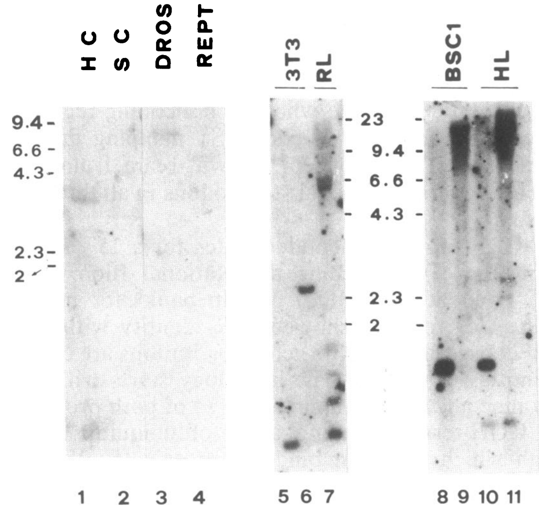
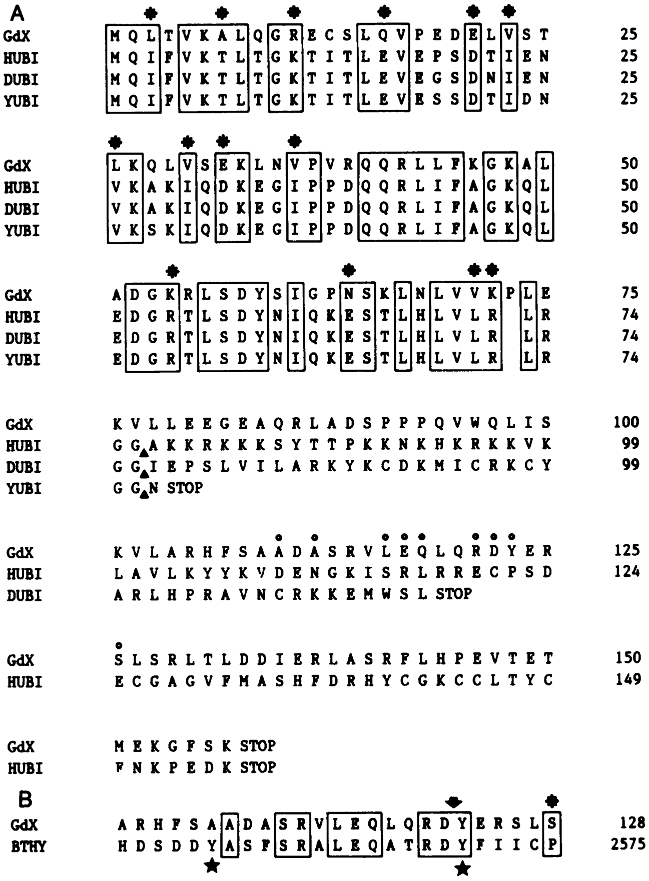


FIG. 5. Blot hybridization of the *GdX* coding region (same probe as in Fig. 4) to DNA from the yeasts *H. capsulatum* (H C) and *S. cerevisiae* (S C), the fruit fly *D. melanogaster* (DROS), the reptile *T. sex lineatus* (REPT), mouse 3T3 cells (3T3), rat liver (RL), monkey BSC1 cells (BSC1), and human leukocytes (HL). DNA samples were digested with *Pst* I (lanes 5, 7, 8, and 10) or *Eco*RI (lanes 1–4, 6, 9, and 11). Markers (kb) were *Hin*dIII fragments of phage λ DNA.



```
A
GdX    M Q L T V K A L Q G R E C S L Q V P E D E L V S T    25
HUBI   M Q I F V K T L T G K T I T L E V E P S D T I E N    25
DUBI   M Q I F V K T L T G K T I T L E V E G S D N I E N    25
YUBI   M Q I F V K T L T G K T I T L E V E S S D T I D N    25

GdX    L K Q L V S E K L N V P V R Q Q R L L F K G K A L    50
HUBI   V K A K I Q D K E G I P P D Q Q R L I F A G K Q L    50
DUBI   V K A K I Q D K E G I P P D Q Q R L I F A G K Q L    50
YUBI   V K S K I Q D K E G I P P D Q Q R L I F A G K Q L    50

GdX    A D G K R L S D Y S I G P N S K L N L V V K P L E    75
HUBI   E D G R T L S D Y N I Q K E S T L H L V L R   L R    74
DUBI   E D G R T L S D Y N I Q K E S T L H L V L R   L R    74
YUBI   E D G R T L S D Y N I Q K E S T L H L V L R   L R    74

GdX    K V L L E E G E A Q R L A D S P P P Q V W Q L I S    100
HUBI   G G A K K R K K K S Y T T P K K N K H K R K K V K    99
DUBI   G G I E P S L V I L A R K Y K C D K M I C R K C Y    99
YUBI   G G N STOP

GdX    K V L A R H F S A A D A S R V L E Q L Q R D Y E R    125
HUBI   L A V L K Y Y K V D E N G K I S R L R R E C P S D    124
DUBI   A R L H P R A V N C R K K E M W S L STOP

GdX    S L S R L T L D D I E R L A S R F L H P E V T E T    150
HUBI   E C G A G V F M A S H F D R H Y C G K C C L T Y C    149

GdX    M E K G F S K STOP
HUBI   F N K P E D K STOP

B
GdX    A R H F S A A D A S R V L E Q L Q R D Y E R S L S    128
BTHY   H D S D D Y A S F S R A L E Q A T R D Y F I I C P    2575
```

FIG. 6. (*A*) Comparison of the GdX amino acid sequence to the sequence of human (HUBI), *Dictyostelium* (DUBI), and yeast (YUBI) ubiquitin, derived from cDNA clones (7–9). Only the COOH-terminal unit of the yeast polyubiquitin precursor is shown, which has an additional amino acid (N) before the stop codon. Circles above the sequence show the region homologous to thyroglobulin. Arrowheads indicate the COOH termini of the ubiquitins. (*B*) Comparison of GdX residues 104–128 to the COOH-terminal hormonogenic site of bovine thyroglobulin (24). Arrow indicates the position of the splice junction. Stars below the tyrosines indicate the iodination sites in thyroglobulin. Homologous amino acids are boxed; conservative amino acid substitutions are indicated with asterisks.

DISCUSSION

The nucleotide sequence reported in this paper identifies an X-linked gene in the *G6PD* cluster at Xq28. This gene, which we call *GdX*, is about 3.5 kb long and interrupted by three introns. Three of the four exons are rather small; the fourth exon is 1927 bp long, corresponding to 83% of the mRNA length. The *GdX* gene is a highly conserved single-copy gene expressed ubiquitously and could code for a cellular housekeeping function. Its nucleotide sequence shows no homology to previously sequenced genes in GenBank.

The sequencing of the *GdX* region has allowed the mapping of one of the CpG clusters previously described (2). From the length of the cDNAs and preliminary nuclease S1 mapping data (not shown), it is very likely that the 5′ end of the GdX mRNA is within a few nucleotides of position 551 in Fig. 3, which corresponds to the cDNAs extending most 5′. Moreover, the genomic region upstream has the features of a promoter of housekeeping genes (1): it is very rich in G+C, lacks signals like "TATA" or "CAT" boxes, and the GGGCGG hexanucleotide or its complement (transcription factor Spl binding site; ref. 29) is present seven times (underlined in Fig. 3). Thus, we can conclude that the CpG cluster is 5′ to the *GdX* gene. However, the clustering of the CpG dinucleotides is not confined to the 5′ flanking region but extends into the gene, to the second intron. This observation suggests that sequences in the body of the gene, as well as 5′ flanking regions, might be responsible for the regulation of the expression of such genes.

From the cDNA sequence we identified an ORF of 471 nucleotides. The starting ATG is the first in the cDNA sequence. The first stop codon in-frame with the ATG leaves a very long and uninterrupted 3′ noncoding region, as has been found in many other genes. The ORF codes for a protein of 157 amino acids. The comparison of the amino acid sequence of the GdX protein to the NBRF data bank showed several interesting features of the GdX gene product. From this analysis, two different moieties can be recognized in the GdX protein. The NH_2-terminal portion, residues 1–74, shows homology to ubiquitin, a highly conserved 76 amino acid protein purified from several animal and plant species (6): 32 amino acids are unchanged (43%) and 14 represent conservative substitutions. The ubiquitin-like moiety is followed by a COOH-terminal moiety of 80 amino acids. Ubiquitin cDNAs have been cloned and sequenced from a variety of eukaryotes, and they revealed that ubiquitin coding elements are typically organized into spacerless head-to-tail arrays (30–32). However, cDNA clones isolated from a human library (7), from *Dictyostelium* (8), and from yeast (9) provide a striking exception to the "polyubiquitin" organization. The DNAs of those clones encode proteins of 156 (7), 117 (8), and 128 and 152 (9) amino acids, respectively; the first 76 residues are identical to those of mature ubiquitin, whereas the remaining residues have no homology to ubiquitin. In every case, the COOH-terminal sequences are rich in basic amino acids and contain several cysteine residues. It has been suggested (7–9) that these proteins are precursors to ubiquitin and that the COOH-terminal sequences could target the ubiquitin to the nucleus. The structural similarity of the protein encoded by the *GdX* gene to the ubiquitin precursors is striking. However, this general similarity in organization does not correspond to sequence homology or to similarity in amino acid composition. We suggest that the function of the COOH-terminal moiety could be the targeting of the NH_2-terminal, ubiquitin-like part to a specific cell compartment or class of protein.

In this respect, the region of the protein homologous to the thyroglobulin hormonogenic site could be important. In thyroglobulin this is a highly specialized amino acid sequence that allows iodination of the tyrosine and the formation of the thyroid hormone (33, 34). The homology of the GdX COOH-terminal part to one of the two thyroglobulin hormonogenic sequences is strengthened by the conservation of an exon/intron splice junction (28). However, it is unlikely that the function of the tyrosine in the GdX protein is the binding of iodine. This sequence as a whole could, however, be envisaged as able to recognize specific cellular components or molecules, possibly through tyrosine modification(s).

The function of ubiquitin may be diverse and is not fully understood (6). Ubiquitin is thought to have a major role in nonlysosomal, ATP-dependent protein degradation (35). But it was also shown that ubiquitin is present in the nucleus, where it binds histone H2A and is found in chromatin (36). More recently, it was discovered that the lymphocyte homing receptor at the cell surface is ubiquitinated (37). Whether or not cell surface ubiquitin has a different function than its intracellular counterpart is not known. Studies of ubiquitin-dependent proteolysis have elucidated a pathway by which ubiquitin-protein conjugates can be formed and subsequently degraded, and they have demonstrated that binding of ubiquitin can function as a signal for proteolysis (6, 25). However, little is known of specific, ubiquitin-dependent proteolytic events that occur *in vivo* and that could be important for cell-cycle control (38), DNA replication (6), and stress response (39, 40). The homology of the GdX NH_2-terminal moiety to ubiquitin suggests that, in addition to ubiquitin, proteins with a similar structure may play specific roles in cellular processes, possibly by protein–protein recognition.

We thank Drs. A. Simeone, J. Chou, and A. Fusco for providing RNA samples; Dr. J. Skowronski for providing RNA and the teratocarcinoma cDNA library; Drs. L. Luzzatto and E. Avvedimento for helpful discussions and critical reading of the manuscript; and Mrs. M. Terracciano, Mrs. C. Salzano, and Mrs. C. Sole for technical assistance. This work was supported by Progetto Finalizzato Consiglio Nazionale delle Richerche "Ingegneria Genetica e Basi Molecolari delle Malattie Ereditarie" and by a grant from the Associazione Italiana per la Ricerca sul Cancro to D.T.

1. Martini, G., Toniolo, D., Vulliamy, T., Luzzatto, L., Dono, R., Viglietto, G., Paonessa, G., D'Urso, M. & Persico, M. G. (1986) *EMBO J.* **5**, 1849–1855.
2. Toniolo, D., D'Urso, M., Martini, G., Persico, M., Tufano, V., Battistuzzi, G. & Luzzatto, L. (1984) *EMBO J.* **3**, 1987–1995.
3. Bird, A. P. (1984) *Nature (London)* **321**, 209–213.
4. McKusick, V. A. (1983) *Mendelian Inheritance in Man* (The Johns Hopkins University, Baltimore).
5. Vuay-Kumar, S., Bugg, C. E., Wilkinson, K. D. & Cook, W. J. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 3582–3585.
6. Finley, D. & Varshavsky, A. (1985) *Trends Biochem. Sci.* **10**, 343–347.
7. Lund, P. K., Moats-Staats, B., Simmons, J. G., Hoyt, E., D'Ercole, A. J., Martin, F. & Van Wyk, J. J. (1985) *J. Biol. Chem.* **260**, 7609–7613.
8. Westphal, M., Muller-Taubenberger, A., Noegel, A. & Gerisch, G. (1986) *FEBS Lett.* **209**, 92–96.
9. Ozkaynak, E., Finley, D., Solomon, M. J. & Varshavsky, A. (1987) *EMBO J.* **6**, 1429–1439.
10. Persico, M. G., Toniolo, D., Nobile, C., D'Urso, M. & Luzzatto, L. (1981) *Nature (London)* **294**, 778–780.
11. Okayama, H. & Berg, P. (1983) *Mol. Cell. Biol.* **3**, 280–289.
12. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
13. Toniolo, D., Persico, M. G., Battistuzzi, G. & Luzzatto, L. (1984) *Mol. Biol. Med.* **2**, 89–103.
14. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
15. Hattory, M., Hidaka, S. & Sakaki, Y. (1985) *Nucleic Acids Res.* **13**, 7813–7827.
16. Collins, S. J., Gallo, R. C. & Gallagher, R. E. (1977) *Nature (London)* **270**, 347–349.
17. Persico, M. G., Viglietto, G., Martini, G., Toniolo, D., Paonessa, G., Moscatelli, C., Dono, R., Vulliamy, T., Luzzatto, L. & D'Urso, M. (1986) *Nucleic Acids Res.* **14**, 2511–2522.
18. Kohler, P. O. & Bridon, W. E. (1971) *J. Clin. Endocrinol. Metab.* **32**, 683–690.
19. Andrews, P. W. (1984) *Dev. Biol.* **103**, 285–293.
20. Skowronski, J. & Singer, M. F. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 6050–6054.
21. Kozak, M. (1986) *Cell* **44**, 283–292.
22. Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472.
23. Cleveland, D. W., Lopata, M. A., MacDonald, R. J., Cowan, N. J., Rutter, W. J. & Kirschner, M. W. (1980) *Cell* **20**, 95–105.
24. Mercken, L., Simons, M.-J., Swillens, S., Massaer, M. & Vassart, G. (1985) *Nature (London)* **316**, 647–651.
25. Bachmair, A., Finley, D. & Varshavsky, A. (1986) *Science* **234**, 179–186.
26. Rawitch, A. B., Chernoff, S. B., Litwer, M. R., Rouse, J. B. & Hamilton, J. W. (1983) *J. Biol. Chem.* **285**, 2079–2082.
27. Marriq, C., Rolland, M. & Lissitzky, S. (1982) *EMBO J.* **1**, 397–401.
28. Di Lauro, R., Obici, S., Condliffe, D., Ursini, V. M., Musti, A., Moscatelli, C. & Avvedimento, V. E. (1985) *Eur. J. Biochem.* **148**, 7–11.
29. Kadonaga, J. T., Jones, K. A. & Tjian, R. (1986) *Trends Biol. Sci.* **11**, 20–23.
30. Izkaynak, E., Finley, D. & Varshavsky, A. (1984) *Nature (London)* **312**, 663–666.
31. Dworkin-Rasti, E., Shrutkowski, A. & Dworkin, M. B. (1984) *Cell* **39**, 321–325.
32. Wiborg, O., Pedersen, M. S., Wind, A., Berglund, L. E., Marcker, K. A. & Vuust, J. (1985) *EMBO J.* **4**, 755–759.
33. Edelhoc, H. & Robbins, J. (1978) in *The Thyroid*, eds. Werner, S. C. & Ingbar, S. H. (Harper & Row, New York), 4th Ed., pp. 62–76.
34. Musti, A. M., Avvedimento, E. V., Polistina, C., Ursini, V. M., Obici, S., Nitsch, L., Cocozza, S. & Di Lauro, R. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 323–327.
35. Hershko, A. (1983) *Cell* **34**, 11–12.
36. Levinger, L. & Varshavsky, A. (1982) *Cell* **28**, 375–385.
37. Siegelman, M., Bond, M. W., Gallatin, W. M., St. John, T., Smith, H. T., Fried, V. A. & Weissman, I. L. (1986) *Science* **231**, 823–829.
38. Finley, D., Ozkaynak, E. & Varshavsky, A. (1987) *Cell* **48**, 1035–1046.
39. Bond, U. & Schlesinger, M. J. (1985) *Mol. Cell. Biol.* **5**, 949–956.
40. Parag, H. A., Raboy, B. & Kulka, R. G. (1987) *EMBO J.* **6**, 55–61.