Volume 14 Number 6 1986 Nucleic Acids Research

Isolation of human glucose-6-phosphate dehydrogenase (G6PD) cDNA clones: primary structure of the protein and unusual 5' non-coding region

M.Graziella Persico*, Giuseppe Viglietto+, Giuseppe Martini+, Daniela Toniolo+, Giacomo Paonessa+, Claudio Moscatelli+, Rosanna Dono+, Tom Vulliamy§, Lucio Luzzatto§ and Michele D'Urso+

Human Genetics Branch, NICHD, National Institutes of Health, Bethesda, MD 20892, USA, +International Institute of Genetics and Biophysics, CNR, Via Marconi 10, 80125 Naples, Italy and §Department of Haematology, Hammersmith Hospital, London W12 0HS, UK

Received 8 January 1986; Revised and Accepted 21 February 1986

ABSTRACT

Glucose-6-phosphate dehydrogenase (G6PD) is an ubiquitous enzyme which by determinig the NADPH level has a crucial role in NADPH-mediated reductive processes in all cells (1). The structural gene for G6PD, Gd , is X-linked in mammals and on the basis of its expression in many tissues, it can be regarded as a typical "housekeeping" gene (2). Over 300 variants of the protein are known, many of which have deficient enzyme activity. Nearly 100 of these variants are polymorphic in various populations (3). The mammalian enzyme is a homodimer or a homotetramer with a subunit molecular weight of ~ 56000 daltons (4). Here we report the isolation of cDNA clones from HeLa cells, SV40-transformed human fibroblasts, human placenta and human teratocarcinoma cell lines. These clones have enabled us to sequence the entire coding region of Gd . Thus, the entire amino acid sequence of human G6PD is provided for the first time. This work is the first step for structural analysis of G6PD variants and for an understanding of the biological features of this enzyme at the molecular level.

INTRODUCTION

The human glucose-6-phosphate dehydrogenase gene, Gd , is of special interest in human genetics because of its high rate of genetic heterogeneity. Of the over 300 variants found around the world nearly 100 are polymorphic in various populations. G6PD is the first enzyme of the pentose phosphate pathway of glucose metabolism which produces pentose sugars required for nucleic acid synthesis. At the same time G6PD produces NADPH molecules required for a variety of biosynthetic and detoxification reactions. G6PD deficiency in man can cause chronic non-spherocytic hemolytic disease or acute hemolytic anemia triggered by ingestion of drugs or fava beans. Despite the importance of G6PD, relatively few structural studies have been reported (1,2,5).

In this paper we describe the isolation and sequence of cDNA clones coding for human G6PD. The sequence of the protein has been derived by extensive sequencing of clones isolated from different cDNA libraries. The analysis of the cDNA sequence reveals particular features of the G6PD mRNA;

the most notable is the presence of a long GC-rich 5′ non-coding sequence. The isolation of the human clones described here is the starting point for the characterization of the numerous existing G6PD variants.

MATERIALS AND METHODS

Screening of cDNA libraries. A cDNA library prepared from HeLa cell mRNA enriched for G6PD mRNA (6,7) was screened using as a probe the 17mer AAGATGATGACNAAGAA deduced from the published peptide KMMTKK (5). The colony hybridization and washing of the filters were done according to published procedures (8). The 170 bp insert of the clone pGD-H-IAIII shown in Fig.1 was used for the screening of three cDNA libraries. The screening of 10^6 clones of the cDNA library made from SV40-transformed fibroblasts mRNA (9) yielded 15 clones. 11 clones were found during the screening of $7x10^5$ phages of the human placenta cDNA library in λgt11 (kindly provided by Dr. F. Gonzales). The screening of $8x10^5$ independent clones of the human teratocarcinoma cell line (Ntera2) (10) cDNA library in λgt10 made from polyA$^+$-enriched cytoplasmic RNA (a generous gift of Dr. J. Skowronsky) gave us 16 independent clones. Small scale phage DNA preparations were analyzed by restriction mapping using standard procedures (11). The EcoRI insert of the phage clones from placenta and teratocarcinoma libraries presented in Fig.1 were subcloned in pUC18 (12) for further analysis.

DNA sequencing. The nucleotide sequence of the clones from HeLa and fibroblast libraries was determined by the procedure of Maxam and Gilbert (13). The sequence of the clones from placenta and teratocarcinoma was determined by the dideoxy method (4) modified to sequence directly from the plasmid pUC18 (Dr. M. Hattory and Dr. Y. Sakaki, personal communication). Confidence in the correctness of the DNA sequence was obtained by performing the sequence on two or more different cDNA clones from each library.

Northern blot analysis. Total RNA from HL60 cells (15) was isolated by cell lysis in 4 M guanidine thiocyanate and sedimentation through 5.7 M CsCl (16). PolyA$^+$ RNA from placenta was a gift of Dr. F. Gonzales, polyA$^+$ RNAs from HeLa and choriocarcinoma cells (JEG) (17) were a gift of Dr. J. Chou. PolyA$^+$ RNA was selected by chromatography on oligo (dT) cellulose (18). 5 μg samples were electrophoresed through a 1.5% agarose gel containing formaldehyde (19). RNA was transferred to nitrocellulose and probed with nick-translated (20) insert of clone pGD-P-2. After hybridization at 65°C in 5x SSPE, 0.1% SDS, 2x Denhardt's and 100 μg/ml

E.coli DNA for 18 h, the filters were washed sequentially in 2x SSC, 0.1% SDS and 0.2x SSC at 65°C for 1 h.

RESULTS

Isolation and analysis of cDNA clones

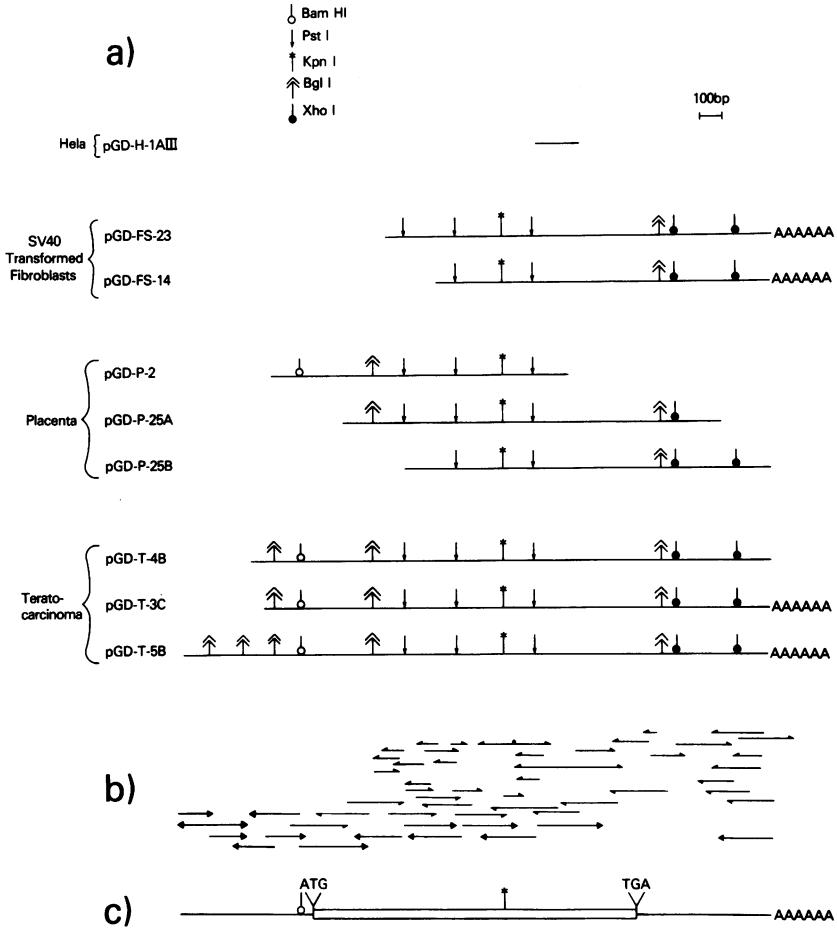
From the published partial amino acid sequence (5) we selected the peptide K-M-M-T-K-K and the corresponding 17mers AA$^{A}_{G}$ATGATGACNAA$^{A}_{G}$AA were synthesized. The mixture of oligonucleotides was used to screen $4x10^4$ clones of a cDNA library prepared from a partially purified G6PD mRNA from HeLa cells (6,7). Among the four clones isolated only one having a 170 bp insert (pGD-H-IAIII) (Fig.1) contained a sequence corresponding to the amino acid sequence of G6PD (5). In order to isolate the full coding sequence for human G6PD, we used the insert of pGD-H-IAIII to screen several different libraries, as described in detail in Materials and Methods and in the legend to Fig.1. Several clones were isolated and analyzed by restrction enzyme mapping and Southern blot hybridization (data not shown). Some representative clones are shown in Fig.1. The size of the major species of G6PD mRNA determined by Northern blot experiments is about 2,600 nucleotides (Fig.2).

Nucleotide sequence of human G6PD cDNAs

The sequence of 2624 nucleotides (Fig.3) was obtained using both chemical (13) and enzymatic (14) sequencing methods on clones from the three different libraries. Each segment was sequenced at least twice. The coding sequence (residue 1 to residue 1440 in Fig.3) was determined more than twice and on each of two or more independent clones, to reduce the possibility of cloning artefacts.

In the longest clone isolated the 5′ untranslated region of G6PD consists of 577 nucleotides and it contains 8 ATG codons which are out of the G6PD translation frame. It is also unusually GC-rich (73.3%), with two subregions more than 80% GC-rich (underlined in fig.3). It is interesting to note that several of these features are also present in other housekeeping genes (21). The first ATG triplet upstream to a long open reading frame was judged to be translation initiation site and it was assigned position No. 1. The sequence -5TCATCATG+3 boxed in Fig.3 is in good agreement with the consensus sequence for eukaryotic translation initiation sites (22).

The function of the 5′ non-coding region of mRNA is still unclear. The length of this region varies from just a few nucleotides to 729 residues (21). No particular sequence has yet been found common to all 5′ non-

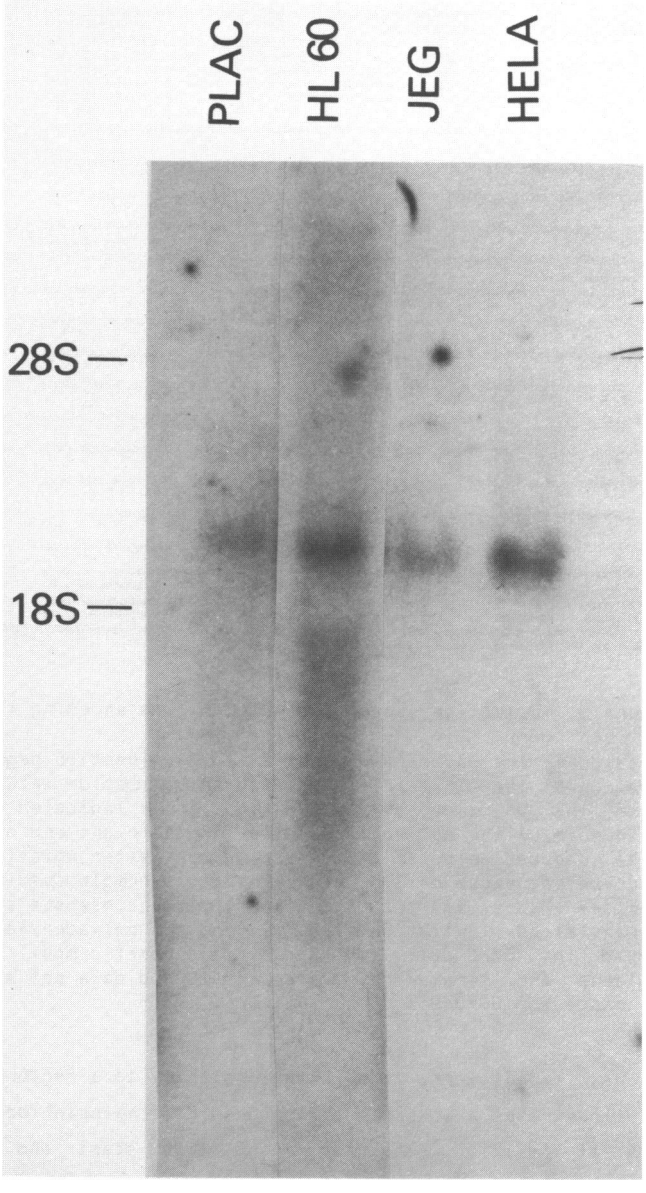


Legend to Figure 1. Restriction map and sequencing strategy of the cDNA clones.

a) The 16 possible 17mers AA$\begin{matrix}A\\G\end{matrix}$ ATGATGACNAA $\begin{matrix}A\\G\end{matrix}$AA were used to screen 4×10^4 clones of a cDNA library prepared from HeLa cells mRNA-enriched for G6PD mRNA (6,7). The 170 bp insert of the clone pGD-H-IAIII was used for the screening of three libraries as described in Materials and Methods. Some representative clones from the cDNA libraries of human fibroblasts (9), human placenta and human teratocarcinoma (Ntera2) (10) are presented in the Figure.

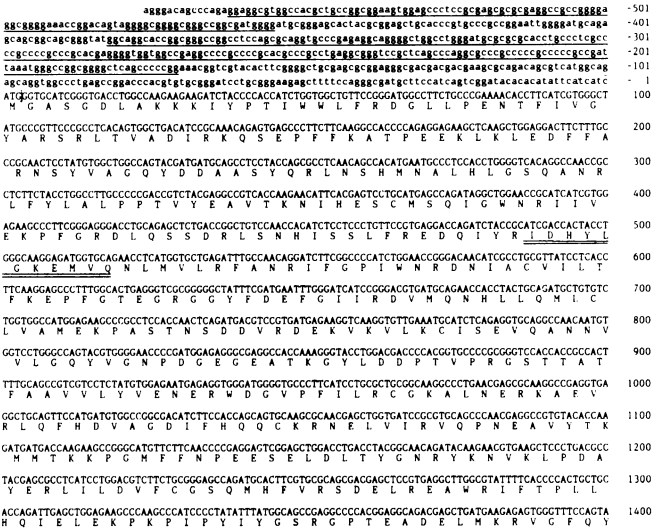
b) Sequencing strategy. The nucleotide sequence of the clone from HeLa cells and fibroblasts (⟶), from placenta (⟶) and teratocarcinoma cells (⟶) was determined as described in Materials and Methods.

c) Schematic representation of the cDNA corresponding to the mRNA coding for the G6PD.



Legend to Figure 2. Northern blot analysis of human Poly A^+ RNAs with G6PD-specific cDNA probe.

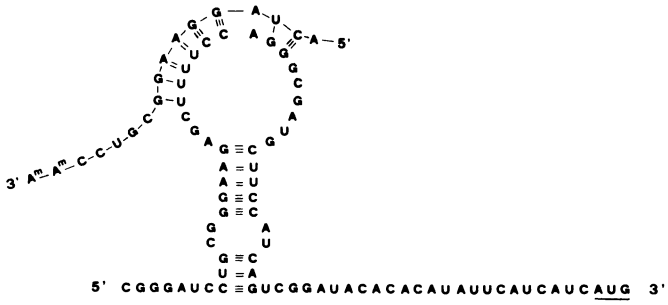
5 μg samples of PolyA^+ RNA from placenta, HeLa, choriocarcinoma (JEG) cells (17) and total RNA from HL60 cells (15) were electrophoresed trough a 1.5% agarose gel containing formaldehyde (19). RNA was transferred to nitrocellulose and probed with a nick-translated (20) insert of clone pGD-P-2. The position of 18S and 28S rRNA are indicated in the Figure.



Legend to Figure 3. Nucleotide sequence of cloned cDNA encoding the human G6PD.

Nucleotide residues are numbered in the 5′ to 3′ direction beginning with the first residue at the ATG triplet encoding the initiation methionine. The nucleotides on the 5′ side of the residue 1 are indicated by negative numbers; the number of the nucleotide residue at the right end of each line is given. The deduced amino acid sequence in one letter notation is shown below the nucleotide sequence. The peptide that is homologous in human and yeast G6PD (see Table II) is underlined. The consensus sequence for eukaryotic translation initiation site (22) and the polyadenylation signal (37) are boxed in. G6PD activities and electrophoretic mobilities of the human fibroblasts and teratocarcinoma cells (10 and data not shown), were in the normal range for G6PD-B.

coding regions, except for a loose complementarity to a region of the 18S rRNA. Fig.4 illustrates a possible stable predicted hairpin-loop structure of G6PD mRNA at position -25 from translation start and hypotetical interaction with the 3′ end of 18S rRNA. It has been proposed that this interaction may function in initiation or presumably in the regulation of translation (23). The 3′ untranslated region of G6PD mRNA is also unusually GC-rich (64.1%) when compared to other eukaryotic mRNAs. A non-canonical polyA addition site ATTAAA is located 13 bases upstream from the poly(A) tail (Fig.3). A high GC content (59.5%) characterizes also the coding



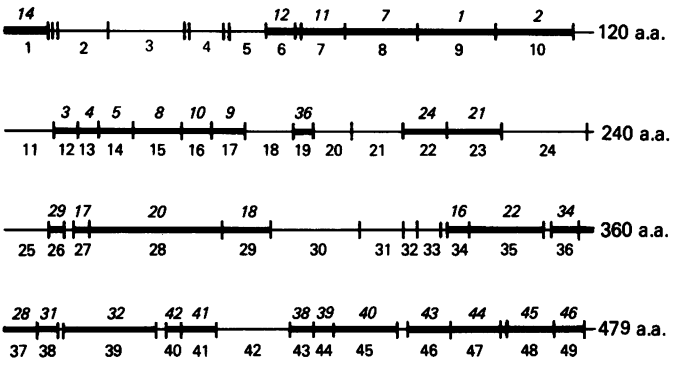
Legend to Figure 4. Possible secondary structure of G6PD mRNA. Hypothetical interaction between a possible hairpin-loop structure in the 5′ untranslated region of G6PD mRNA and the 3′ end of 18S rRNA. The stability of the stem of the structure and the association with 18S rRNA were calculated by the method of Tinoco et al. (38). The AUG starting of the translation is underlined.

portion of the G6PD mRNA. This result in a considerable deviation from the codon usage reported in the literature (24), based on a sample of 13,000 nucleotides of coding sequences correspoding to 22 proteins. Table I shows the codon usage for G6PD (column c) and for the random sample (column d).

Table I

a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d
TTT	Phe	35	17	TCT	Ser	17	4	TAT	Tyr	47	33	TGT	Cys	30	14
TTC	Phe	65	83	TCC	Ser	26	40	TAC	Tyr	53	77	TGC	Cys	70	85
TTA	Leu	5	0	TCA	Ser	11	13	TAA	End	0	0	TGA	End	100	100
TTG	Leu	9	4	TCG	Ser	7	9	TAG	End	0	0	TGG	Trp	100	100
CTT	Leu	11	2	CCT	Pro	24	4	CAT	His	42	9	CGT	Arg	9	18
CTC	Leu	22	28	CCC	Pro	41	79	CAC	His	58	90	CGC	Arg	19	55
CTA	Leu	7	2	CCA	Pro	24	8	CAA	Gln	26	0	CGA	Arg	10	3
CTG	Leu	46	63	CCG	Pro	11	8	CAG	Gln	74	100	CGG	Arg	15	14
ATT	Ile	23	15	ACT	Thr	20	10	AAT	Asn	34	12	AGT	Ser	11	4
ATC	Ile	64	80	ACC	Thr	47	75	AAC	Asn	66	88	AGC	Ser	29	27
ATA	Ile	13	3	ACA	Thr	21	5	AAA	Lys	45	10	AGA	Arg	24	11
ATG	Met	100	100	ACG	Thr	12	10	AAG	Lys	55	89	AGG	Arg	23	14
GTT	Val	13	3	GCT	Ala	31	10	GAT	Asp	38	33	GGT	Gly	15	13
GTC	Val	27	24	GCC	Ala	44	75	GAC	Asp	62	77	GGC	Gly	44	54
GTA	Val	9	0	GCA	Ala	17	13	GAA	Glu	40	5	GGA	Gly	17	3
GTG	Val	50	72	GCG	Ala	12	0	GAG	Glu	60	94	GGG	Gly	24	29

a, codon; b, amino acid; c, per cent codon utilization in a sample of human proteins (24); d, per cent codon utilization in human G6PD.



Legend to Figure 5. Comparision between the amino acid sequence deduced from cDNA and published one (5).
The line represents schematically the amino acid sequence deduced from cDNA. Vertical bar symbols indicate Arg and Lys residues delimiting the 49 hypothetical tryptic peptides (3 and more a.a. long) (numbers below the line). We have numbered the tryptic peptides of the published amino acid sequence (5) considering only the one, whose sequences are reported as definitive (numbers above the line). Of these, 34 out of 46 match with the sequence present in this paper. It is to be noted that the N terminal peptide deduced from cDNA is the 14th peptide obtained from the tryptic peptide sequencing. The N-terminal peptide of the previously published protein sequence occupies the 9th position in the protein deduced from the cDNA.

The most evident differences are seen with respect to the codons CAC and GAC.

Analysis of the amino acid sequence

The deduced amino acid sequence (in one letter notation) of G6PD is shown in Fig.3 below the nucleotide sequence. The predicted polypeptide consists of 479 amino acids and corresponds to a molecular weight of 55,126 daltons, in good agreement with previous reports (1). The sequence of several peptides of human G6PD has already been reported (5,25). As shown in Fig.5, the G6PD sequence deduced from the cDNA agrees with individual peptides of the published sequence (5). The two sequences differ markedly in their overall arrangement (described in detail in the legend to Fig.5). Because of the different methodologies used in DNA and protein sequencing, we consider the DNA-based sequence more likely to be correct.

It has been reported that the amino terminus of G6PD is blocked to Edman degradation (26-28). From our data, the predicted sequence of terminal peptide is M-G-A-S-G-D-L-A-K. Indeed, this peptide was identified in G6PD (5). We are carrying out the analysis of the protein at the NH_2-end of G6PD to define if post-translational modification occurs after the

Table II

	Similarities between yeast and human G6PD										
Yeast	I	D	H	Y	L	G	K	E	L	V	K
									*		*
Human	I	D	H	Y	L	G	K	E	M	V	Q
cDNA	ATC	GAC	CAC	TAC	CTG	GGC	AAG	GAG	ATG	GTG	CAG

Reaction of yeast G6PD with acetylsalicylic acid results in the inactivation of the enzyme and modification of only one lysine residue per subunit (30). The modified lysine is underlined. The difference in amino acids are marked by asterisks. The nucleotide sequence of the corresponding human cDNA is presented under the amino acid sequence.

synthesis. Two different C terminal peptides have been previously reported for G6PD (5,25,29). The sequence of G6PD derived from the cDNA sequence agrees with the C - terminal peptide sequence published by De Scalzi-Cancedda et al. (25).

An 11-residues tryptic peptide from Saccharomyces cerevisiae (30) and an 8-residues peptide from Leuconostoc mesenteroides (31) are the only available data on the sequence of G6PD from other species. Strikingly similar are the sequences of Baker's yeast G6PD peptide, which contains a lysyl residue essential for catalytic activity (Table II), and the peptide underlined in Fig.3. Both differences betweeen the two peptide sequences can be accounted for by two single base pair changes. An identical peptide (underlined in Fig.3) is present also in the amino acid sequence derived from the nucleotide sequence of Drosophila melanogaster (J.F. Manning, personal communication). In addition, considerable sequence homologies exist also in other regions of the human and Drosophila G6PD (preliminary results).

DISCUSSION

The amino acid sequence of the human G6PD provided here has been deduced by extensive sequencing of the coding region of cDNA clones derived from three different sources. Even if the sequence per se does not show any particular feature for the localization of the functional sites, G6P and NADP binding sites, comparison with the sequence of deficient variants and with G6PD of other species will help in the clarification of the structure of the G6PD. This analysis will be important for the understanding of the biological features of this enzyme.

The nucleotide sequence reported in this paper corresponds to the

longest isolated clone (pGD-T-5B). The GGCGGG hexanucleotide regarded as characteristic of the SV40 promoter region, occurs five times in the 5′ untranslated region and its complement CCCGCC occurs four times. These sequences have been observed also in the promoter region (35) as well as in the 5′ UT region (36) of several other genes. S_1 mapping and primer extension experiments are in progress to define the 5′ end of mature G6PD mRNA.

It is to be noted that the 3′ end of the cDNA described in this paper is different from the one previously reported (pGD6405), selected on the basis of its ability to hybridize to G6PD mRNA (6,7). The insert of the clone pGD6405 was used to isolate genomic DNA clones (32 and data to be published elsewhere) and was shown to be located on Xq28 where Gd is mapped (33,34). We have now estabilished that the cDNA presented here and the clone pGD6405, both single copy per haploid genome, are located on the genomic DNA clones about 40 kb apart (data to be published elsewhere). The identification of pGD6405 as a G6PD clone allows many possible explanations. For istance, a common transcription unit starting upstream of the Gd gene can be hypothesized; alternative processing of this common transcript would result in an mRNA containing the G6PD coding portion and different 3′ untranslated regions.

ACKNOWLEDGEMENT

We thank Dr. M. Zasloff for generous hospitality to M.G. Persico; Dr. J. Chou for providing RNA; Dr. J. Skowronski, Dr. F. Gonzales and Dr. H. Okayama for providing the cDNA libraries; Tim Harris for providing the oligonucleotide. We also thank Ms. M. Terracciano, Ms. C. Salzano and Ms. C. Sole for excellent technical assistance. We are grateful to Dr. R. Di Lauro, Dr. R.G. Martin, Dr. B. Paterson, Dr. E. Boncinelli, Dr. F. Graziani, Dr. M. Zasloff and Dr. L. Rosner for helpful discussions and critical reading of the manuscript. This work was supported by Progetto Finalizzato "Ingegneria Genetica e Basi Molecolari delle Malattie Ereditarie" of the Consiglio Nazionale delle Ricerche, Italy, and by a Program Grant by MRC, Great Britain. G.M. was recipient of an EMBO short-term fellowship.

*Permanent address: International Institute of Genetics and Biophysics, CNR, Via Marconi 10, 80125 Naples, Italy

REFERENCES

1. Levy, H.R. (1979) Adv. Enzym. 48, 97-184.
2. Luzzatto, L. and Testa, U. (1978) Current Topics in Haematol. 1, 11-70.
3. Luzzatto, L. and Battistuzzi, G. (1984) Adv. Human. Genet. 14, 217-329.
4. Shreve, D.S. and Levy, H.R. (1977) Biochem. Biophys. Res. Commun. 78, 1369-1375.
5. Beutler, E. (1983) in The Metabolic Basis of the Inherited Diseases, Stanbury, J.B. et al. Eds., 5th edn. pp. 1629-1653, McGraw-Hill, New York.
6. Persico, M.G., Toniolo, D., Nobile, C., D'Urso, M. and Luzzatto, L. (1981) Nature 294, 778-780.
7. Toniolo, D., Persico, M.G., Battistuzzi, G. and Luzzatto, L. (1984) Mol. Biol. Med. 2, 89-103.
8. Hanaham, D. and Meselson, M. (1983) Methods in Enzymology 10, 333-342.
9. Okayama, H. and Berg, P. (1983) Molec. Cell. Biol. 3, 280-289.
10. Andrews, P., Damjanov, I., Simon, D., Banting, G.S., Carling, C., Dracopoli, B. and Fogh, J. (1984) Laboratory Investigation 50, 147-162.
11. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning. A Laboratory Manual. Published by Cold Spring Harbor Laboratory Press, N.Y.
12. Yanisch-Perron, C., Vieira,J. and Messing, J. (1985) Gene 33, 103-109.
13. Maxam, A.M. and Gilbert, W. (1980) Methods in Enzymology 65, 499-560.
14. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5473-5477.
15. Collins, S.J., Gallo, R.C. and Callagher, R.E. (1977) Nature 270, 347-349.
16. Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) Biochemestry 18, 5294-5304.
17. Kolher,P.O. and Bridon, W.E. (1971) J. Clin. Endocrinol. Metab. 32, 683-690.
18. Aviv, H. and Leder, P. (1972) Proc. Natl. Acad. Sci. USA 69, 1408-1412.
19. Lebrach, H., Diamond, D., Worney, J.M. and Boedtker, H. (1977) Biochemistry 16, 4743-4750.
20. Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113, 237-251
21. Reynolds, G.A., Basu, S.K., Osborne, T.F., Chin, D.J., Gil, G., Brown, M.S., Goldstein, J.L. and Luskey, K.L. (1984) Cell 38, 275-285.
22. Kozak, M. (1984) Nucl. Acid Research 12, 857-872.
23. Gorin, M.B., Yancey, S.B., Cline, J., Revel, J.-P. and Horwitz, J. (1984) Cell 39, 49-59
24. Lathe, R. (1985) J. Mol. Biol. 183, 1-12.
25. De Scalzi-Cancedda, F., Caruso, C., Romano, M., di Prisco, G. and Camardella, L. (1984) Biochem. Res. Commun. 118, 332-338.
26. Kahn, A., Bertrand, O., Cottreau, D., Boivin, P. and Dreyfus J.-C. (1976) Biochem Biophys. Acta 445, 537-548.
27. Weiner, A.M., Platt, T. and Weber, K. (1976) J. Biol. Chem. 247, 3242-3251.
28. Yoshida, A. (1972) Anal. Biochem. 49, 320-325.
29. Kahn, a., Bertrand, O., Cottreau, D., Boivin, P. and Dreyfus, J.-C. (1977), Biochem. Biophys. Res. Commun. 77, 65-72.
30. Jeffrey, J., Hobbs, L. and Jornvall, H. (1985) Biochemestry 24, 666-671.
31. Haghighi, B., Flynn, T.G. and Levy, H.R. (1982) Biochemestry 21, 6415-6420.
32. Toniolo, D., D'Urso, M., Martini, G., Persico, M., Tufano, V., Battistuzzi, G. and Luzzatto, L. (1984) The EMBO J. 3, 1987-1995.

33. Szabo, P., Purrello, M., Rocchi, M., Archidiacono, N., Alhadeff, B., Filippi, G., Toniolo, D., Martini, G., Luzzatto, L. and Siniscalco, M. (1984) Proc. Natl. Acad. Sci. USA 81, 7855-7859.
34. Martin-De Leon, P.A., Wolf, S.F., Persico, G., Toniolo, D., Martini, G. and Migeon, B.R. (1985) Cytogenet. Cell Genet. 39, 87-92.
35. Dynan, W.S., Sazer, S., Tjian, R. and Shimke, R.T., (1986) Nature 319, 246-248.
36. Hall, A. and Brown, R. (1985) Nucleic Acid Res. 14, 5255-5268.
37. Proudfoot, N.J. and Brownlee, G.C. (1976) Nature 263, 211-214.
38. Tinoco, I.Jr., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973) Nature 246, 40-41.