Phosphoglucomutase 1: a gene with two promoters and a duplicated first exon

Wendy PUTT, Jane H. IVES, Martine HOLLYOAKE, David A. HOPKINSON, David B. WHITEHOUSE and Yvonne H. EDWARDS*

MRC Human Biochemical Genetics Unit, The Galton Laboratory, University College London, 4 Stephenson Way, London NW1 2HE, U.K.

In view of its central role in glycolysis and gluconeogenesis and its polymorphic genetic variability, the phosphoglucomutase 1 (PGM1) gene in man has been the target of protein structural studies and genetic analysis for more than 25 years. We have now isolated genomic clones containing the complete PGM1 gene and have shown that it spans over 65 kb and contains 11 exons. We have also shown that the sites of the two mutations which form the molecular basis for the common PGM1 protein polymorphism lie in exons 4 and 8 and are 18 kb apart. Within this region there is a site of intragenic recombination. We have discovered two alternatively spliced first exons, one of which,

exon 1A, is transcribed in a wide variety of cell types; the other, exon 1B, is transcribed in fast muscle. Exon 1A is transcribed from a promoter which has the structural hallmarks of a housekeeping promoter but lies more than 35 kb upstream of exon 2. Exon 1B lies 6 kb upstream of exon 2 within the large first intron of the ubiquitously expressed PGM1 transcript. The fast-muscle form of PGM1 is characterized by 18 extra amino acid residues at its N-terminal end. Sequence comparisons show that exons 1A and 1B are structurally related and have arisen by duplication.

INTRODUCTION

Phosphoglucomutase (PGM; EC 5.4.2.2) is a phosphotransferase which functions in both glycolysis and gluconeogenesis by catalysing the transfer of phosphate between the 1- and 6-positions of glucose. The isoenzymes of PGM are all monomeric with a molecular size of about 60 kDa, and in man are encoded by a family of three, possibly four, structurally related genes, encoding PGM1, PGM2, PGM3 and PGM4 (Parrington et al., 1968; Harris and Hopkinson, 1976). In most cell types the isoenzymes of PGM1 predominate, representing about 90 % of the total PGM activity, except in red cells where PGM2 is a major isoenzyme, and the lactating mammary gland where PGM4 is the sole isoenzyme (Cantu and Ibarra, 1982). It is suspected that PGM2 and PGM3 may have other major roles in metabolism which are so far unrecognized, a view supported by the finding that PGM2 shows a remarkable phosphopentomutase activity unique to this isoform (Quick et al., 1972).

In view of its central role in metabolism and its widespread tissue distribution, PGM1 has been the target of structural and genetic studies. The complete amino acid sequence of rabbit PGM1 was reported in 1983 (Ray et al., 1983), and more recently the crystal structure has been determined at 2.7 Å (0.27 nm) resolution (Dai et al., 1992). These studies show that PGM1 is arranged into four α/β -domains with the active site buried at the confluence of the domains. We have recently isolated the cDNAs for both rabbit and human PGM1 (Whitehouse et al., 1992) and comparisons of the nucleotide sequences show that, in addition to the active site region, the entire coding sequence from the two species is highly conserved (92 % sequence identity).

PGM1 has also achieved prominence as a key marker in genetic linkage mapping and in forensic science (Sensabaugh, 1982) due to the very high level of both common and rare genetic variation shown by the protein (Roychoudhury and Nei, 1988). We have identified two point mutations at separate sites in the

PGM1 sequence which give rise in various combinations and by intragenic recombination to the 10 common phenotypes characteristic of this locus worldwide (March et al., 1993).

In this paper we describe the complete structure of the PGM1 gene and discuss its architecture in relation to the domain structure of the protein product. We consider the sites of mutation underlying the genetic polymorphism and the location of a site of recombination within the PGM1 gene. We also describe the occurrence of an alternative first exon in the PGM1 gene which is transcribed from a promoter distinct from the usual promoter. This latter finding explains the molecular origins of an unusual transcript of PGM1 found in rabbit fast skeletal muscle and recently described by Young Sup Lee et al. (1992).

MATERIALS AND METHODS

Genomic DNA clones for human PGM1 were isolated from a human genomic library constructed in the vector λ 2001 (LeFranc et al., 1986) using the PGM1 cDNA as probe (Whitehouse et al., 1992). Restriction enzyme fragment mapping of insert DNA was carried out using standard procedures. Fragments were ligated into M13mp18 and transfected into JM101. DNA sequences across exon/intron junctions and in flanking regions were determined by the dideoxy chain termination method of Sanger et al. (1977) and Biggin et al. (1983). The distances between all the exons were estimated by restriction mapping, and those between exons 2, 3, 4, 5, 6 and 7 were confirmed by PCR amplification using primers derived from both exon and intron sequences.

The position of exon 1B in the human genomic sequence was determined by PCR using the rabbit oligonucleotide primers, designated RPGM1-FM1 and RPGM1-FM2, based on the rabbit cDNA sequence (Young Sup Lee et al., 1992), and various human genomic λ clones (see Figure 1) as templates. The PCR conditions were 15 s at 94 °C, 30 s at 53 °C and 30 s at 72 °C for

Figure 1 Physical map of the human PGM1 gene

Exons are indicated by filled boxes. The mutations which give rise to the PGM1 2/1 and PGM1 +/− alleles are in exons 4 and 8 respectively and the distances between these sites are indicated at the top of the figure. The distances between exons 1A and 1B and between exons 1A and 2 are indicated in a similar manner. The positions of an Alu repeat element and a (CA)_n repeat are marked by arrows. The nine λ clones used to construct this map are shown below the map. Restriction endonuclease sites: Sm, *SmaI*; H, *HindIII*; X, *XbaI*; B, *BamHI*; S, *SacI*.

35 cycles. Exon 1B was sequenced using the same primers and single-stranded DNA prepared from a 2.4 kb *XbaI* fragment from PGMλ.3. Internal primers designated HPGM1-FM1 and HPGM1-FM2 were used to sequence the 5′ and 3′ ends of exon 1B and flanking sequences. The primer sequences are as follows: RPGM1-FM1, TGGAAGAAGGTCCTCTCCCTCT; RPGM1-FM2, ATTGGCTGCCGCCATCTGCAAT; HPGM1-FM1, AGTGCAGATGGCAGCTG; HPGM1-FM2, CTCCAGATAGCATGGCT.

RESULTS

The human PGM1 gene: isolation and exon/intron structure

Nine recombinant clones, PGM1λ.1–9, which spanned the human PGM1 gene were isolated. DNA fragments containing PGM1 exons were identified by Southern blot analysis of single and double *XbaI*, *SacI* and *XhoI* digests of each clone using the cDNA (Whitehouse et al., 1992) as probe. These fragments were subcloned into M13mp18 and pUC18 and the precise locations of the exon/intron boundaries were determined by further restriction enzyme mapping and sequence analysis.

The detailed restriction map of the whole PGM1 gene shown in Figure 1 was obtained by isolating *SacI* and *XbaI* fragments from each clone and further digestion with *HindIII*, *SmaI*, *EcoRI*, *AccI*, *BamHI*, *HincII*, *KpnI*, *SalI*, *SphI* and *PstI* (for clarity only the positions of *SacI*, *XbaI*, *HindIII*, *SmaI* and *BamHI* sites are shown in Figure 1).

These analyses show that the gene for human PGM1 is more than 65 kb long and that more than half of this length is attributed to an unusually large first intron. The ubiquitously expressed PGM1 comprises 11 exons, of which the most 5′ is designated exon 1A in Figure 1 and Table 1. The sequences at the 5′ and 3′ junction of each intron and exon are shown in Table 1. The exons vary in size between 116 bp (exon 7) and 659 bp (exon 11). The introns vary in size from 0.55 kb (IVS 2) to > 35 kb (IVS 1). The very large IVS 1 means that exon 1A and exon 2 are separated from each other in unspliced RNA by an unusually large distance.

All of the intron/exon junctions conform to the consensus eukaryotic structure beginning with the dinucleotide GT and ending with AG, except IVS 3 which ends in AG but begins with GC. This dinucleotide at the 5′ splice site is a rare but recognized alternative; in a survey of 3700 splice junctions there were nine examples of non-conforming 5′ splice sites, five of which were GC (Shapiro and Senapathy, 1987). At this splice site PGM1 has C at −3, where C and A can occur.

5′ end of the PGM1 gene

The upstream sequences of the PGM1 gene (Figure 2) exhibit a number of features which are considered to be characteristic of housekeeping genes; in particular, this region is G + C-rich (Bird, 1986). In the 900 bp region from −240 bp to +600 bp the G + C content is 70 %, compared with 40 % in bulk DNA. The incidence of CpG dinucleotides is also high such that the CpG/GpC ratio is 0.79, compared with 0.10 for bulk DNA. This is reflected in the occurrence of multiple ($n = 11$) sites for the restriction enzyme *HpaII* (CCGG), the hallmark of so-called ‘HTF islands’ described by Bird (1986).

The elements normally associated with genes transcribed by RNA polymerase II, i.e. the TATA box and the CCAAT box, do not occur, a feature noted before in a number of housekeeping genes. However, the core consensus binding element for the ubiquitous transcription factor Sp1, CCCGCC, occurs six times in the 900 bp region around exon 1A. Three of these sites are in the immediate 5′ flanking region; two are contiguous in exon 1A and one is in IVS 1 (Figure 2).

3′ end of the PGM1 gene

The most 3′ exon, 11, contains 572 bp of untranslated sequence with a polyadenylation signal, AATAAA, 20 bp from the poly(A) addition site. The nucleotide sequence extending 250 bp downstream of the 3′ end of the gene was determined and shown to contain an element TTGTCGTCTT, at −27 bp from the poly(A) addition signal, which is similar to the consensus sequence YGTGTTY required for efficient formation of mRNA 3′ termini (McLauchlan et al., 1985).

Restriction mapping data suggest that a CpG island occurs 3′ to the human PGM1 gene. Within 3–5 kb downstream from the 3′ end of the PGM1 gene, three sites for the rare cutter *SmaI* occur. This region has not yet been sequenced but the relatively high CpG content is marked by multiple *HpaII* sites in the small *SmaI* DNA fragments generated by digestion of cloned genomic DNA from this region (Figure 3). We have not yet determined whether this CpG island is integral to the PGM1 gene or whether it is associated with an as yet unidentified gene lying downstream of PGM1.

The fast-muscle isoform of PGM1

In addition to the ubiquitously expressed isoform of PGM1, a second isoform, which is expressed in fast skeletal muscle, has also recently been described in the rabbit (Young Sup Lee et al.,

Table 1 Structural details of the PGM1 gene

Each exon/intron junction is illustrated by giving on the left the 3' end of an exon and on the right the 5' sequences beginning the next exon (capital letters). The corresponding amino acids are given above and are numbered in terms of the complete human PGM1 amino acid sequence (Whitehouse et al., 1992). The beginning and end of each intron are shown in lower case lettering. The sizes of each exon and intron (IVS) are given in parentheses. For exon 1B, amino acid numbering is in terms of the derived amino acid sequence for the PGM1-FM isoform, such that amino acid 97 in PGM1-FM 1B corresponds to amino acid 80 in PGM1 1A.

Exon	3' sequence	Intron sequences	5' sequence of next exon
	80		85
Exon 1A (304 bp) (58 bp non coding)	Ala Ala Ala Asn Gly GCT GCC GCC AAC GGG	gtaagggatgcgcgccccc	IVS 1 (> 35 kb) ttctggatttcttctcctag Ile Gly Arg Leu ATC GGT CGC TTG
	135		139
Exon 2 (163 bp)	Ile Ser Asn Gly G ATT TCT AAT GGA G	gtgagtttgctgtcattttg	IVS 2 (0.55 kb) gtttgctgtttggtttccag ly Pro Ala Pro GT CCT GCT CCA
	182		188
Exon 3 (147 bp)	Lys Pro Phe Thr V AAA CCC TTC ACA G	gcatgtttacttttctctc	IVS 3 (1.1 kb) gatgttgcttgtttctcacag al Glu Ile Val TG GAA ATT GTG
	226		230
Exon 4 (126 bp)	Ala Met His Gly V GCT ATG CAT GGA G	gtatacaatcattttcttttc	IVS 4 (3 kb) ttattccatgctgtatatag al Val Gly Pro TT GTG GGA CCG
	289		294
Exon 5 (191 bp)	Phe Asp Gly Asp Gly TTT GAT GGA GAT GGG	gtgggtataagtgcatttaa	IVS 5 (1.3 kb) gatcctctcttctccccaag Asp Arg Asn Met GAT CGA AAC ATG
	341		345
Exon 6 (155 bp)	Gly Ala Leu Asp Ar GGT GCT CTG GAC CG	gtaggtgtctccattccctt	IVS 6 (2.3 kb) acttagttcagccttgaaag g Val Ala Ser G GTG GCT AGT
	380		384
Exon 7 (116 bp)	Ser Phe Gly Thr G AGC TTC GGG ACC G	gtaagtcacactcctgtgct	IVS 7 (9.1 kb) gcttgetgtccccctccag ly Ser Asp His GT TCT GAC CAC
	425		429
Exon 8 (136 bp)	Asn Phe Phe Thr Ar AAT TTC TTC ACC AG	gtgagccacagcccagctgg	IVS 8 (3.2 kb) aacttcgtgacttcttccag g Tyr Asp Tyr G TAT GAT TAC
	486		491
Exon 9 (184 bp)	Ile Ser Arg Asn Gln ATT TCA AGA AAT CAG	gtagaaacagaccggtgtgt	IVS 9 (2.8 kb) tctctgcttatcttttccag Gly Leu Arg Leu GGC TTG CGC CTC
	541		546
Exon 10 (133 bp)	Asn Gln Asp Pro Gln AAC CAG GAC CCC CAG	gtaacgcccagccctgtgcc	IVS 10 (5.1 kb) tctctctatgtcttctcctag Val Met Leu Ala GTC ATG TTG GCC
Exon 11 (659 bp) (572 bp non coding)			
	98		103
Exon 1B (684 bp) (384 bp non coding)	Ala Ala Ala Asn Gly GCA GCT GCC AAT GGG	gtgggtatatgataagtatt	IVS 1 (6 kb) ttctggatttcttctcctag Ile Gly Arg Leu ATC GGT CGC TTG

1992). Comparison of the published cDNA sequences for the ubiquitous PGM1 (Whitehouse et al., 1992) and the fast skeletal muscle isoform, here designated PGM1-FM, shows that the two sequences are identical except at their 5' ends. This suggests two possible explanations for the occurrence of the PGM1-FM isoform: either it is a transcript from a gene which is closely related but distinct from PGM1, or it is transcribed from the same PGM1 gene but is processed differently at the 5' end.

Support for the second hypothesis comes from the recognition that the PGM1-FM and PGM1 sequences diverge at precisely the boundary between exon 1A and IVS 1, which suggests the possibility of alternative first exons in the PGM1 gene. In order to test this idea we searched the entire human PGM1 gene for the 5' sequence unique to PGM1-FM. PCR amplification of the human PGM1 clones using primers derived from the unique 5' region of the rabbit sequence identified the product in PGM1.3 and PGM1.4 (see Figure 1 and Figure 4). This was subsequently isolated in a 2.4 kb *Xba*I fragment which was subcloned and sequenced. Detailed analysis confirmed the presence of PGM1-

FM sequence, designated exon 1B, within the human PGM1 gene at a position more than 29 kb downstream of exon 1A (the first exon of the ubiquitous PGM1 transcript) and 6 kb upstream of exon 2, which is common to both PGM1-FM and PGM1.

In their investigation of the rabbit fast-muscle isoform cDNA, Young Sup-Lee et al. (1982) designated a particular methionine codon, which roughly corresponds to the terminus of the PGM1 isoform, as the translation start site of this isoform. However, closer inspection of the cDNA sequence and comparison with the human sequence in this region suggests a start site further upstream, effectively extending the protein coding sequence by a further 42 bp and the amino acid sequence by 14 residues, (Figures 5 and 6). This start site seems to be realistic, since it is embedded in very similar sequences, TGAAAGATGA in the rabbit and TGAAGGATGA in man, and both motifs conform to examples described by Kozak (1987) for initiation of translation.

Thus we conclude that exon 1B (PGM1-FM) differs from exon 1A (PGM1) by coding for 18 extra amino acids at the N-terminal

```

-240 gcctcttttt acagtggata ggattccgaa gggagagtaa atcagtaagg
-190 ctccacccat ttcacaggag gcctagagac cggcgctccc ctttccctc
-140 ccgccgacct gccaggaggt gggctggcgc ggagggaggg cctgtcccc
-90 tgtcccttta aggaggaggg ccaaagccg gcctagagtg cgcgtagccc
-40 ccacccgccg tgcctcacc ccagagcagc tgcagcctca GCCGGCCGCC
 11 CCTCCGCCAG CCAAGTCCGC CGCTCTGACC CCCGGCAGCA AGTCGCCACC
 61 ATGGTGAAGA TCGTGACAGT TAAGACCCAG GCGTACCAGG ACCAGAAGCC
111 GGGCACGAGC GGGCTGCGGA AGCGGGTGAA GGTGTTCCAG AGCAGCGCCA
161 ACTACGCGGA GAACTTCATC CAGAGTATCA TCTCCACCGT GGAGCCGGCG
211 CAGCGGCAGG AGGCCACGCT GGTGGTGGGC GGGGACGGCG GGTTCTACAT
261 GAAGGAGGCC ATCCAGCTCA TCGCTCGCAT CGCTGCCGCC AACGGGgtaa
311 gggatgcgcg gcccccccg cgtgtgcacc ctggcgctg tgcgacgtgc
361 ggcccgcggc gccctccctc gctcggggcc ggccgcttcc gcgcgtgcc
411 gcctcggttt ccacctcccg ctctccctc tccttcgcgc tcgctctct
461 ggctcggagg ccgacggagg cggccggctg ggaaagtggc cacggaccgg
511 gcagacctgc tcgcctgact cccggggcgc ggggaggtgc gggcgcgagg
561 ccaccaagg tcacgcccga ctctccgtgt ctaccgctg ctttccctct
611 gccctcccc cgccccctcc gaggcgtctg acatttgccc taaccttgca

```

Figure 2 Nucleotide sequence at the 5' end of the PGM1 gene and encompassing exon 1A

The number 1 indicates the extreme 5' end of the PGM1 cDNA (Whitehouse et al., 1992). Exon 1A sequence is given in upper case lettering. The ATG initiation codon is underlined and six CCCGCC-Sp1 elements are shown in bold type.

Figure 3 *Smal* and *HpaII* digests of the 8 kb *SacI* fragment at the 3' end of the human PGM1 gene

The 5' *SacI* site of this fragment corresponds to the most 3' *SacI* site shown in Figure 1. *Smal* (Sm) digestion cuts the 8 kb parent (P) band into four smaller fragments of 2.3 kb, 1.3 kb, 1.2 kb and 3.2 kb. Each of these contains more than one *HpaII* (Hp) site.

end (Figure 6). There are, however, striking similarities between the two exons, with 58 % identity between the DNA sequences and 51 % identity, with 74 % similarity, between the amino acid sequences (Figure 6). Furthermore, there are three relatively large blocks of complete amino acid identity comprising 11, 5 and 5 contiguous residues. The clear sequence similarities between exons 1A and 1B indicate that they are related to each other and are likely to have arisen by duplication.

The 5' untranslated region of rabbit PGM1-FM is relatively very long (384 bp) in comparison with those of rabbit PGM1

Figure 4 PCR amplification of DNA from the human genomic clones PGMλ.1-9 using the RPGM1-FM specific primers

The product of 254 bp was amplified in PGMλ.3 and PGMλ.4. C is a no-DNA control.

(39 bp) and human PGM1 (62 bp). The sequence appears to be conserved, since there is 71 % identity between rabbit and human in this region.

Our very preliminary studies suggest that the expression of the PGM1-FM transcript will be limited to a restricted number of tissues (J. H. Ives and Y. H. Edwards, unpublished work), and thus would be transcribed from a non-housekeeping promoter. In keeping with this view, and in contrast to the promoter flanking exon 1A, the sequences immediately flanking exon 1B are not G+C-rich (47 %) and have a relatively low CpG/GpC ratio (0.24). Furthermore, there are no binding sites for the Sp1 transcription factor (Figure 5). No TATA box or CAAT box could be identified. In those cases where these elements are absent, commonly in housekeeping promoters and occasionally in tissue-specific promoters, it has been proposed that a 'tethering factor' functions to recruit the TFIID (transcription factor IID) complex to the promoter (Gill and Tjian, 1992, and references cited therein).

```

-274 ttgcatgtgt gaagtaaaaa taatttctta ggtttcaaaa tgacagtgat
-224 gtaagaaaac caggggtaaa ttaagctccc tcacttccat ttttaagattg
-174 acctttaact tgctaggtgg catggggcaa ctttagctgc agatggctct
-124 cagagatgat gccagagagg ctctggttta agggatggca ggagcagtga
-74 gagccacacc aggggtgtctt gttcacaat ctgcttcgtg ggactgcatt
-24 tgccctcttt gcctgggctc tgagAGTGGA GCCTGTCAGC CACCTGGTGT
27 TACAGCTGAA GTTGGCTGCA TGCAGCACTC CTTGGAGCTA TCTCTACCTA
77 ACGGAATGAG ACAAATAAT TGAAGGGTAT TTGGCTTGGT TAAGTGGGCA
127 GAAGTGTCTT CATCAAAACA ACTTTGAAGA CCAGCTTTGT GTGTGGGTGC
177 GAGTGGGTGT GTGGCCCTTA ACTTGTTCAC AAGGGACAAC AACACAAGA
227 ATTAACCAAG CTTTCTCATT GAGAGTGATC GTCCATGCAA ACCCTCTATT
277 TTAGTTACTC ATACGACACT GTTGCAGCTT CAGGTGGGAC AGACCTATTT
327 TGGAGTCCCT TGACTGCTTT CTGGACACCC TCACTGAGCC AGTCGGTGGG
377 AGTGAAGGAT GAGTGATTTT GAAGAATGGA TTTCTGGGAC ATATAGAAAA
427 ATGGAAGAAG GTCCTCTCCC TCTGTGACT TTTGCTACAG CTCCCTACCA
477 CGATCAGAAA CCAGGAACAA GTGGATTACG GAAGAAAACC TATTATTTTG
527 AGGAAAAGCC ATGCTATCTG GAGAATTTCA TCCAGAGTAT ATTCTTTTCC
577 ATAGACCTAA AAGATCGCCA GGGATCATCA CTGGTGGTTG GTGGAGATGG
627 GCGGTACTTT AATAAATCAG CAATAGAAAC AATAGTGCAG ATGGCAGCTG
677 CCAATGGGgt gggtatatga taagtattgt tatgtttctg gctctccaaa
727 taaccttaat tttacagaag tatacaaacg tacaaggatt tatatgtggt

```

Figure 5 Nucleotide sequence encompassing exon 1B

The transcribed sequence is shown in upper case and the proposed ATG initiation methionine is underlined. The transcription start site, numbered 1, has been defined by comparison with the published rabbit cDNA sequence (Young Sup Lee et al., 1992).

Figure 6 Amino acid sequence comparisons

(a) Comparison of the amino acid sequences encoded by exon 1B in human and rabbit PGM1-FM. Only those residues in the rabbit sequence which differ from those in the human sequence are shown. (b) Comparison of the amino acid sequences encoded by human exon 1A and exon 1B, using the BESTFIT sequence alignment program. Number 1 refers to the initiation methionine in exon 1B. Bold vertical lines indicate complete identity, * indicates a degree of similarity.

Search for mutation in the exon 1B sequence

We have recently sequenced in full PGM1 cDNAs (including the ubiquitously used exon 1A) from individuals of known PGM1 phenotype and have identified the molecular basis of the common PGM1 protein polymorphism in exons 4 and 8 (March et al., 1993). In order to complete our genetic analysis of the PGM1 gene, exon 1B has been fully sequenced using DNA from five individuals of various PGM1 protein phenotypes: 1+, 2+

(n = 2), 2-1-, and 1+1-. No mutations were found in the 1B exon in these individuals.

DISCUSSION

When the architecture of a gene is analysed in relation to the three-dimensional structure of its protein product, a mixed picture emerges. For certain genes such as those encoding alcohol dehydrogenase, serum albumin and the immunoglobulins (Tonegawa, 1983; Brändén et al., 1984; Minghetti et al., 1986), the exons coincide with the structural units of multidomain proteins, whereas for others such as those coding for amylase, α_1 -antitrypsin and carbonic anhydrase (MacDonald et al., 1980; Leicht et al., 1982; Venta et al., 1985), they do not. Dai et al (1992) have described a four-domain structure for PGM1, with three domains containing a characteristic $\alpha_3\beta_4\alpha_1$ motif and a fourth with a different β -sheet topology. The amino acid residues which flank the domains are as follows: domain I, Val-1 to Asp-189; II, Ser-190 to Phe-301; III, Phe-302 to Gly-420; IV, Arg-421 to Thr-561. There is no evidence that the PGM1 exons encode polypeptide domains that fold independently, since the positions of the domain boundaries do not coincide with any of the intron/exon junctions. It is of interest, however, that all but one of the intron/exon junctions fall outside the extensive α -helices, perhaps pointing to the possibility that the individual exons were once micro-genes (Seidel et al., 1992).

Our studies have shown that there are two types of PGM1 transcript; one is ubiquitously expressed whereas the other, PGM1-FM, is found in fast muscle and at relatively low levels. The differences in their patterns of expression and in their 5' leader sequences strongly suggest that the two transcripts are transcribed from distinct promoters. The occurrence of alternative first exons transcribed from independent promoters is

relatively common and is found for example in the amylase, α_1 -antitrypsin and aldolase genes (Young et al., 1981; Perlino et al., 1987; Izzo et al., 1988). In some instances, for example aldolase (Izzo et al., 1988) and porphobilinogen deaminase (Chretien et al., 1988), one of the alternative promoters is housekeeping and the other tissue-specific; this also appears to be the case for PGM1.

As yet, the functional significance of the PGM1-FM transcript and its protein product is not clear. The long 5' untranslated region may be of some importance; it is not G+C-rich and so is probably not highly structured, and such regions are believed to facilitate translation by allowing the queuing of several 40 S rRNA molecules (Kozak, 1991). However, there is no evidence for favoured translation in rabbit skeletal muscle since only one out of four clones isolated from the cDNA library corresponded to the PGM1-FM transcript (Young Sup Lee et al., 1992).

PGM1 appears to be unusual in that the alternative first exon encompasses both untranslated and coding sequence and appears to have arisen by intragenic gene duplication. Such duplication has been a relatively common event in gene evolution and is implicated in the origin of the present-day structures and numerous genes, including those for albumin, transferrin and haptoglobin, for example (Bowman and Kurosky, 1982; Murray et al., 1983; Park et al., 1985). As yet we have not defined the boundaries of the duplication event in the PGM1 gene. A preliminary comparison of the genomic sequences immediately flanking exon 1A and exon 1B reveals only 38% identity in these regions. It is of interest to note, however, that the distance between these two exons would be consistent with a complete gene duplication followed by the loss of functional exons 2–11 in the upstream copy. This possibility is being investigated and our very preliminary studies suggest that sequences homologous to exon 2 occur 4 kb downstream of exon 1A within the large first intron.

The occurrence of such a large first intron in the ubiquitously expressed PGM1 transcript raises issues concerning the transcription and splicing of the PGM1 mRNA. The major question is how the correct donor and acceptor splice sites are recognized over a long distance. This problem is of course by no means unique; long first introns have been described for red cell carbonic anhydrase I and the c-abelson gene, and brain dystrophin has an extremely large first intron of 400 kb (Bernards et al., 1987; Fraser et al., 1989; Boyce et al., 1991). It seems likely that the transcription of such genes might be discontinuous, involving looping of chromatin to juxtapose the correct splice sites, but there is still a need for experimental evidence to support this idea.

Our recent studies (March et al., 1993) of the PGM1 mRNA sequence in the human population have confirmed the prediction that the PGM1 protein polymorphism arose by a combination of point mutation and recombination (Carter et al., 1979; Takahashi et al., 1982). Our present genomic structural studies have shown that the sites of the two common mutations at positions 723 bp and 1320 bp lie in exons 4 and 8 respectively and are thus 18 kb apart. This distance presumably encompasses the site of recombination. Preliminary investigation of this region has revealed a non-polymorphic (CA)_n repeat. It is not yet clear whether this is of functional significance, but poly(CA) repeats are known to form Z DNA (Wahls et al., 1990) and promote recombination (Treco and Arnheim, 1986). We are currently working towards identifying more precisely the site of recombination by structural analysis and by allelic association studies.

Note added in proof (received 13 October 1993)

Since this paper was accepted for publication, the genomic structure of human PGM1 has been completed. The size of IVS 1 is 38.5 kb.

M.H. is supported by a grant from the Home Office, U.K.

REFERENCES

- Bernards, A., Rubin, C. M., Westbrook, C. A., Paskind, M. and Baltimore, D. (1987) *Mol. Cell. Biol.* **7**, 3231–3236
- Biggin, M. D., Gibson, T. J. and Hong, H. F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3963–3965
- Bird, A. (1986) *Nature (London)* **321**, 209–213
- Bowman, B. H. and Kurosky, A. (1982) *Adv. Hum. Genet.* **12**, 189–193
- Boyce, F. M., Beggs, A. H., Feener, C. and Kunkel, L. M. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88**, 1276–1280
- Brändén, C.-I., Eklund, H., Cambillau, C. and Pryor, A. J. (1984) *EMBO J.* **3**, 1307–1310
- Cantu, J. M. and Ibarra, B. (1982) *Science* **216**, 639–640
- Carter, N. D., West, C. M., Emes, E., Parkin, B. and Marshall, W. H. (1979) *Ann. Hum. Biol.* **6**, 221–230
- Chretien, S., Dubart, A., Beaupain, D., Raich, N., Grandchamp, B., Rosa, J., Goossens, M. and Romeo, P.-H. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 6–10
- Dai, J. B., Liu, Y., Ray, W. J., Jr. and Konno, M. (1992) *J. Biol. Chem.* **267**, 6322–6337
- Fraser, P., Cummings, P. and Curtis, P. (1989) *Mol. Cell. Biol.* **9**, 3308–3313
- Gill, G. and Tjian, R. (1992) *Curr. Opin. Genet. Dev.* **2**, 236–242
- Harris, H. and Hopkinson, D. A. (1976) *Handbook of Enzyme Electrophoresis in Human Genetics*, North-Holland, Amsterdam
- Izzo, P., Costanzo, P., Lupo, A., Rippa, E., Paolella, G. and Salvatore, F. (1988) *Eur. J. Biochem.* **174**, 569–578
- Kozak, M. (1987) *Nucleic Acids Res.* **15**, 8125–8147
- Kozak, M. (1991) *J. Cell Biol.* **115**, 887–903
- LeFranc, M.-P., Forster, A., Baer, R., Stenson, M. A. and Rabbitts, T. H. (1986) *Cell* **45**, 237–246
- Leicht, M., Long, G. L., Chandra, T., Kurachi, K., Kidd, V. J., Mace, M., Davie, E. W. and Woo, S. L. C. (1982) *Nature (London)* **297**, 655–659
- MacDonald, R. J., Crerar, M. M. and Swain, R. L. (1980) *Nature (London)* **287**, 117–122
- March, R. E., Putt, W., Hollyoake, M., Ives, J. H., Lovegrove, J. U., Hopkinson, D. A., Edwards, Y. H. and Whitehouse, D. B. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90**, in the press
- McLauchlan, J., Gaffney, D., Whitton, J. L. and Clements, J. B. (1985) *Nucleic Acids Res.* **13**, 1347–1368
- Minghetti, P. P., Ruffner, D. E., Kuang, W.-J., Dennison, D. E., Hawkins, J. W., Beattie, W. G. and Dugaiczky, A. (1986) *J. Biol. Chem.* **261**, 6747–6757
- Murray, J. C., Demopoulos, C. M., Lawn, R. M. and Motulsky, A. G. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 5951–5955
- Park, I., Schaeffer, E., Sidoli, A., Baralle, F. E., Cohen, G. N. and Zakin, M. M. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 3149–3153
- Parrington, J. M., Cruickshank, G., Hopkinson, D. A., Robson, E. B. and Harris, H. (1968) *Ann. Hum. Genet.* **32**, 27–34
- Perlino, E., Corlese, R. and Ciliberto, G. (1987) *EMBO J.* **6**, 2767–2771
- Quick, C. B., Fisher, R. A. and Harris, H. (1972) *Ann. Hum. Genet.* **35**, 445–448
- Quick, C. B., Fisher, R. A. and Harris, H. (1974) *Eur. J. Biochem.* **42**, 511–517
- Ray, W. J., Jr., Hermodson, M. A., Jr., Puvathingal, J. M. and Mahoney, W. C. (1983) *J. Biol. Chem.* **258**, 9166–9174
- Roychoudhury, A. K. and Nei, M. (1988) *Human Polymorphic Genes World Distribution*, pp. 105–110, Oxford University Press, New York
- Sanger, F., Nicklen, S. and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467
- Seidel, H. M., Pompliano, D. L. and Knowles, J. R. (1992) *Science* **257**, 1489–1490
- Sensabaugh, G. F. (1982) *Forensic Science Handbook*, pp. 338–415
- Shapiro, M. B. and Senapathy, P. (1987) *Nucleic Acids Res.* **15**, 7155–7175
- Takahashi, N., Neel, J. V., Satoh, C., Nishizaki, J. and Masunari, N. (1982) *Proc. Natl. Acad. Sci. U.S.A.* **79**, 6636–6640
- Tonegawa, S. (1983) *Nature (London)* **302**, 575–581
- Treco, D. and Arnheim, N. (1986) *Mol. Cell. Biol.* **6**, 3934–3947
- Venta, P. J., Montgomery, J. C., Hewett-Emmett, D., Wiebauer, K. and Tashian, R. E. (1985) *J. Biol. Chem.* **260**, 12130–12135
- Wahls, W. P., Wallace, L. J. and Moore, P. D. (1990) *Mol. Cell. Biol.* **10**, 785–793
- Whitehouse, D. B., Putt, W., Lovegrove, J. U., Morrison, K., Hollyoake, M., Fox, M. F., Hopkinson, D. A. and Edwards, Y. H. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 411–415
- Young, R. A., Hagenbuchle, O. and Schibler, U. (1981) *Cell* **23**, 451–458
- Young Sup Lee, Marks, A. R., Gureckas, N., Lacro, R., Nadal-Ginard, B. and Kim, D. H. (1992) *J. Biol. Chem.* **267**, 21080–21088