ARTICLE

Comparative analysis of the phosphomannomutase genes PMM1, PMM2 and PMM2ψ: the sequence variation in the processed pseudogene is a reflection of the mutations found in the functional gene

Els Schollen1, Els Pardon1, Leen Heykants1, Jan Renard1, Norman A. Doggett2, David F. Callen3, Jean-Jacques Cassiman1 and Gert Matthijs1,*

1Center for Human Genetics, University of Leuven, Campus Gasthuisberg O&N 6, Herestraat 49, B-3000 Leuven, Belgium, 2Life Science Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA and 3Department of Cytogenetics and Molecular Genetics, Women’s and Children’s Hospital, North Adelaide, SA 5006, Australia

Received September 23, 1997; Revised and Accepted November 21, 1997

The search for the carbohydrate-deficient glycoprotein syndrome type I (CDG1) gene has revealed the existence of a family of phosphomannomutase (PMM) genes in humans. Two expressed PMM genes, PMM1 and PMM2, are located on chromosome bands 22q13 and 16p13, respectively, and a processed pseudogene PMM2ψ is located on chromosome 18p. Mutations in PMM2 are the cause of CDG type IA whereas no disorder has been associated with defects in PMM1 as yet. Here, we describe the genomic organization of these paralogous genes. There is a 65% identity of the coding sequence, and all intron/exon boundaries have been conserved. The processed pseudogene is more closely related to PMM2. Remarkably, several base substitutions in PMM2 that are associated with disease are also present at the corresponding positions in the pseudogene. Thus, mutations that occur at a slow rate in the active gene in the population have also accumulated in the pseudogene.

INTRODUCTION

Carbohydrate-deficient glycoprotein syndrome type I (CDG1) is the paradigm of a group of genetic multi-system disorders characterized by a deficiency in the glycosylation pathway [see (1) for a recent review]. CDG1 is an autosomal recessive disorder, with a major disease locus located on chromosome band 16p13. Van Schaftingen and Jaeken (2) have previously identified a deficiency of phosphomannomutase (PMM) activity in CDG1 patients. The enzyme isomerizes mannose 6-phosphate into mannose 1-phosphate, which is then converted into GDP-mannose. The latter is required for the synthesis of dolichol-P-oligosaccharides in the endoplasmic reticulum. The deficiency has now been confirmed in the vast majority of CDG1 cases (3).

The search for the CDG1 gene led to the discovery of two genes in the human genome that encode active PMMs. Their identification was based on their sequence similarity to yeast PMM (SEC53) (4). The first human PMM gene, PMM1, is localized on chromosome band 22q13 and could therefore not harbor the primary defect in CDG1 patients (5). More recently, PMM2 was mapped to the CDG1 candidate region on chromosome 16p13 (6). Mutations in PMM2 have been identified in CDG1 patients with a documented PMM deficiency, providing evidence that this gene is the CDG1 gene (6).

The genomic structure of PMM1 and PMM2 has now been determined and compared. The data are presented here. In the course of the experiments, a processed pseudogene, PMM2ψ, was identified. The availability of the flanking sequences has also allowed an extensive analysis of PMM2 in CDG1 patients (G. Matthijs et al., submitted). Interestingly, PMM2ψ carries a number of mutations at positions corresponding to those where mutations were found in PMM2 in patients.

RESULTS

Fine mapping of the CDG1 syndrome and precise localization of the PMM2 gene

The CDG1 gene had been localized to chromosome band 16p13 by linkage analysis (7,8). Critical cross-overs in patients and in carriers have allowed us to reduce the candidate region to a region of <1 cM between markers D16S406 and D16S404 (Fig. 1, and

*To whom correspondence should be addressed. Tel: +32 16 34 60 70; Fax: +32 16 34 59 97; Email: gert.matthijs@med.kuleuven.ac.be
Figure 1. Delineation of the candidate region for CDG1 by genetic analysis of families with a documented PMM deficiency in the proband. The genetic map of the CDG1 region is shown, and the polymorphic markers were ordered according to the available maps (10,11), with the telomere to the left, centromere to the right. The previously published candidate region extends from D16S406 to D16S405 and spans 13 cM (7). The regions represented by thin lines were excluded by cross-overs between the marker alleles and the disease in patients or carriers in the different families (ni, not informative). The minimal region is defined by markers D16S406 and D16S404, that are <1 cM apart on the genetic map.

Figure 2 presents the final mapping data on the PMM2 gene and our contribution to the physical map of chromosome 16 (10). A YAC contig covering this minimal region was constructed, starting from the physical map which had been constructed for chromosome 16 at the Center for Human Genome Studies [Los Alamos National Laboratory, Los Alamos, NM; see (11)]. As shown in Figure 2, the YAC contig spans the region of interest, be it that only a single YAC, 802G4, spans the gap between markers D16S406 and D16S3087, and is positive for marker s54A6. The order of the markers on the Los Alamos map was accepted, but markers D16S513 and D16S406 might be switched. There are conflicting data in the literature about the order of these markers and their position relative to D16S495 and D16S502 [(9), and data from the Whitehead Institute and from Los Alamos].

YAC 802G4 was unstable in our hands, and a new screening of the CEPH MEGA YAC library with marker D16S3020 did not identify any other YAC clones to bridge the region. Therefore, a BAC library was screened with markers s52C5, D16S3087, s54A6, D16S3020, D16S406 and D16S502, and several BACs could be positioned on the map and linked to the YAC contig. Overlaps between different BACs (as indicated in Fig. 2) were identified by hybridization of Southern blots of the BACs with end fragments, generated by vectorette PCR. No BACs were identified for D16S502.

As soon as the PMM2 cDNA was available (6), it was used as a probe for hybridization of Southern blots of the YACs: positive signals were obtained for YAC clones 802G4 and 909F5. These results were confirmed by PCR on YAC DNA with primers specific for exon 4 of PMM2. This result localizes the gene between D16S3020 and D16S502/D16S406. Marker D16S3020 is close to (within 20 kb) and upstream (5′) of the PMM2 gene, because positive signals were obtained by PCR with the BAC clones and several of the cosmids that contain the 5′ end of the gene (see below); the chromosomal orientation of the gene has not been determined. On the high-resolution somatic cell hybrid panel of chromosome 16 (11–13), the gene was mapped to the interval between CY197/CY196 and CY198. The BAC clones isolated with the PMM2 cDNA and localized on 16p by FISH analysis (41M16, 27D21, 342N15) (6), have been integrated into the map (Fig. 2).
PMM1, PMM2 and a processed pseudogene on chromosome 18

PMM1 has been unambiguously localized to 22q13 (5). Hybridizations at low stringency with a PMM1 cDNA probe on a chromosomal mapping panel did not reveal any other signals. PMM2 was identified after a cDNA sequence similar to but different from PMM1 was identified among the IMAGE data in GenBank (6). Several BAC clones were isolated using a PMM2 cDNA probe. However, the results of Southern blot analysis of the BAC clones could only be reconciled if a different chromosomal origin of several of the clones was invoked.

On a genomic Southern blot, the same cDNA probe identified two EcoRI fragments of >12 and 0.8 kb (data not shown). The larger band was also observed with DNA isolated from the cell lines GM10567, which contains chromosome 16 as the sole human chromosome in a mouse background, and GM10487, known to contain chromosome 14 and material derived from chromosome 16p. The 0.8 kb fragment originates from chromosome 18, because it was only observed with DNA from GM11010, a cell line with human chromosome 18 in a hamster background (not shown). FISH analysis revealed that BAC clones 27D21, 41M16 and 342N15, which contain the >12 kb EcoRI fragment, mapped to the candidate region on the short arm of chromosome 16 (16p13), and contained (part of) the PMM2 gene, while BAC clones 147L11, 110G4, 322C1 and 311N20, containing the 0.8 kb EcoRI fragment (Fig. 3A), were derived from the short arm of chromosome 18 (18p) (Fig. 3B).

Cloning and sequencing of the 0.8 kb EcoRI fragment revealed a sequence closely related to the PMM2 cDNA sequence, indicating that it represented a processed pseudogene on chromosome 18. Additional sequence of the processed pseudogene was obtained by cycle sequencing on BAC DNA with specific primers (Fig. 4). The overall sequence similarity between the coding region of PMM2 and the corresponding regions of the intronless, processed pseudogene (PMM2ψ) is 88%. When compared with PMM2, several base substitutions and single base insertions or deletions are present, suggesting that this processed pseudogene has been inactivated by mutations (Fig. 4). The open reading frame is disrupted by an insertion at nucleotide 837 in PMM2ψ which results in a frameshift and leads to a stop at which would be codon 143 of the pseudogene. The region upstream of PMM2ψ has no apparent characteristics of a promoter region, but contains two short repeats and an Alu sequence (not shown). We did not find repetitive elements typically flanking retrotransposed genes (14). Downstream of the position corresponding to the stop codon in the PMM2 transcript, the sequence homology is weak, which suggests that only a partial cDNA has been transposed.

Genomic structure of PMM1 and PMM2

To determine the genomic structure of PMM1 and PMM2, cosmids were isolated by hybridization of arrayed chromosome-specific cosmids libraries from chromosome 22 (15) and chromosome 16 (16) with the respective full-length cDNA probes.

Southern blot analysis of EcoRI, HindIII and double digests of cosmid DNA, hybridized with several probes derived from the cDNAs or from genomic PCR products, and with exon-specific oligonucleotides, and partial HindIII digestions (in the case of the chromosome 16 cosmids) were combined to determine the
Figure 4. Alignment of the PMM2 cDNA sequence with the genomic sequence of the processed pseudogene PMM2ψ on chromosome 18. The translation of the reading frame for PMM2 is given above the DNA sequence (single letter codes). The PMM2 sequence is numbered from the ATG codon, the sequence of the pseudogene starts at the upstream EcoRI site, which was used for cloning (not shown). Due to insertions and deletions in the PMM2ψ sequence, the reading frame is interrupted or altered several times. The open reading frame of the pseudogene ends at the position corresponding to codon 143 in PMM2. The homology was calculated from the alignment shown in the figure. When the sequence difference would lead to a different codon in the putative reading frame of the pseudogene, the corresponding amino acid letter code is represented in bold. The codons in PMM2 that were found mutated in CDG1 patients, and two polymorphisms [(6), and G. Matthijs et al., submitted] are boxed. An asterisk denotes that the same nucleotide change is seen in the patients and in PMM2ψ, and an arrowhead indicates the nucleotide of interest. An EcoRI site, present in the coding region of PMM2 (exon 5) and in the processed pseudogene, is underlined.
Figure 5. Comparison of the genomic structure of the PMM1 (top) and PMM2 genes (bottom). A restriction map for EcoRI (E) and HindIII (H) has been established for both genes. Both genes are composed of eight exons, represented by black boxes. The different cosmid and BAC clones that were used to determine the genomic structure are schematically represented.

Figure 6. Alignment of the exon/intron boundaries for PMM1 and PMM2. The coding sequence is represented in uppercase. Values between square brackets represent additional nucleotides present in one gene as compared with the other.

The genomic structure of both genes. Figure 5 shows the genomic structure of PMM1 and PMM2 with a detailed EcoRI and HindIII map. The position of the different cosmids is indicated. For PMM2, the overlapping BAC clones have been included.

All exon/intron boundaries and partial intron sequences were obtained by cycle sequencing on cosmid or BAC DNA. Both PMM1 and PMM2 are composed of eight exons, and the major difference resides in exon 8 in which the 3' untranslated region is contained (540 bp for PMM1 and 1599 bp for PMM2). The intron/exon junctions conform with the eukaryotic consensus sequences for splice donors and acceptors (Fig. 6).

The intron lengths have been determined by PCR. They vary from 160 bp (intron 3) to 4.9 kb (intron 5) in PMM1 and from 0.5 kb (intron 5) to >4 kb (intron 7) in PMM2. The exact size of the last intron in PMM2 has not been precisely determined: none of the available cosmids contained exon 7 plus exon 8, and (long) PCR on genomic or BAC DNA with primers in exon 7 and 8 has not been successful.

The PMM1 gene thus spans ~13 kb of genomic DNA, whereas the size of PMM2 is at least 17 kb.

**DISCUSSION**

We describe a novel family of genes in the human genome, encoding PMMs and represented by PMM1 on 22q13 and PMM2 on 16q13. A processed pseudogene is present on chromosome 18p. PMM2 is the CDG1 gene, as mutations in this gene have been identified in patients with the syndrome (6). There is currently no disease associated with defects in PMM1.

A YAC contig and partial BAC contig were constructed across the CDG1 minimal region, which had been reduced to an interval on the genetic map of <1 cm between markers D16S406 and D16S404. The positional cloning approach and the physical mapping effort were suspended due to the availability of a candidate gene, i.e. a cDNA for PMM2 (6). PMM2 has now been precisely mapped to the candidate region for CDG1, within 20 kb of D16S3020. PMM1 has previously been isolated by a similar approach and mapped to bin16.2 on chromosome 22 (5).

In functional assays, both proteins display PMM activity, be it that the purified enzymes have different kinetics and a distinct substrate specificity ([17], and M. Pirard and E. Van Schaftingen, unpublished data). The differences indicate that PMM1 and PMM2 may have distinct functions. At the amino acid level, the
proteins are strongly conserved. An alignment was published in Matthijs et al. (6). PMM2 is shorter than PMM1 and yeast and Candida albicans PMM, and the difference is mainly due to a deletion of 7 amino acids in the N-terminal part of the protein. An insertion of 2 amino acids, after position 63 and between conserved domains, is unique to PMM1. These variations do not coincide with exon boundaries. At present, nothing is known about functional domains in the proteins. For comparison, five monomeric phosphoglucomutase (PGM) isoforms are known, all with a different substrate specificity and thermostability (18,19). However, the sequence similarity between the PGM proteins is low.

Comparison of the genomic structure of PMM1 and PMM2 indicates that these genes have arisen by gene duplication. There are 78 differences at the amino acid level between PMM1 and PMM2, not including the four mutational events that have led to differences in length of the proteins. At the DNA level, the silent sites vary in >50% of the positions, suggesting that the ancestral gene was duplicated before the mammalian radiation that occurred 75–110 million years ago. Indeed, we have cloned Pmm1 and Pmm2, the orthologs of PMM1 and PMM2 in the mouse, and these genes are located on syntenic chromosomal regions in the human and the mouse genome (unpublished data).

The identification of PMM2 on 16p13 and PMM1 on 22q13 has prompted us to look for other paralogous genes in these regions. There are now at least three genes on 16p13 that have homologs on 22q12–q13: CREB-binding protein (CBP, OMIM 600140), mutated in the Rubinstein–Taybi syndrome (OMIM 190849) and located on 16p13 (20–22) has a functional homolog, p300, on 22q13 (23), and the HMOX-1 (heme oxygenase-1, OMIM 141250) and HMOX-2 (OMIM 141252) genes are on 22q12 and 16p13.3, respectively (24). It is thus very likely that these chromosomal regions are paralogous regions that have arisen by duplication.

In view of the theory of Ohno, in which it is hypothesized that the genome of higher organisms has arisen by a round or two of tetraploidization (25), one now wonders whether up to four linkage groups or paralogous regions exist for 16p13. However, thus far no other members of the PMM family have been identified in humans nor have other proteins related to the HMOX-1 and -2 genes been described. At least two proteins, related to CBP/p300 are known (26,27) but the corresponding genes have, to our knowledge, not been located in the human genome. On the other hand, the MYH11 and MYH9 genes on 16p13.13 and 22q11–q13 may also be paralogous genes and related MYH genes are clustered on 14q12 and 17p13, whereas members of the somatostatin receptor family have been mapped to 22q13 (SSTR3), 16p13 (SSTR5), 14q13 (SSTR1), 17q24 (SSTR2) and 20q11.2 (SSTR4). Thus, regions on 14q and on 17p–17q are candidate regions in the context of Ohno’s suggestions.

The gene on chromosome 18 is a processed pseudogene of PMM2. It is closely related to the PMM2 cDNA sequence and the absence of homology in the 5’ upstream region suggests that it has arisen by retrotransposition. The 3’ tail of the processed pseudogene has not been fully sequenced, so it is not known whether the poly(A) stretch, characteristic for processed pseudogenes, is present. The presence of a stop codon at position 143 implies that this gene cannot be translated into a functional enzyme. Most likely, this processed pseudogene has never been actively transcribed. The fact that the pseudogene has accumulated mutations at silent and replacement sites at the same rate fits with this assumption: there is a 10.5% sequence divergence at all sites and a 10.8% divergence at silent sites. If a unit evolutionary period (UEP = time needed for a 1% divergence) of 2.1 million years is accepted for synonymous or silent sites, one can estimate that this processed pseudogene arose ~23 million years ago (28). It is therefore unique to humans (and possibly primates).

Of the 22 missense mutations and two polymorphisms thus far identified in CDG1 patients (6, G. Matthijs et al., submitted), seven were also present at the corresponding positions in the processed pseudogene. This is an important observation, because these mutations might interfere with certain mutation detection strategies, e.g. dot-blot analysis, and probes need to be designed carefully. In the case of the frequent R141H (CGC→CAG) mutation, and of the A108V (GCC→GTC), R123G (CGA→GAA) and T237M (ACG→ATG) mutations, the corresponding base is unchanged, but a mutation occurred in the same CpG dinucleotide, probably caused by a G→T transition in the opposite strand. Since the mutations in the PMM2 gene are single base pair changes, without variations in the neighboring sequences as in the pseudogene, they cannot be explained by gene conversion. It rather indicates that the inactive pseudogene accumulated the same mutations that arose independently in the active PMM2 gene where they cause disease. The C→T and G→A transitions prevail, which is in accordance with the propensity of 5-methylcytosine to undergo deamination to form thymine (29). In the case of the frequent R141H mutation, we infer from this comparison that all carrier chromosomes must originate from a single mutation event, because the equally probable R141C mutation seen in the pseudogene has not been observed among CDG1 patients. Similar inferences can be made for other mutations.

In conclusion, the precise mapping of the PMM1 and PMM2 genes allows these genes to be integrated in the physical maps of the respective chromosomes. Their gene structure has been determined, and strongly supports their origin by duplication. The latter suggestion sheds light on a common origin of the chromosomal regions of 16p13 and 22q13 in humans. Also, the identification of a processed pseudogene is interesting from the point of view of molecular evolution, in that it seems to have acted as a sink for mutations since its creation by retrotransposition. Mutations at the corresponding nucleotides in PMM2pseudogene in the human genome and in PMM2 in patients may inadvertently interfere with molecular diagnosis of CDG1.

MATERIALS AND METHODS

YAC analysis

YAC clones were obtained from ‘mega’ YAC libraries constructed at CEPH. The selected YAC clones were grown on uraltrp plates and individual colonies were picked, grown in selective AHC medium, and analyzed by pulsed-field gel electrophoresis (PFGE) to check the size of the insert. Yeast DNA was prepared in agarose plugs (200 µl) according to Ragoussis (30). The STS content of the YACs was checked by PCR. For PCR analysis, a quarter of a plug was dissolved in 1 ml H2O and 5 µl of this solution was used in 25 µl reaction volume, under standard PCR conditions.

A PFGE gel with YACs 820D10, 948AR, 931F5, 912C5, 905F10, 909F5, 802G4, 944F4, 951F4, 925B10 and 936B6 was blotted and the filter was hybridized with the PMM2 cDNA probe under standard conditions (31). The cDNA probe contained the
entire coding, and part of the 3’ untranslated region (to nucleotide 1077).

Isolation of cosmid and BAC clones

The cosmid clones for PMM1 were isolated and kindly provided by M.L. Budarf (5). Cosmids for PMM2 were identified by hybridization of a chromosome 16-specific arrayed cosmid library (16). BAC clones for markers s52C5, D16S3087 and D16S4046 and PMM2 were obtained by hybridization of ‘high density human BAC colony DNA membranes’ (Research Genetics). The probe for s52C5 was generated by PCR from the ampiclon of s52C5 and random primed labeling. For the polymorphic markers D16S4046 and D16S3087, the specific PCR primers were used as probes for oligonucleotide hybridization. The primer sequences were obtained from GDB. For PMM2, the insert of the PMM2 cDNA clone was purified and labeled. The BAC clones for marker s54A6 were isolated by PCR analysis of BAC DNA pools (Research Genetics).

BAC ends were rescued by vectorette PCR. Vectorette libraries were constructed as described by Riley et al. (32) with minor modifications. In brief, 200 ng DNA was digested with 5 U Rsal and ligated to 6 pmol Rsal vectorette cassette (top strand: 5’-CAAGGAGAGAGACGCTTGGTCAAGG-3’; bottom strand: 5’-CTCTCCTCTTGGAATCCGTCGTACGAGAA TC -3’). The end fragments were then amplified by PCR with the universal vectorette primer (5’-CGAATTCGTAACCGTTCGTACGAGAA TCGCT-3’) and the SP6 or T7 primer. The purified PCR product was used as a probe in Southern hybridization.

Analysis of BACs and cosmid clones

BAC clones for PMM2 and cosmids for PMM1 and PMM2, were analyzed by hybridization of Southern blots of EcoRI and HindIII digestions. cDNA or genomic probes were labeled with [α-32P]dCTP by random primed labeling and used for hybridization overnight in 50% formamide, 5× SSPE, 10× Denhardt’s solution, 2% SDS and 100 μg/ml heparin. Filters were washed for 30 min in 0.1× SSPE and 0.1% SDS at 62°C. Oligonucleotide probes, derived from the cDNA sequence and used for the determination of the genomic structure, were typically 20 bp long and 5’-labeled using [γ-32P]ATP and T4 PNK according to established protocols (31). Hybridizations with oligonucleotide probes were done in 6× SSPE, 5× Denhardt’s solution, 0.5% SDS and 200 μg/ml heparin. Filters were washed in 2× SSPE, 0.1% SDS at 45°C for 5 min.

A detailed HindIII map of cosmids 428D1, 408C7, 422F4 and 404H6 was generated by partial digestion. Four μg of purified cosmid DNA was digested to completion with NotI (20 U) in 20 μl, and further digested with HindIII (5 U) in 100 μl. Aliquots of 25 μl were taken at 2, 5, 10 and 60 min. The reaction was stopped by adding 1 μl 0.1 M EDTA. Samples were analyzed by field-inversion gel electrophoresis (FIGE), blotted and hybridized with T3 and T7 oligonucleotides.

Gene-specific primers have been used for genomic PCR on cosmids, BACs and total human DNA under standard conditions. The Long Expand PCR kit was used as prescribed by the manufacturer (Boehringer Mannheim).

Fluorescently labeled primers (fluorescein-isothiocyanate, FITC) were used for cycle sequencing of cosmid DNA using Amersham’s Thermo-sequenase kit and 2–6 μg of DNA.

ACKNOWLEDGEMENTS

These investigations have been supported by the Interuniversitary Network for Fundamental Research of the Belgian Federal Service for Scientific, Technical and Cultural Affairs (IUAP) and by the Flanders Foundation for Scientific Research (FWO-Vlaanderen). We are indebted to Denis Le Paslier and Laetitia Gressin for the YAC screenings. We thank Emile Van Schaftingen, Peter Marynen and Rachel Giles for critical comments and helpful suggestions.

REFERENCES


