Molecular and evolutionary investigation of the phosphoglucomutase gene family.

by

Janine Tomkins

July 1996

A thesis submitted for the degree of
Doctor of Philosophy
in the University of London

MRC Human Biochemical Genetics Unit,
Galton Laboratory,
University College London.

ProQuest Number: 10106760

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

ProQuest 10106760

Published by ProQuest LLC(2016). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code.
Microform Edition © ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

This thesis describes molecular and evolutionary investigations of the phosphoglucomutase (PGM) gene family. The PGM loci (*PGM1*, *PGM2* and *PGM3*) widely expressed in man are thought to be the products of a diverged gene family. Following the cloning of *PGM1* in 1992, the primary aim of this project was to investigate approaches for cloning the other members of the gene family.

The strategies investigated include the use of anti-PGM1 antibodies, low stringency PCR, degenerate primer PCR and searching EST databases. A variety of resources were used, including the human cell line K562. This cell line is devoid of PGM1 activity and the deficiency was found to be associated with a marked reduction in PGM1 mRNA, thereby providing a useful resource.

Two novel DNA sequences, *hyhbf* and human ESTI have been partially characterized. *Hyhbf* was identified by degenerate primer PCR of human cDNA. Although it is a member of the PGM gene family, no evidence could be obtained to confirm the sequence was human and it is suspected to be of bacterial origin. The human ESTI sequence, however, represents a widely expressed gene, which shows alternative transcripts and a related sequence. Evidence suggests it is a candidate for *PGM2*.

Evolution of the *PGM1* gene was investigated in mammals. Nucleotide analysis of the great apes showed the PGM1*1+ is ancestral since the ape homologues have the same characteristic amino acid substitutions as man. Extensive phylogenetic analysis of prokaryotic and eukaryotic sequences identified through conserved functional protein domains was undertaken. Eight distinct evolutionary pathways were identified, two of which, represented by *Mycoplasma pirum PMM* and *Saccharomyces cerevisiae AGM* are thought to reflect the divergent evolution of *PGM2* and *PGM3.*

ACKNOWLEDGEMENTS

I would first like to thank David for all his advice, encouragement and support throughout my PhD, both in his role as a supervisor and a friend. I would also like to thank Hoppy for his ideas and comments, particuarly regarding the production of this thesis.

I would like to thank Jenny Parrington and Margaret Fox for carrying out the cytogenetic analysis of K562, and Owen McMillan for his help and advice on the phylogenetic analysis. I thank Jane Sowden for providing the K562 cell line, Ben Carrit for the PGD primers, Kay Taylor for the primate DNA and Michael Miles for the *Typanosome cruzi* DNA.

I am most grateful to Jenny for all the help and support she has provided during the last four years, and also to other members of the Galton, both past and present, who have provided friendship and inspiration, particularly Helen, Alex, Martine, Jane, Alex and Rod. I would also like to thank those in Newcastle who have been supportive over the last few months. I am very grateful.

Finally, I would like to thank Kate, Clare, and my family for their patience and understanding throughout it all.

This work was funded by an MRC Research Studentship.

ABBREVIATIONS

A	adenine
bp	base pairs
C	cytosine
cDNA	complementary DNA
chr	chromosome
cps	counts per second
der	derivative
DNA	deoxyribonucleic acid
dNTPs	2' deoxyribonucleotide triphosphate
EDTA	ethylenediaminetetraacetic acid
G	guanine
GDP	guanosine diphosphate
hnRNP	heterogeneous nuclear ribonucleoprotein particles
IgG	immunoglobulin G
IVS	intervening sequence
kb	kilobase
mRNA	messenger RNA
mw	molecular weight
nt	nucleotide
OD	optical density
pI	isoelectric point
RNase	ribonuclease
RNA	ribonucleic acid
rRNA	ribosomal RNA
T	thymine
Tris	tris(hydroxymethyl)aminomethane
UDP	uridine diphosphate
UV	ultraviolet

CONTENTS

FIGURES

CHAPTER SIX: DATABASE SEARCH FOR MEMBERS OF THE PGM GENE FAMILY 145 - 162

CHAPTER ONE:

INTRODUCTION

The research described in this thesis focuses on molecular investigations of the phosphoglucomutase (PGM) gene family. The three PGM loci widely expressed in humans are thought to be the products of an ancient gene family, evolved from a single gene, which has undergone duplications and translocations. Characterization of members of the gene family at both the DNA and protein level will allow the evolutionary relationship between the three loci to be investigated. Following the cloning of the *PGM1* gene in 1992, the primary aim of the research was to investigate approaches for the cloning of other members of the PGM gene family.

In addition, the evolution of the *PGM1* gene has been investigated by comparative studies of DNA sequence from PGM1 homologues in primates, rabbits and rats. The analysis focuses on exons which are known to contain genetic polymorphisms in the human population, including those which underlie the PGM1 protein polymorphism. Finally, phylogenetic analysis of PGM and PGM-related sequences has been carried out to investigate the evolution of an apparent ancestral gene which has given rise to genes present in both prokaryotes and eukaryotes. The function of these genes is not always conserved, but certain protein motifs characteristic of the ancestral gene are evident.

The next section reviews the current literature on the evolution of proteins generally and on specific topics which have particular relevance to PGM.

1.1 EVOLUTION OF PROTEINS

Many biochemical pathways are conserved between the three major taxonomic kingdoms of archaebacteria, eubacteria and eukaryotes. Enzymes catalyzing identical chemical reactions are identifiable in species from each of these kingdoms. Analysis of these proteins at the amino acid level may show sequence conservation, indicating that they are homologous: that they are derived from a common ancestral gene and are functionally conserved. Other proteins may show conservation of amino acid sequence, but during evolution, may have diverged to perform a different function. In this case, the proteins are termed orthologous: they are derived from a common ancestral gene. Thus all homologues are orthologues. Comparison of amino acid sequences of orthologous proteins from different species allows the construction of phylogenetic trees, indicating the divergent evolution of the protein and provides an indication of the constraints upon the protein if it is to retain its structure or biochemical function. Phylogenetic analysis may also be used to reflect the evolution of the species involved.

1.1.1 GENE FAMILIES

Analysis of protein sequences can identify conserved proteins within species. These proteins are termed paralogous, as they are derived from a single ancestral gene by a duplication event, (rather than the speciation event which gives rise to orthologous proteins). Where paralogous genes are identified in a single genome they are classified as a gene family (Creighton, 1993).

The initial duplication event which gives rise to a gene family may have occurred in a variety of ways: from non-homologous chromosomal breakage and reunion, unequal but homologous crossing over between two repeated sequences either side of the ancestral gene or by RNA mediated transposition. Once duplicated, the genes are liable to diverge through mutations, such as point mutations and small frameshifts. The level of divergence is restricted by recombination between the loci, with both unequal homologous crossing-over giving rise to or loss of further copies, or gene conversions transferring lengths of nucleotides between the loci. Generally, this genetic exchange contributes to the maintenance of homogeneity in members of a multigene family. However, if one of the loci fulfills the cell's requirements, the other(s) can gain a new function or regulation, or be silenced to become a pseudogene, especially if the level of divergence becomes too great for recombination (reviewed by Maeda & Smithies, 1986).

Gene families can be subdivided depending upon the functions of the proteins which they encode. One family may catalyze an identical reaction, but show regulation in site of expression, for example, the carbonic anhydrases. Alternatively, the members of the gene family may show functional divergence, such as the serine protease inhibitors with regard to their substrate specificity. Members of a gene family may even acquire auxillary functions far removed from those of the other members, such as enzymes serving a role as a structural protein in the eye. These examples will be discussed briefly.

The carbonic anhydrase (CA) family contains seven genes which exhibit a characteristic pattern of tissue expression. Whilst CAII is widely expressed, CA I is highly expressed in erythrocytes, CA III is expressed in muscle and liver, CA IV is expressed as a membrane bound form in lung and kidney, CA V in mitochondria, and CA VI in the salivary glands (Lowe et al, 1990). The seventh gene, CA VII, has been identified as a member of the CA gene family on sequence data alone. Localization of some of the genes provides evidence for translocation following gene duplication, with CA I, CA II and CA III found on chromosome 8, whilst CA VI and CA VII are on chromosomes 1 and 16 (Tashian, 1989).

Members of the serine protease inhibitor (serpin) gene family have evolved in parallel with their substrates, the serine proteases. The archetypal member of the gene family is α_1-antitrypsin (AAT), which is an inhibitor of neutrophil elastase. It shows 30% identity at the protein level with antithrombin III (Doolittle, 1985), yet a single amino acid change of Met to Arg at the reactive centre of AAT changes its protease inhibiting activity from elastase to thrombin (Carrell et al, 1989). High conservation is seen between AAT and α_1-antichymotrypsin, not only with respect to amino acid sequence but also in genomic structure, with conservation of the positions of the introns. Interestingly, the conservation of these particular introns is also seen in angiotensinogen, which shows very low amino acid conservation and does not possess protease inhibitor activity, yet is obviously a member of the gene family (Bao et al, 1987).

Ovalbumin has been identified as a member of the serpin gene family, showing conservation of amino acid sequence with AAT, yet its role as the food storage protein of egg white is far removed from that of the other members of the gene family. Other examples of variation in function can be seen within the δ-crystallin gene family. In chickens, δ1 is specialized for lens expression and produces >95% of the lens δ-crystallin (Piatigorsky & Wistow, 1991). However, the tandemly repeated gene δ2 encodes the enzymatically active argininosuccinate lyase, which shows greater expression in non-lens tissue than δ1. Interestingly, in ducks both genes are expressed in the lens (Wistow & Piatigorsky, 1990).

1.1.2 GENE SHARING

In contrast to gene duplications giving rise to novel proteins, the genome also shows adaption of a single protein to two distinct roles. Examples are seen among the enzyme crystallins, with both the ε-crystallin and τ-crystallin identified as lactate dehydrogenase-B and α-enolase respectively, in non-lens tissue (Piatigorsky & Wistow, 1991). Another gene with recognized dual roles is that of glyceraldehyde-3-phosphate dehydrogenase, which functions as a transfer RNA binding protein in the nucleus (Singh & Green, 1993).

1.1.3 CONVERGENT EVOLUTION

In contrast to divergent evolution where proteins have evolved from a single common ancestor and show homology in their sequences, convergent evolution is the independent evolution of the same catalytic function on different structural frameworks. This is exemplified by the superoxide dismutases (SOD). There are two forms, with most eukaryotes expressing both (Smith & Doolittle, 1992). The first is Cu-Zn SOD, which contains one atom of copper and one atom of zinc. The other is Mn/Fe SOD which contains either a manganese or an iron atom. The two forms have distinctive amino acid sequences which give rise to different 3-dimensional structures, and different mechanisms of action. They occupy different compartments of the cell, with the Cu-Zn SOD localized to the cytosol and the Mn SOD in the mitochondria. The Fe SOD is found in bacteria and fungi, and has also been identified in tobacco chloroplasts.

1.2 PHOSPHOGLUCOMUTASE

Phosphoglucomutase (PGM) is a soluble, intracellular enzyme which catalyzes the interconversion of glucose-1-phosphate (Glc-1-P) and glucose-6-phosphate (Glc-6-P). It acts at the threshold of glycolysis, its role pivotal to both the utilization and synthesis of glycogen. PGM is universally expressed in a wide variety of prokaryotes and eukaryotes, from *Esherichia coli* and *Saccharomyces cerevisiae* to flounder and the chloroplasts of peas (Joshi & Handler, 1964, McCoy & Najjar, 1959, Hashimoto & Handler, 1966, Salvucci et al, 1990). In eukaryotes, multiple loci for PGM have been described. These have been best studied in man, where there are three independent loci. Based upon the similarity in molecular weight, isozyme patterns and enzymatic activity, the isozymes are thought to be the products of an ancient gene family.

1.2.1 EARLY STUDIES OF PGM

PGM was first described in 1938 during investigations on the breakdown of glycogen in mammalian tissues and yeast extracts (Cori et al, 1938a; Cori et al, 1938b). Studies showed PGM activity to require magnesium ions (Cori & Cori,

1937) and glucose-1,6-diphosphate as a cofactor (Leloir et al, 1948). Following the discovery that the active enzyme is phosphorylated (Jagannathan & Luck, 1949), a mechanism of action for PGM was proposed (Najjar & Pullman, 1954). The phosphoryl group from the enzyme is transferred to the C-6 of glucose-1-phosphate to form glucose-1,6-phosphate (Glc-1,6-P). The phosphoryl group on C-1 of the intermediate is then transferred to the dephospho-enzyme, resulting in glucose-6-phosphate and a regenerated phosphoenzyme:

(i) Glc-1-P + phospho-enzyme = Glc-1,6-P + dephospho-enzyme

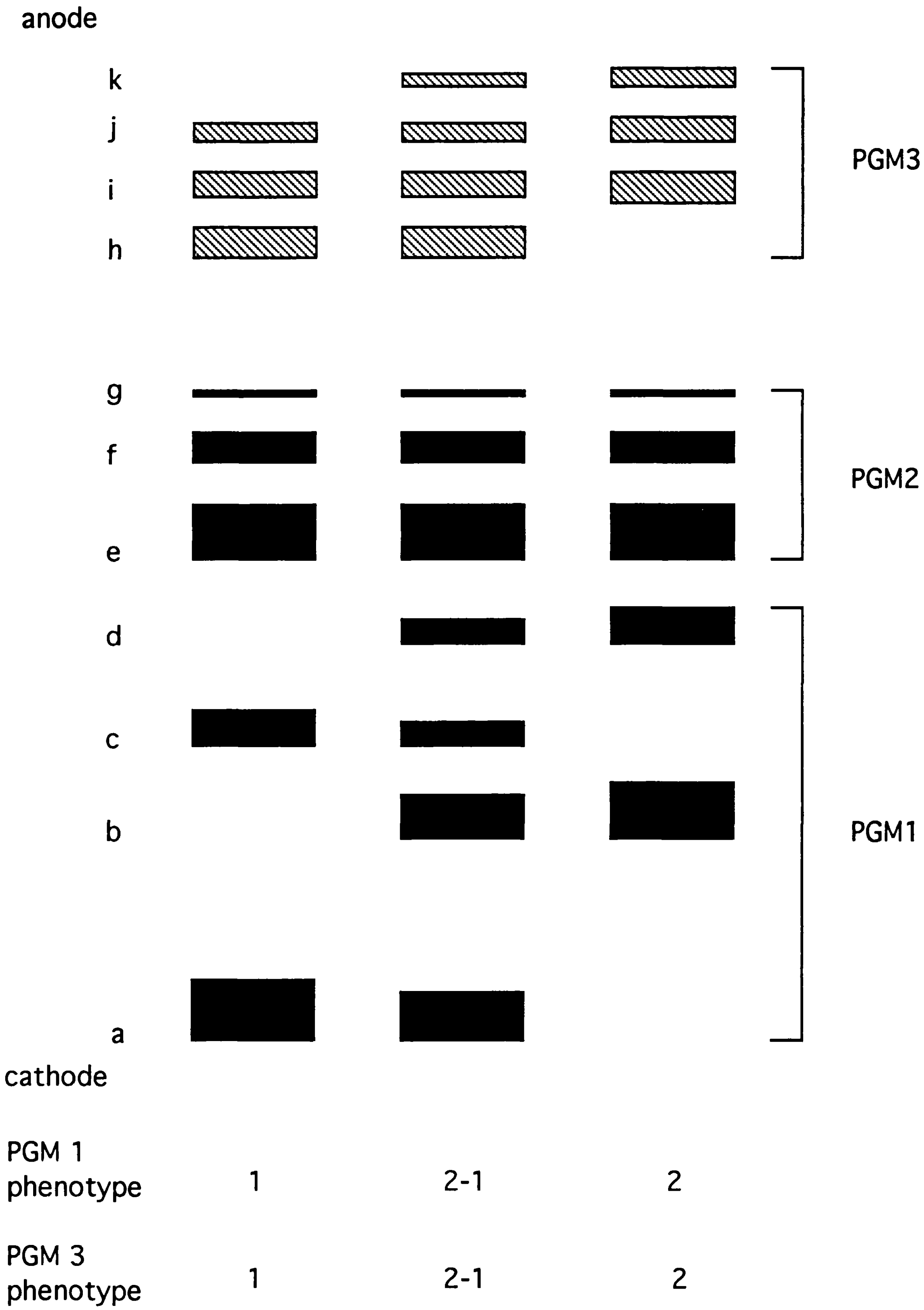
(ii) Glc-1,6-P + dephospho-enzyme = Glc-6-P + phospho-enzyme

In 1957, Anderson and Jolles investigated the linkage of the phosphate group to the protein in PGM (Anderson & Jolles, 1957). Following partial acid hydrolysis, paper chromatography identified the phosphate containing substance as a phosphoserine. The amino acid sequence of the active site surrounding the phosphoserine was first determined by Milstein and Sanger, (1961). The pentapeptide of -Thr-Ala-SerP-His-Asp(Asn)- was identified from crystalline PGM from rabbit skeletal muscle. (The Asp residues could not be distinguished from Asn because of technical difficulties at that time.) Subsequent peptide analysis of rat and yeast PGM indicated that the pentapeptide was common to all three (Milstein, 1961). Conservation of this pentapeptide was seen in flounder, with publication of the sequence **-Thr-Ala-SerP-His-Asp**-Pro-Gly-Gly-Pro-Asp-Asp-Gly-Phe- (Hashimoto et al, 1966). In 1968, the amino acid sequence around the active site was extended to 23 residues (Milstein & Milstein, 1968). The full amino acid sequence of rabbit muscle PGM was published in 1983 and confirmed the work of these early studies (Ray et al, 1983).

1.2.2 PGM LOCI IN MAN

Electrophoretic techniques to separate proteins coupled with enzyme activity detection assays revealed the polymorphic nature of PGM (Spencer et al, 1964). Three distinct and reproducible patterns were seen in red cell extracts, indicating person to person variation. Family studies indicated that the patterns were consistent with the segregation of two alleles (PGM1*1 and PGM1*2) at a single locus: homozygotes for the two alleles showed distinctive two banded patterns, whilst heterozygotes possessed all four bands (Figure 1.1). Three additional bands of activity, located toward the anode of the gel, were invariant. However, further investigations revealed two unusual phenotypes concerning

Figure 1.1 Diagrammatic representation of PGM isozymes separated by starch gel electrophoresis. Primary isozymes for the three loci are a and b for PGM1, e for PGM2 and h and i for PGM3. In addition, faster migrating secondary isozymes are observed for each allele (Fisher & Harris, 1972).



these faster migrating bands (Hopkinson & Harris, 1965). Family studies showed these phenotypes were independent of PGM1, encoded by a second structural locus, designated PGM2.

During investigations of PGM expression in a wide range of tissues, a third structural locus, PGM3, was identified (Hopkinson & Harris, 1968). The PGM3 isozymes only accounted for 1-2% of total PGM activity in placental extracts and were found to be polymorphic, showing three distinct electrophoretic patterns (Figure 1.1). These three phenotypes were determined by two common alleles, PGM3*1 and PGM3*2, and showed no association with PGM1 phenotype.

A fourth locus (PGM4) was reported to be expressed in human milk and appeared to be highly polymorphic (Cantu & Ibarra, 1982). Four alleles were proposed to account for the variation in the milk isozyme patterns which appeared to be independent of the PGM1 and PGM2 phenotypes.

1.2.2.1 Polymorphic and Variant Alleles of PGM

The common PGM1 polymorphic alleles, PGM1*1 and PGM1*2 are found in all populations, but other alleles have been identified. Of these, the PGM1*3 and PGM1*7 alleles reach polymorphic frequencies in the Asian-Pacific area (Blake & Omoto, 1975), the PGM1*3 allele common in New Guinea (10%) and the Western Caroline Islands (11%), and the PGM1*7 allele, also common in the Western Caroline Islands (4-8%), and in the Chinese in Indonesia. The PGM1*7 allele is also found at lower frequency in Japan, China, Thailand and West Malaysia. Other rare PGM1 alleles, many of which are restricted either by geographic, or ethinic distribution, have been identified; by 1985 there were 30 PGM1 alleles described (Dykes et al, 1985).

In contrast to PGM1, PGM2 is monomorphic in most populations where the PGM2*1 allele predominates. However, the PGM2*2 allele in the heterozygous PGM2*2-1 phenotype, originally described in black Africans (Hopkinson & Harris, 1966), reaches polymorphic frequencies of up to 5% in certain sub-Saharan populations (Blake & Omoto, 1975). In addition, other rare PGM2 phenotypes identified by starch gel electrophoresis in the Asian-Pacific area indicate a total of 12 PGM2 alleles. As with the rare PGM1 alleles, the existence of these alleles is restricted to specific populations and regions, where their incidence may reach polymorphic levels.

The polymorphic PGM3 locus possesses two alleles, PGM3*1 and PGM3*2. The PGM3*1 allele is the most frequent in European and other populations with gene frequencies of 73-88% (reviewed by Corbo et al, 1980), whilst the PGM3*2 allele is more frequent in the Nigerian population with a frequency of 66% (Hopkinson & Harris, 1968). It is interesting to note, that only these two common alleles have so far been detected with no rare variants identified. This may be due to the data not being as extensive as for the other two loci. This is primarily because of the difficulty in detecting the PGM3 isozymes, and therefore its use as a genetic marker in population studies is restricted.

Isoelectric focusing (IEF) gels, which separate proteins according to their isoelectric points (pI), allow a higher resolution than starch gel electrophoresis. The use of IEF gels showed that the two alleles PGM1*1 and PGM1*2 can each be subdivided into two, 1+ and 1-, 2+ and 2-, with the '+' being more anodal than the '-' (Bark et al, 1976, Kuhnl et al, 1977). The four common allelomorphs give rise to the ten phenotypes observed on isoelectric focusing gels (Figure 1.2). The PGM1*1 homozygous phenotype seen by starch gel electrophoresis focuses either as a 1+, 1- or a 1+1- heterozygote. The PGM1*2 phenotype can similarly be subdivided into three phenotypes, with the 2 allele focusing at a lower pI. The PGM1*2-1 phenotype subdivides as one of four phenotypes on IEF: 2+1+, 2+1-, 2-1+ or 2-1-.

In 1979, following measurement of the pI of the four alleles, Carter and collegues hypothesized that rather than three independent mutations giving rise to the four alleles, they evolved by two independent nucleotide substitutions, and an intragenic recombination event to form the fourth allele (Carter et al, 1979). As the 1+ is the most frequent allele observed in the human population, and resembled the PGM1 isozymes seen in primates, it was proposed that this was the ancestral allele from which the other three evolved. Two mutations would give rise to the 1- and 2+ alleles, with intragenic recombination between the two mutation sites forming the 2- allele (Figure 1.3). This phylogeny is supported by the additive nature of the pI values and the general distribution of the allele frequencies in most human populations.

Isoelectric focusing of the PGM1*3 and PGM1*7 variants from the Japanese population also showed that they subdivide into '+' and '-' forms (Takahashi et al, 1982). The pI values of the isozymes encoded by these four alleles (3+, 3-, 7+, 7-), as well as those determined by the common four alleles (1+, 1-, 2+, 2-) were measured. The phylogeny of Carter et al was then extended to include all eight alleles. In addition to two mutations and single intragenic recombination

Figure 1.2 Diagram showing how the three PGM1 phenotypes observed on starch gel electrophoresis subdivide into ten phenotypes on isoelectric focusing gels.

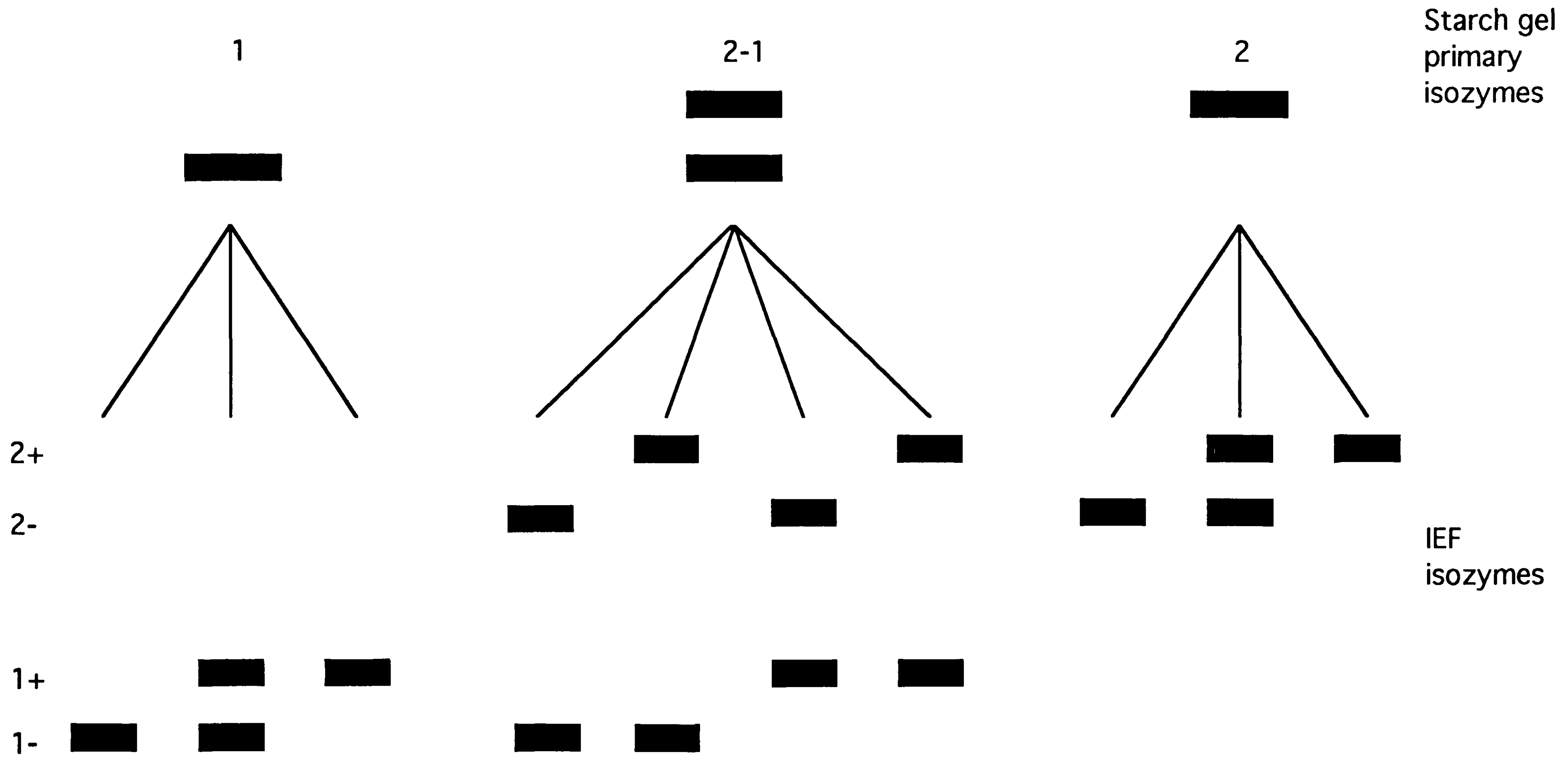


Figure 1.3 Phylogeny of the four PGM1 alleles (Carter et al, 1979). μ indicates a mutation, X an intragenic recombination event.

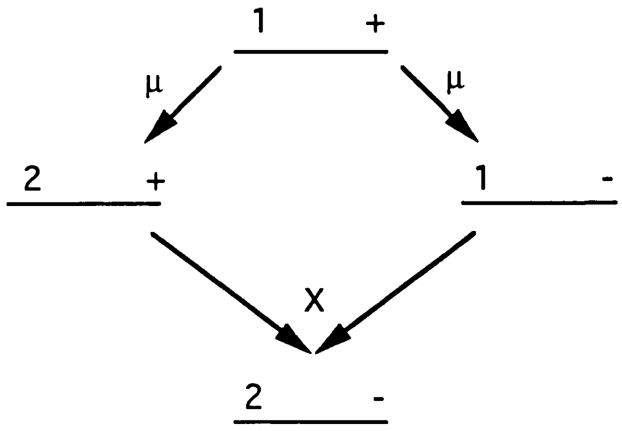
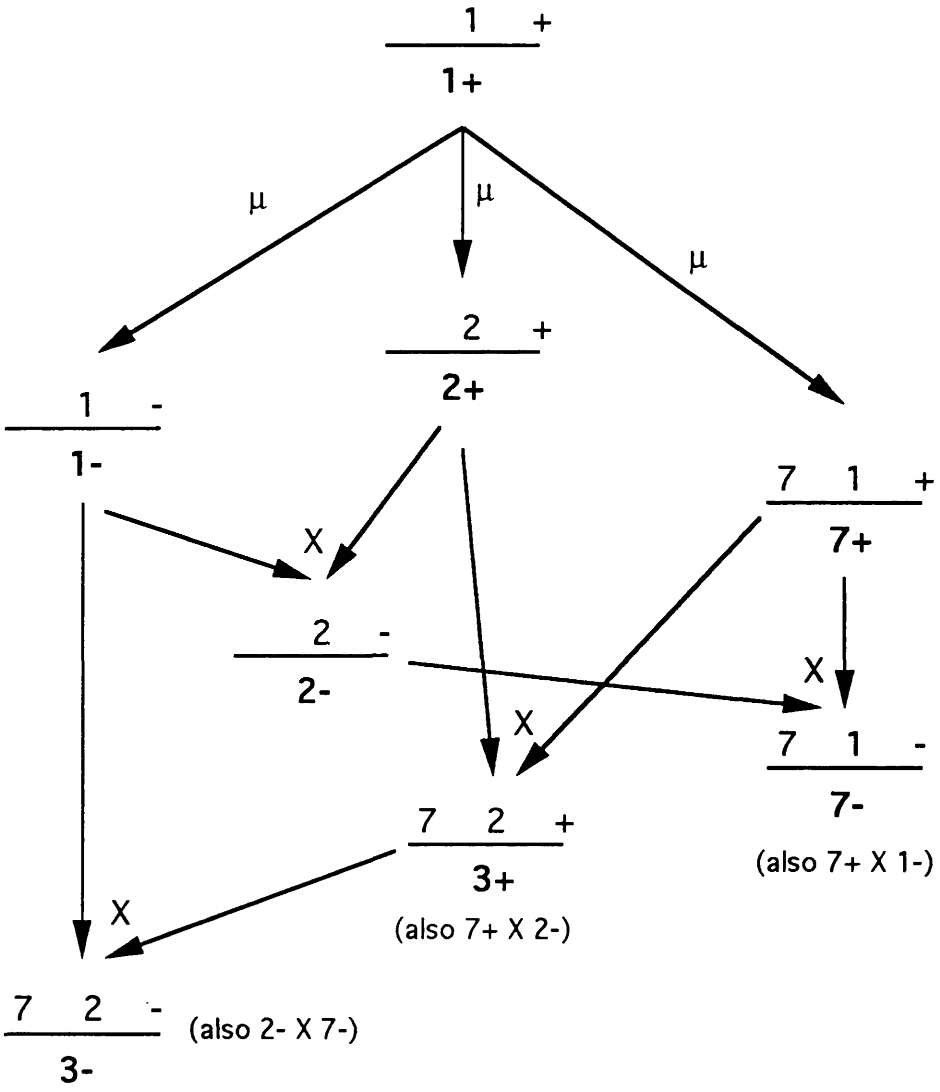


Figure 1.4 Phylogeny of the eight PGM1 alleles (Takahashi et al, 1982). μ indicates a mutation, X an intragenic recombination event.



event already proposed, a further mutation with three intragenic recombination events could give rise to these seven PGM1 alleles from the ancestral 1+ allele (Figure 1.4).

1.2.2.2 Chromosome Localization of the PGM Loci

PGM1 was the first of the three loci to be localized to a human chromosome. In 1972, using human-mouse somatic cell hybrids and chromosome banding techniques, *PGM1* was localized to chromosome 1, along with peptidase-C (*Pep-C*) and, by inference from its syntenic relationship to *PGM1*, 6-phosphogluconate dehydrogenase (*PGD*) (Ruddle et al, 1972). The position of *PGM1* on chromosome 1 was mapped using a somatic cell hybrid clone containing a human chromosome 1 in which most of the long arm had been deleted. Presence of *PGM1* and *PGD* and absence of *Pep-C* activity after electrophoresis, indicated the *Pep-C* locus was sited on the long arm, whilst the *PGM1* and *PGD* loci were either on the short arm or proximal part of the long arm of chromosome 1 (Jongsma et al, 1973). The construction of a genetic linkage map for the short arm of chromosome 1 localized *PGM1* to 1p22 (Dracopoli et al, 1988), although it was stated that a somewhat more distal localization could not be excluded (Bruns & Sherman, 1989). Following the cloning of the *PGM1* gene, the precise location was determined to be 1p31 (Whitehouse et al, 1992).

PGM2 was localized to chromosome 4 in 1975 using human-hamster somatic cell hybrids (McAlpine et al, 1975). The ability of PGM2 to catalyze ribose-1-phosphate distinguished it from the co-migrating hamster PGM. A hybrid containing chromosome 4 with a deletion below 4q26 retained PGM2 expression. Subsequent analysis of somatic cell hybrids containing deletions in chromosome 4 have mapped *PGM2* to lie between 4p14 and 4q12 (McAlpine et al, 1990).

PGM3 was localized to chromosome 6 using human-mouse somatic cell hybrids (Jongsma et al, 1973), and allowed the syntenic markers malate dehydrogenase 1 (*ME1*) and indophenol oxidase-B (*IPO-B*) (van Someren et al, 1974), and multiple histocompatibility complex, locus *HLA-A* (Lamm et al, 1970), to be mapped by inference. Linkage analysis of a family in which there was a cross-over in the HLA region mapped the *PGM3* locus proximal to the HLA loci in 6p or in 6q (Lamm, 1981). Further use of somatic cell hybrid clones including one containing chromosomes resulting from a translocation between 6q12 and 4p13, assigned *PGM3* to 6q12 (Meera Khan et al, 1984).

1.2.2.3 Properties of the PGM loci

The tissue specificity, thermostability, and substrate specificity of each of the PGM loci will briefly be discussed, along with immunological studies to determine the cross reactivity of an anti-rabbit PGM IgG fraction, and molecular weight estimates for the four loci. A summary of these properties is shown in figure 1.5.

Tissue specificity: The three isozymes, PGM1, PGM2 and PGM3, are expressed in most tissues. Generally, 85-95% of the activity is attributable to the PGM1 enzyme, with 2-15% due to PGM2 (McAlpine et al, 1970a). PGM3 contributes only 1-2% of total activity. Exceptions include red blood cells, where PGM 1 and PGM2 enzymes contribute equally to the total activity, PGM3 being undetectable, and cultured fibroblasts where PGM3 isozymes appear to be equal, if not greater, in activity than those of PGM2, contributing >6% of the total PGM activity. PGM4 is the only isozyme which shows tissue specific expression, with activity only detectable in human milk.

Thermostability studies: PGM3 is the least stable, and PGM2 the most stable (McAlpine et al, 1970b). Thus it was hypothesized that the differences observed in PGM activity in red cell and cultured fibroblast extracts may reflect the *in vivo* stability of the proteins, with the anucleate red cells showing loss of PGM3 and reduced PGM1 activity compared with PGM2.

Substrate specificity: Investigations of the PGM isozyme kinetics and the substrate specificity showed that PGM1 was the true phosphoglucomutase, and PGM2 a phosphopentomutase, catalyzing the interconversion of ribose and deoxyribose 1- and 6-phosphates (Quick et al, 1972). The ability of PGM2 to act as a phosphoglucomutase was found to be dependent upon the concentration of the cofactor, Glc-1,6-P. At high concentrations of Glc-1,6-P PGM2 was as efficient as PGM1 in its phosphoglucomutase activity. The PGM2 isozymes, and to a lesser extent, those of PGM1, were also shown to act as phosphomannomutases (Mareneh, 1973). To date, no other substrate for PGM3 has been identified, yet silver staining of 2D electrophoretic gels has shown it to be a fairly abundant protein (Goldman et al, 1985).

Immunological studies: The cross reactivity of a polyclonal anti-rabbit muscle PGM IgG fraction (anti-rabbit PGM) to the members of the human PGM gene family was determined by immunoprecipitation experiments (Drago et al, 1991).

Figure 1.5 Properties of the PGM loci. n/a = data not available.

	PGM1	PGM2	PGM3	PGM4
Tissue Specificity:				
Most Tissues	85-90%	2-15%	1-2%	None
Red Blood Cells	50%	50%	none	None
Fibroblasts	87%	6%	7%	None
Milk	?	None	None	100%
Themostability:	Intermediate	Most Stable	Least Stable	n/a
Substrate Specificity:				
Phosphoglucomutase	Good	Good/Poor	Poor	Good
Phosphomannomutase	Poor	Good	None	n/a
Phosphoribomutase	Poor	Good	None	n/a
Phosphodeoxyribomutase	Poor	Good	None	n/a
Cross reactivity of Anti-rabbit PGM	Strong	None	None	Strong
Molecular Weight				
Gel Filtration	51,000	62,000	53,000	n/a
Ultracentrifugation	62,000	71,000	n/a	n/a
SDS-PAGE	60.000	n/a	n/a	60,000

Following starch gel electrophoresis, the gels were stained for PGM activity. Red cell lysates incubated with the anti-rabbit PGM removed the PGM1 activity, whilst PGM2 was unaffected. In placental extracts, PGM3 was also shown to be unaffected by the presence of the anti-rabbit PGM. The anti-rabbit PGM was shown, however, to cross-react with PGM4 isozymes in milk.

Molecular weight studies: Molecular weight estimates of PGM appear to be dependent upon the technique employed and the tissue under investigation. Ultracentrifugation methods gave estimates of PGM to be 62,000mw in human muscle (Joshi et al, 1967), and 71,000mw in human red cells (Santachiara, 1969). Gel filtration of human red cell PGM provided estimates of PGM1 and PGM2 to be 51,000mw and 58,000mw respectively (Monn, 1969). McAlpine and collegues obtained similar estimates of 51,000mw and 62,000mw using gel filtration for PGM1 and PGM2 respectively, and additionally estimated the molecular weight of PGM3 as 53,000mw (McAlpine et al, 1970c). The commercially available rabbit PGM from Boehringer Mannheim was estimated to be 51,000mw. Thus, although there is agreement in the sizes with respect to each of the isozymes, ultracentrifugation methods gave higher estimates than gel filtration. In 1983, the amino acid sequence of rabbit muscle PGM was published and the molecular weight was determined to be 61,600mw (Ray et al, 1983). Thus, the ultracentrifugation method appears to have given the most accurate estimates. SDS-PAGE of human milk samples gave an estimate of 60,000mw for PGM4 (Drago, 1992).

1.2.3 PGM LOCI IN OTHER SPECIES

In addition to man, multiple isoforms of PGM have been identified in all vertebrates tested, sweet and white potatoes, and yeast (Ray & Peck, 1972, Joshi et al, 1967). This suggests an initial duplication of the ancestral gene is deeply rooted in the evolution of eukaryotes, as only a single PGM locus appears to be present in bacteria, such as *E.coli* (Joshi & Handler, 1964).

In vertebrates, other than man, the most detailed studies have been carried out on the mouse. There are three multiple isoforms observed and their homologies with the three human isozymes have been established: *Pgm2* is the homolog of *PGM1* in humans, *Pgm1* is homologous with *PGM2* and *Pgm3* is homologous with *PGM3.* This was deduced from shared substrate specificities, cofactor requirements and conservation of linkage groups between the species (Quick et al, 1972, Lalley et al, 1978, Nadeau et al, 1981). This suggests that the presence of three isozymes is a feature of mammalian PGM and indicates that the origin of the three loci is rooted in the evolution of vertebrates.

Other species exhibit multiple isozymes. In spinach two isoforms exist, but are distinguishable by their subcellular distribution (Muhlbach & Schnarrenberger, 1978). Isozyme 1 is localized in the chloroplasts where it provides a link between the Calvin cycle and the storage of starch and isozyme 2 is found in the cytoplasm where it performs a role in sucrose metabolism. Both isozymes appear to have a molecular weight of 60,000. Recently the chloroplast PGM was cloned and the coding sequence found to include a putative transit peptide for chloroplast import (Penger et al, 1994). Thus the PGM isozyme 1 has evolved to perform its sole role in the chloroplast.

Two isozymes are also seen in *Saccharomyces cerevisiae*, where PGM not only plays a role in the glycogen metabolism and galactose utilization, but is also on the biosynthetic pathway of the UDP-Glc. This precursor is essential for the synthesis of glucan, one of the cell wall polymers. Presence of two isozymes was first proposed in 1964, when mutants unable to grow on galactose exhibited vastly reduced, but not abolished PGM activity (Tsoi & Douglas, 1964). This was followed with the identification of two activities, PGM1 and PGM2, with *gal5* mutants lacking the major, PGM2, activity. Confirmation of two genes was found in 1969, when two alleles were detected for the minor PGM1 activity (Bevan & Douglas, 1969). No linkage was observed between the two isozymes.

PGM in bacteria was first investigated by Joshi and Handler in an attempt to compare an enzyme responsible for a single reaction in a wide diversity of organisms (Joshi & Handler, 1964). They identified a single protein in *E.coli* with both a similar molecular weight and requirement for magnesium ions and Glc-1,6-P as rabbit and yeast PGM. PGM was subsequently found in *Micrococcus lysodeikticus* and *Bacillus cereus* (Hanabusa et al, 1966).

1.3 MOLECULAR ANALYSIS OF PHOSPHOGLUCOMUTASE

With the advances in molecular biology, the study of PGM moved from the level of the protein, to the DNA. These studies are not confined to man, but include the cloning of other mammalian PGM genes, two PGM genes from *S.cerevisiae*, and many PGM genes from bacteria.

1.3.1 PGM1 IN MAN

In man, PGM1 is the major isozyme of phosphoglucomutase and therefore possesses a greater accessibility for study than the other PGM isozymes. Cloning of human PGM1 in 1992 has led to the further characterization of this gene at the DNA level, with the genomic structure and the molecular basis of the protein polymorphism being elucidated.

1.3.1.1 Cloning of *PGM1*

In 1992, cDNA clones for both the rabbit and human *PGM1* genes were characterized (Whitehouse et al, 1992). The rabbit *PGM1* cDNA was isolated by screening an expression library with a polyclonal antiserum raised against rabbit muscle PGM. The rabbit cDNA inserts were then used to screen a human muscle cDNA library to identify the homologue in man. The cDNA inserts were 2320bp for human *PGM1* and 2279bp for rabbit *PGM1*, with both containing an open reading frame of 1686bp (Figure 1.6), encoding 561 amino acids. The translated cDNA sequence of rabbit was identical to that published by Ray in 1983. Sequence analysis showed a high level of conservation between rabbit and human, with an identity of 88% at the nucleotide level and 97% at the amino acid level. The molecular weight of human PGM1 deduced from the cloned sequence was 61,300.

1.3.1.2 The Genomic Structure of *PGM1* in Man

The genomic structure of human *PGM1* was elucidated in 1993 (Putt et al, 1993). The protein was found to be encoded by 11 exons which covered 65kb (Figure 1.7). Exon sizes varied from 116bp (exon 7) to 659bp (exon 11), and intron sizes from 0.55kb (IVS 2) to 38.5kb (IVS 1). It was noted that two *PGM1* isoforms in rabbit fast muscle differed in sequence at the boundary of exons 1 and 2 (see section 1.3.2.1) (Lee et al, 1992a), thereby suggesting alternative splicing as the mechanism to produce these isoforms. An alternative exon 1 (exon 1B) was identified in humans 6kb upstream of exon 2. It showed conservation at the nucleotide level of 58% with exon 1A (the ubiquitously expressed first exon). The similarity between the two exons at the amino acid level was 74%, suggesting that the two exons arose by duplication (Putt et al, 1993).

Figure 1.6 Diagram of human and rabbit *PGM1* complete cDNAs. (All subsequent numbering of amino acid residues and nucleotides are taken from the reported *PGM1* cDNA sequence in Whitehouse et al, 1992.)

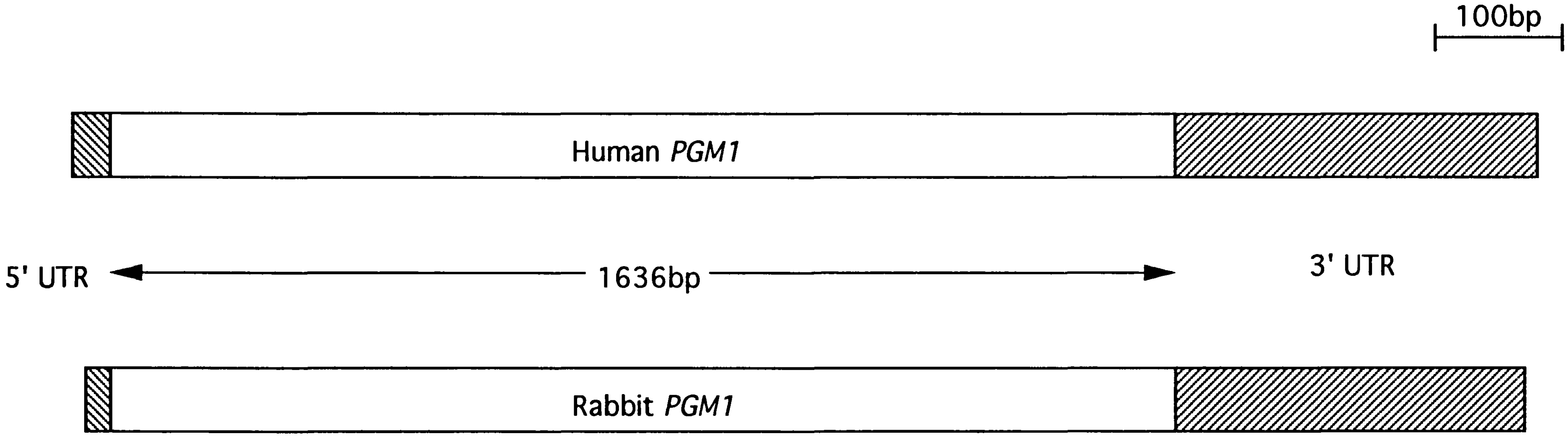
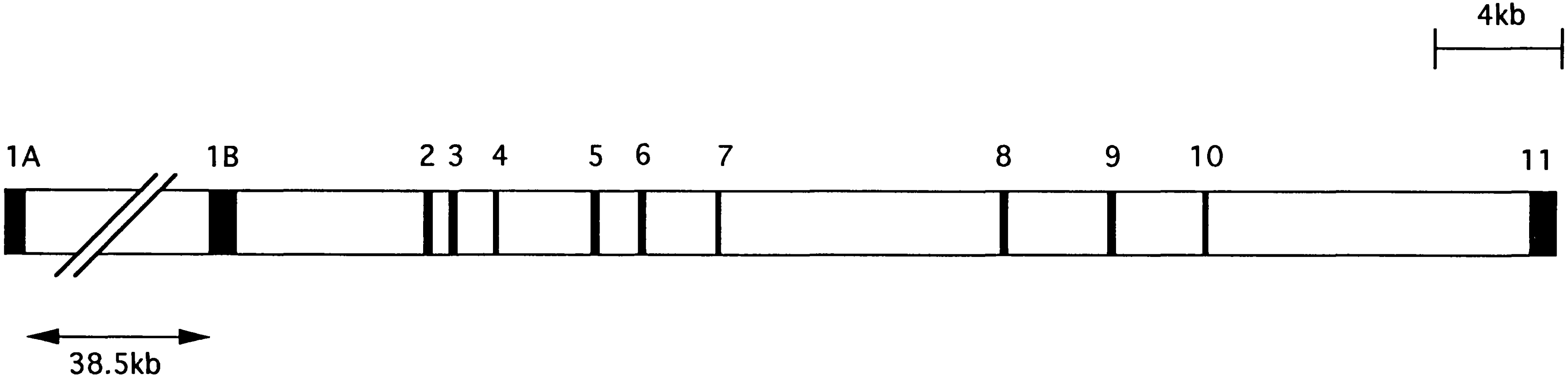


Figure 1.7 Genomic structure of *PGM1*.



1.3.1.3 Molecular Basis of the PGM1 Protein Polymorphism

Cloning of the human *PGM1* cDNA allowed the molecular basis of the PGM1 polymorphism to be defined (March et al, 1993a). A mutation in exon 4, is the basis of the 1/2 polymorphism. Transition of C to T at nt 723 leads to a missense mutation at codon 220 of Arg to Cys: individuals showing the PGM1*1 isozyme carry the Arg codon CGT, whilst those showing the PGM1*2 isozyme carry the Cys codon TGT. A second mutation, located in exon 8, is the basis of the +/- polymorphism. A transition of T to C at nt 1320 leads to a substitution of Tyr to His at codon 419: individuals with the PGM1*+ isozyme carry the Tyr codon TAT, whereas individuals with the PGM1*- carry the His codon CAT. The charge changes predicted by these amino acid substitutions are consistent with the pI values seen on isoelectric focusing and the identification of only two mutations confirmed the idea of Carter et al (1979). Thus, the four alleles arose by two point mutations and an intragenic recombination event occurring in the 18kb separating these two sites.

Investigations in the Japanese population confirmed the mutations which underlie the common polymorphism and also showed the molecular basis of the four less frequent alleles: 3+, 3-, 7+ and 7- (Takahashi & Neel, 1993). In exon 1A, an A to T transversion at nt 265 leads to an amino acid substitution of Lys to Met at residue 67: individuals with the PGM1 isozymes 1+, 1-, 2+ and 2- carry the Lys codon AAG, whereas individuals with the PGM1 isozymes 3+, 3-, 7+ and 7- carry the Met codon ATG. The nucleotides located at each of the mutation sites in each of the eight alleles is illustrated in figure 1.8

Figure 1.8 The polymorphic PGM1 isozymes with the underlying nucleotide substitutions.

Allele	Exon1 nt 265	Exon 4 nt 723	Exon 8 nt 1320
1+	A	C	T
1-	A	C	C
2+	A	T	T
2-	A	T	C
3+	T	C	T
3-	T	C	C
7+	T	T	T
7-	T	T	C

The finding of these sites fulfilled the predictions of the phylogeny put forward over ten years earlier of three mutations and four recombination events generating the PGM protein polymorphisms (Takahashi et al, 1982). A second site of intragenic recombination therefore must exist between exon 1A and exon 4, and it is plausible that the duplication of exon 1 may have also generated this second site for recombination (Putt et al, 1993).

1.3.1.4 The 3' Untranslated Region Polymorphism

In addition to the classical PGM1 polymorphism, a polymorphism has also been described in the 3' untranslated region (3' UTR) of exon 11 (March et al, 1993b). Four alleles were demonstrated by single stranded conformational polymorphism (SSCP) analysis and DNA sequencing identified three mutations: a C to T transition at nt 1773, an A to G transition at nt 1788 and an A to C transversion at nt 1844. Of the eight possible haplotypes, only four were observed. These are shown in figure 1.9.

Figure 1.9 Alleles and haplotypes of the *PGM1**3' UTR polymorphism.

Allele	Allele Frequency	nt 1773	nt 1788	nt 1844	Haplotype
1	82%	C	G	A	+++
2	7%	T	G	C	-+-
3	5%	T	A	C	---
4	6%	C	A	C	+--

The 3' UTR polymorphism was strongly associated with the +/- polymorphism, but not with the 1/2 polymorphism. The *PGM1**3'1 allele was found in high frequency with the + allele, whilst the *PGM1**3'2, 3 and 4 alleles were associated with the - allele.

1.3.1.5 The *TaqI* Polymorphism

Two diallelic restriction fragment length polymorphisms (RFLP) were found in genomic DNA digested with *TaqI* (Hollyoake et al, 1992). At the first site, the two alleles were A1 and A2 with gene frequencies of 18% and 82% respectively, and at the second more 3' site, the two alleles of B1 and B2 showed frequencies of 79% and 21% respectively. The A1 and B2 alleles showed strong allelic association with the *PGM1***- allele and the A2 and B1 with the *PGM1***+ allele.

1.3.2 PGM IN OTHER SPECIES

1.3.2.1 Isoforms of PGM1 in Eukaryotes

In addition to the human and rabbit *PGM1* genes isolated by Whitehouse et al (1992), two isoforms of *PGM1* were identified in rabbit skeletal muscle (Lee et al, 1992a), differing in length and compostion at the amino terminus. The first clone, type 1 *PGM*, encoded a protein of 566 amino acids in which the first 81 amino acids were different from that of the first 77 of the type 2 *PGM*. Analysis of the genomic structure of human *PGM1* determined that these two forms were due to alternatively spliced first exons (Putt et al, 1993). The role of the type 1 PGM was not determined. However, type 2 was identified as PGM1, and appeared to be identical to a Ca^{2+}/calmodulin dependent phosphoprotein found in the sarcoplasmic recticulum that is believed to regulate Ca^{2+} release by its phosphorylation and dephosphorylation.

Purification of a 62,000mw phosphoglycoprotein in rat liver followed by sequential Edman degradation, identified 11 out of 12 amino acids were identical to the N-terminal peptide of rabbit PGM1 (type 2) (Auger et al, 1993). This sequence data was used in combination with data from rabbit PGM1 to obtain a probe which was used to isolate rat *PGM1* from a rat liver cDNA library (Rivera et al, 1993). Partial cDNA sequence has also been obtained from *Pgm2* in the mouse, which is the homologue of human *PGM1* (Friedman, personal communication). The clone contains 780bp corresponding to nt 621 in exon 4 to 1400 in exon 9 of the human *PGM1* gene.

Nucleotide and amino acid sequence comparisons of mammalian PGM1 are tabulated in figure 1.10. As expected, the level of identity at the nucleic acid level is lower than that seen at the amino acid level since most of the nucleotide substitutions are in the third base positions of the codons. The figures for the mouse are based on incomplete sequence data.

The genes for the two PGM isozymes observed in yeast were cloned in 1994, using complementation of *pgm* mutants (Boles et al, 1994). *PGM2* (*gal5*) encoded for the major isozyme activity, and *PGM1* for the minor. The two genes showed a high level of conservation at both the nucleotide and amino acid levels (Figure 1.11). This is higher than the level of identity with human

Figure 1.10 Amino acid and nucleotide identities of mammalian PGM1 sequences.

		Amino Acid Identity (%)			
		Human	Rabbit	Rat	Mouse
Nucleic Acid Identity (%)	Human	-	96.8	96.4	97.7
	Rabbit	88.2	-	97.3	98.5
	Rat	88.3	88.5	-	98.5
	Mouse	89.0	89.1	94.6	-

Figure 1.11 Amino acid and nucleotide identities between human and *S.cerevisiae* PGM sequences.

		Amino Acid Identity (%)		
		Human PGM1	*S.cerevisiae* PGM1	*S.cerevisiae* PGM2
Nucleic Acid Identity (%)	Human PGM1	-	51.5	52.4
	S.cerevisiae PGM1	58.2	-	79.1
	S.cerevisiae PGM2	58.4	72.5	-

PGM1, suggesting that the duplication occurred subsequent to the divergence of yeast and man. This is supported by the observation of conserved sequences downstream of each of the PGM coding sequences. These encode closely related protein kinase genes, *YPK1* and *YPK2*. Thus the duplication event giving rise to the duplication of yeast *PGM* appears to be over an extended region of DNA.

1.3.2.2 PGM in bacteria

In 1994, studies with *pgm* mutants of *E.coli* led to the cloning of a gene for PGM (Lu & Kleckner, 1994). The protein of 58,360mw showed sequence conservation of 48.8% similarity at the amino acid level with human PGM1, and confirmed Joshi and Handler's early work on the presence of a homologue in *E.coli* (Joshi & Handler, 1964).

In addition to having a role at the threshold of glycolysis, the bacterial PGM also plays an essential role in the production of sugars for lipopolysaccharide and exopolysaccharide precursors. Mutants deficient in PGM have been identified in *Agrobacterium tumefaciens* and *Acetobacter xylinum* by their inability to synthesize glucan and succinoglycan in *A.tumefaciens* (Uttaro et al, 1990) and extracellular cellulose in *A.xylinum* (Fjærvik et al, 1991). These genes, *pgm* (*exoC*) in *A.tumefaciens* and *celB* in *A.xylinum* were cloned and in both cases it was shown that the sole catalytic activity of the protein was phosphoglucomutase (Uttaro & Ugalde, 1994, Uttaro & Ugalde, 1995, Brautaset et al, 1994). This however, is not the case in *Xanthamonas campestris* and *Pseudomonas aeroginosa*, where single genes, *xanA* and *algC* respectively, are responsible for both the PGM and phosphomannomutase activities of the bacterium (Ye et al, 1994, Koplin et al, 1992, Zielenski et al, 1991, Harding et al, 1993). Searching of the Genbank and EMBL databases revealed other PMM sequences, similar to *xanA* and *algC.* A summary of these PGM genes is presented in figure 1.12, together with the PMM genes which are discussed in the next section. Further and more detailed analysis of these bacterial genes is presented in Chapter Eight of this thesis.

1.3.3 PHOSPHOMANNOMUTASE IN BACTERIA

Phosphomannomutase (PMM) plays a vital role in bacteria in the production of the mannose residues present in the O-antigen oligosaccharide of the outermembrane lipopolysaccharide . PMM catalyzes mannose-6-phosphate to mannose-1-phosphate, which is catalyzed to GDP-mannose by GDP-mannose

Figure 1.12 Bacterial PGM and PMM sequences. Human *PGM1* is included for comparison

Organism (serotype group)	Gene	Function	Length (No. AA)	AA similarity to HPGM1 (%)	AA identity to HPGM1 (%)	Reference
E.coli	*pgm*	PGM	545	48.8	25.7	Lu & Kleckner, 1994
A.tumefaciens	*exoC/pgm*	PGM	541	73.6	55.8	Uttaro & Ulgade, 1994
A.xylinum	*celB*	PGM	554	48.2	26.3	Brautaset et al, 1994
X.campestris	*xanA*	PGM/PMM	447	45.9	24.7	Koplin et al, 1992
P.aeroginosa	*algC*	PGM/PMM	462	48.7	23.3	Zielenski et al, 1991
S.enterica (B)	*cpsG*	PMM	455	49.5	25.1	Stevenson et al, 1991
S.enterica (B)	*rfbK*	PMM	476	48.9	23.3	Jiang et al, 1991
S.enterica (C)	*rfbK*	PMM	455	49.7	25.5	Lee et al, 1992b
E.coli	*cpsG*	PMM	455	49.9	24.7	Aoyma et al, 1994
E.coli (07)	*rfbK*	PMM	452	49.8	24.5	Marolda & Valvano, 1993
E.coli (09)	*rfbK*	PMM	459	47.1	21.6	Sugiyama et al, 1994
Human	*PGM1*	PGM	561	-	-	Whitehouse et al, 1992

pyrophosphorylase (GDP-MPP). The genes for PMM (*rfbK*) and GDP-MPP (*rfbM*) are adjacent, and together they form the mannose synthesis pathway region of the *rbf* gene cluster (Jayaratne et al, 1994). (Other genes in the *rfb* pathway are concerned with the synthesis of the other repeating hexoses, and transferases.)

In addition to this cluster, PMM may also be found in the *cps* gene cluster. This encodes genes for the biosynthesis of the M-antigen (colanic acid) in the capsular polysaccharide of many enteric bacteria (Stevenson et al, 1991). Although mannose itself is not found in the colanic acid, it does contain fucose, and the precursor GDP-fucose is synthesized from GDP-mannose. To date, most of the work has been carried out on the serotype groups of *Salmonella enterica* and *E.coli* (Figure 1.12).

1.3.4 CONSERVATION OF PROTEIN MOTIFS

In all of the peptide sequences, whether the proteins possess PGM, PMM or both activities, there is evidence of amino acid conservation throughout the sequence. However, regions of higher identity are seen and conservation of these protein motifs throughout the taxa suggest that these are required to enable enzymatic activity as a phosphomutase. The crystal structure of rabbit muscle PGM1 was published in 1992, and the PGM1 protein was shown to consist of four sequence domains, based upon structural and spatial considerations (Dai et al, 1992). Domains I-IV comprised of residues 1-188, 189-301, 302-420 and 421-561. All four domains were found to contribute to the deep cleft forming the active site of the protein, and the conserved protein motifs are shown to be located on loops which are exposed in the active site cleft of the protein.

The crystal structure located the active site peptide of -^{114}Thr-Ala-Ser-His-Asn-Pro^{119}- in domain I, the loop protruding in to the active site cleft. Comparison of this peptide in the cloned PGM and PMM genes indicates complete conservation of four of the residues including Ser^{116} (Figure 1.13).

The phosphoglucomutase reaction requires magnesium ions and the metal binding loop is located at the bottom of the active site crevice, beneath Ser^{116}. The loop which protrudes from domain II consists of the residues -^{287}Asp-Gly-Asp-Gly-Asp-Arg^{292}-. This motif shows complete conservation of all three Asp and the Arg residues. Also there is only a very low level of variation seen in the Gly^{288} (Figure 1.13).

Figure 1.13 Amino acid comparisons of the active site, magnesium binding loop and glucose binding loop (or equivalent) from cloned PGM and PMM genes

Organism	Gene	Active Site	Magnesium Binding Loop	Glucose Binding Loop
Human	*PGM1*	TASHNP	FDGDGDR	GEESFG
Rabbit	*PGM1*	TASHNP	FDGDGDR	GEESFG
Rat	*PGM1*	TASHNP	FDGDGDR	GEESFG
Mouse	*PGM1*	no data	FDGDGDR	GEESFG
S.cerevisiae	*PGM1*	TASHNP	**S**DGDGDR	GEESFG
S.cerevisiae	*PGM2*	TASHNP	**S**DGDGDR	GEESFG
E.coli	*pgm*	T**P**SHNP	**N**DGD**Y**DR	GEES**A**G
A.tumefaciens	*exoC/pgm*	**S**ASHNP	**S**DGDGDR	GEESFG
A.xylinum	*celB*	T**P**SHNP	**N**D**T**D**A**DR	GEES**A**G
X.campestris	*xanA*	TASHNP	**W**DGD**F**DR	GEMSAH
P.aeroginosa	*algC*	T**G**SHNP	FDGDGDR	GEMS**G**H
S.enterica (B)	*cpsG*	TASHNP	FDGD**F**DR	GEMSAH
S.enterica (B)	*rfbK*	T**G**SH**I**P	**T**DGDGDR	GEMSAH
S.enterica (C)	*rfbK*	TASHNP	FDGD**F**DR	?
E.coli	*cpsG*	TASHNP	FDGD**F**DR	GEMSAH
E.coli (07)	*rfbK*	TASHNP	FDGD**F**DR	GEMSAH
E.coli (09)	*rfbK*	TASHNP	FDGD**F**DR	GEMSAH

The loop in domain III is shorter, and almost certainly provides binding specificity by interacting with the glucose ring. This amino acid sequence of -^{377}Glu-Ser-Phe^{379}- is embedded in a motif which is conserved in PGM proteins as -^{375}Gly-Glu-Glu-Ser-Phe-Gly^{380}-. As would be expected, this motif is not evident in the proteins which show phosphomannomutase activity. However, an equivalent sequence has been identified, -Gly-Glu-Met-Ser-Ala-His- (Figure 1.13).

Three loops from domain IV lie on the opposite side of the cleft wall to Ser^{116}. However, none of these appear to form the other expected feature of the active site, that of the distal phosphate binding site. Neither a positive end of a helical dipole, nor a phosphate gripper, with the motif -Gly-X-Gly-X-X-Gly, is present on any of the loops. This sequence is seen in domain IV, but is present on the external surface of the protein. It is conserved in the eukaryotic PGM sequences (except spinach) and also *A.tumefaciens pgm*. An equivalent motif is also seen in the majority of bacterial mutases. However, rather than being at the carboxyl end of the protein it is located between the active site and the magnesium binding loop. Thus this may be an indication, if it is a functional motif, that there are two similar but distinct structural frameworks upon which the conserved amino acid sequence catalyzes the reaction.

1.4 DIVERGENCE OF FUNCTION

In addition to PGM and PMM showing conservation of specific protein motifs, three other proteins have been identified which either contain the active site and magnesium binding loop motifs, or exhibit high amino acid conservation through the cross-reactivity of "specific" monoclonal antibodies to rabbit PGM. However, none of these proteins are phosphoglucomutase or phosphomannomutase. They are N-acetylglucosamine-phosphate mutase from *S.cerevisiae* , parafusin from *Paramecium tetraurelia* and aciculin from man. Each will be discussed briefly below.

1.4.1 N-acetylglucosamine-phosphate mutase in *S.cerevisiae*

During the cloning of PGM1 and PGM2 in *S.cerevisiae* a further gene was identified which, when over-expressed on a multi-copy plasmid, restored the ability of the *pgm1/2* double deletion mutants to grow on galactose (Boles et al, 1994). Partial amino acid sequencing identified a peptide of 10 amino acids which showed high similarity to the active site of PGM and PMM. The complete cDNA (*AGM1*) was identified and found to encode a protein of 557 amino acids (Hofmann et al, 1994).

Deletion mutants of *agm1* progressed through approximately five cell cycles to form a string of undivided, morphologically abnormal cells. A similar phenotype is observed in glucosamine auxotrophic mutants starved of glucosamine. The combination of this observation and the similarity of the amino acid sequence at the active site with PGM, suggested that the function of the protein is as another hexosephosphate mutase, involved in the formation of UDP-N-acetylglucosamine, the precursor in the synthesis of the cell wall polymer chitin. Mutation studies showed that *AGM1* did indeed encode for N-acetylglucosamine-phosphate mutase.

1.4.2 Parafusin in *Paramecium tetraurelia*

Parafusin is a cytosolic phosphoprotein that plays a role in regulated exocytosis in *P.tetraurelia*. The 583 amino acid protein shows 54.6% identity and 71.9% similarity at the amino acid level to human PGM1. The sequence contains four insertions and two deletions, in comparison to PGM1 (Subramanian et al, 1994). One insertion is just downstream of the active site motif, and this disruption may account for the absence of PGM activity in parafusin. The active site and the glucose binding loop motifs found in PGM sequences are identical in parafusin, and a single amino acid change of Gly to Ala occurs in the magnesium binding loop. Southern blotting analysis, enzyme activity assays and Western blotting using parafusin specific antibody indicates that the loci encoding parafusin and PGM are distinct in paramecium (Subramanian et al, 1994, Andersen et al, 1994).

1.4.3 Aciculin in man

The third protein, which shows amino acid sequence similarity to PGM1 in man, yet has a distinct biological function and no PGM activity, is aciculin. This protein was identified by monoclonal antibodies raised against uterine smooth muscle, during studies on the molecular architechture and function of the cytoplasmic domain of adherens junctions (Belkin et al, 1994). Five monoclonal antibodies recognize a doublet of 60/63,000mw, three of which were found to show cross-reactivity with rabbit PGM.

Partial amino acid sequencing of four peptides located throughout the protein, show a high level of amino acid identity to human PGM1. A peptide from near the amino terminus showed 21 out of 22 residues to be identical, two overlapping peptides located in the middle of the protein showed 25 out of 38 residues to be identical and a peptide at the carboxyl end showed identity at 12 out of 20 residues. None of the peptides were located at an identifiable conserved motif, such as the active site.

1.5 CONVERGENT EVOLUTION OF PHOSPHOMANNOMUTASE

The cell wall of yeast consists of three types of structural polysaccharide: glucans, which are polymers of glucose, mannans, heavily glycosylated proteins containing mannose, and chitin, a linear polymer of N-acetylglucosamine. The precursors of these glycans are the nucleotide sugars UDP-Glc, UDP-Man and UDP-GlcNAc, which have additional roles in the glycosylation of other proteins. Each of these nucleotide sugars are synthesized from hexose-6-phosphates, which are converted to hexose-1-phosphates by the action of three different hexosephosphate mutases (Boles et al, 1994). These three enzymes have been identified in *S.cerevisiae*, and both the phosphoglucomutases and N-acetylglucosamine-phosphate mutase have been discussed in the previous section. The third enzyme, phosphomannomutase, is quite distinct from the other two, and provides a good example of the convergent evolution of protein function.

The phosphomannomutase gene (*sec53*) was identified during investigations of the genes involved in the secretory pathway in yeast (Bernstein et al, 1985). The *sec53* mutants showed an early block in the mechanism of protein transport, with the accumulation of secreted precursors in the endoplasmic recticulum (ER). Cloning of the gene and characterization of the gene product identified a hydrophillic, cytoplasmic protein of 29,000mw. This was subsequently identified as phosphomannomutase, through direct assays of activity on multi-copy plasmids carrying the *sec53* gene, and the cofractionation of the *sec53* gene product and phosphomannomutase activity following gel filtration and DEAE chromatography (Kepes & Schekman, 1988).

A homologue of the *sec53* gene has been identified in the pathogenic fungus *Candida albicans* (Smith et al, 1992). The *PMM1* gene shows high similarity to *sec53* at both the nucleotide and amino acid level with sequence similarities of 76.2% and 77.7% respectively. Neither of these genes show any sequence homology to the PGM sequences of yeast. However, the *sec53* gene product has also been shown to restore growth on galactose of a double deletion mutant of *pgm1/pgm2* when expressed on a multicopy plasmid (Boles et al, 1994). Therefore this second structural framework which catalyzes the phosphomannomutase reaction is also able to catalyze the interconversion of Glc-1-P and Glc-6-P. The reciprocal experiment, transformation of a temperature-sensitive *sec53* mutant with multicopy plasmids carrying the *PGM1*, *PGM2* and *AGM1* genes from *S.cerevisiae*, did not complement the mutation.

Therefore, there are four distinct loci, with three specific activities in *S.cerevisiae*, each with individual physiological functions, yet all four, despite the divergence of the AGM1 and the covergent evolution of the PMM, are all able to catalyze the interconversion of Glc-1-P and Glc-6-P.

1.6 SUMMARY OF AIMS

The molecular investigations of the phosphoglucomutase gene family can be divided into two main areas of research. The first describes approaches investigated for the identification of other members of the PGM gene family and related sequences:

Chapter Three details the characterization of a cell line, K562, which is devoid of PGM1 activity. The molecular basis of the deficiency has been carried out at the level of the protein, the gene and the RNA to assess its usefulness as a resource for cloning. In combination with these studies, the use of two anti-human PGM1 polyclonal antibodies as screening tools has also been investigated.

Chapter Four describes two PCR-based approaches using primers designed to conserved regions of the protein to identify of other members of the PGM gene family. The low stringency PCR strategy identifies closely related sequences, whereas the degenerate primer PCR strategy allows for a greater level of divergence between sequences.

Chapter Five details the investigations carried out on the novel PGM-related sequence identified by degenerate primer PCR. The sequence, named *hyhbf* due to the high level of similarity with the *yhbf* gene from *E.coli*, was partially characterized by RT-PCR, genomic DNA PCR, and Southern blot analysis.

Chapter Six describes an alternative strategy of gene identification, by utilizing the expressed sequence tag (EST) databases to search for PGM-related sequences. One of the ESTs identified, human ESTI has been investigated and partially characterized by RT-PCR, genomic DNA PCR, and Southern and Northern blot analysis.

The second area of research looks into the evolution of the PGM gene:

Chapter Seven covers the evolution of the *PGM1* gene in mammals. Comparative analysis of exons 1A, 4, 5, 8 and 11 from primates, rodents and rabbits was carried out to investigate the level of sequence conservation and to determine if the PGM1*1+ like protein in primates has the same characteristic amino acid substitutions as man. This would support the idea of the PGM1*1+ as the ancestral isozyme.

Chapter Eight considers the evolution of an apparent ancestral PGM gene. Phylogenetic analysis was carried out on PGM and PGM-related amino acid sequences from a wide variety of species. The identification of possible alternative pathways in the phylogeny may be suggested to represent the evolution of PGM2 and PGM3.

Finally, Chapter Nine, discusses the broader aspects of the gene identification approaches, the identity of the two novel PGM-related sequences, the conservation of a PGM1 homologue in other species, and the evolution of the phosphohexomutases in man.

CHAPTER TWO:

MATERIALS AND METHODS

2.1 MATERIALS

2.1.1 GENERAL REAGENTS

Unless otherwise stated, standard laboratory chemicals were obtained from BDH, Fisons or Sigma. Enzymes were obtained from Boehringer Mannheim, restriction endonucleases from Gibco-BRL or NEB, MMLV-reverse transcriptase was obtained from Gibco-BRL and Taq polymerase from Promega. Radioactive nucleotides were obtained from Amersham or NEN, Dupont.

2.1.2 CELL CULTURE

The Roswell Park Memorial Institute (RPMI) 1640 media, Dulbecco's minimal essential media (DMEM), and fetal calf serum (FCS) were obtained from Gibco-BRL.

2.1.3 ELECTROPHORESIS MATERIALS

For the protein studies, starch was obtained from Sigma, and the acrylamide:bis-acrylamide from Biorad. For DNA sequencing, a "ready to use sequencing gel solution" was obtained from Severn Biotech. NNN'N'-tetramethylethylenediamine (TEMED) was obtained from BDH and ammonium persulfate (AMPS) from Biorad. Standard agarose of type I, low EEO, was obtained from Sigma, and NuSieve GTG agarose from FMC Bioproducts, Flowgen.

2.1.4 COMMONLY USED SOLUTIONS

Solutions commonly used for protein work were:

TEMM	0.1M Tris; 0.1M maleic anhydride; 10mM $MgCl_2$; 10mM EDTA
TGM	25mM Tris-HCl, pH 8.3; 0.192M glycine; 20% methanol
TGS	25mM Tris-HCl, pH 8.3; 0.192M glycine; 0.1% SDS
1X PBS	0.137M NaCl; 2mM KCl; 8mM Na_2HPO_4; 1.5mM KH_2PO_4

Solutions commonly used for DNA work were:

1X TBE	86mM Tris; 1.9mM EDTA; 90mM Boric Acid; pH 8.3
1X TAE	40mM Tris; 20mM NaOAc; 1mM EDTA; pH 8.0
1X SSC	0.15mM NaCl; 1.5mM tri-sodium citrate; pH 7.0
10mM TE	10mM Tris-HCl; 1mM EDTA; pH 8.0
Phenol	Phenol equilibrated to pH 7.5 with 10mM TE
Chloroform	Chloroform:isoamylalcohol (IAA) mixed 24:1
Phenol/ Chloroform	Equal volumes of phenol and chloroform:IAA
100X Denhardts	2% bovine serum albumin; 2% Ficoll 400; 2% polyvinylpyrollidone

All solutions were made using water purified by reverse osmosis (MilliRO). Solutions used in the preparation of DNA and RNA, and water used for PCR experiments, were autoclaved prior to use.

2.1.5 CELL CULTURE MEDIA

DMEM media: 1x DMEM, diluted using sterile deionized water supplemented with 10% fetal calf serum, 2mM glutamate, 60μg/ml streptomycin, 100μg/ml penicillin (final concentrations).
RPMI media: 1x RPMI, diluted using sterile deionized water supplemented with 10% fetal calf serum, 2mM glutamate, 60μg/ml streptomycin, 100μg/ml penicillin (final concentrations).

2.1.6 MICROBIOLOGICAL MEDIA

L-broth:	1% tryptone (Difco), 0.5% yeast extract (Difco), 0.5% NaCl.
L-agar:	1% tryptone, 0.5% yeast extract, 0.5% NaCl, 1.5% agar (Difco).

2.1.7 PGM SAMPLES

The erythroleukaemic cell line, K562, a gift from Dr. J. Sowden, had been cultured intermittently over the past few years. The lymphoblastoid cell lines 6997, 7014 and 7057 are CEPH cell lines obtained from Nigel Spurr at the ICRF.

Full term placenta had been collected during a previous study and tissue samples had been stored at -70°C.

A panel of 46 blood samples from unrelated individuals were available, along with corresponding DNA samples, extracted on an Applied Biosystems nucleic acid extractor (model 340A).

Primate blood, white cells and cell lines were available in the laboratory. Gorilla Sampson, chimpanzee Masikini and orangutan Henry DNA samples were donated by Dr K. Taylor.

2.1.8 BACTERIAL STRAINS

E.coli INVαF':	endA1, recA1, hsdR17, supE44, λ-, thi-1, gyrA, relA1, ø80lacZΔM15Δ(lacZYA-argF), deoR, F'

2.2 METHODS

2.2.1 CELL CULTURE

The lymphoblastoid cell lines were cultured in DMEM media and K562 in RPMI media, and grown in a moist 5% CO_2 atmosphere.

The cells were harvested by centrifugation and washed in 0.9% saline. They were either used immediately or flash frozen in liquid nitrogen, prior to storage at -70°C.

2.2.2 PREPARATION OF PLACENTAL AND CELL EXTRACTS

2mg of placenta was homogenized with an equal weight/volume of distilled water on a Silveson homogenizer. Following centrifugation at 9789g for 10mins at 10°C, the supernatant was stored at -70°C. Cells were resuspended in an equal weight/volume of distilled water and sonicated by three cycles of 5sec on / 5sec off on an MSE Soniprep 150. Following centrifugation at 12,000g the supernatant was stored at -70°C.

2.2.3 PROTEIN ELECTROPHORESIS TECHNIQUES

2.2.3.1 Starch Gel Electrophoresis of PGM

11% starch gels (300 x 160 x 5mm) were made from a 1 in 10 dilution of TEMM bridge buffer. They were prepared by heating the mixture with continuous stirring. Once the mixture thickened, it was heated for a further minute and then degassed using a vacuum pump until the formation of bubbles had decreased. The gel was poured into the mould, a sheet of Melanex film was lain over and it was allowed to set.

50μl of sample was applied to pieces of No.17 Whatman paper (7 x 5mm) which were inserted vertically into the gel, 6cm from the cathode. The samples were electrophoresed at 5V/cm for 17hrs at 4°C. Starch gels were sliced in half in preparation for either PGM activity staining or electroblotting onto nitrocellulose.

2.2.3.2 Starch Gel Electrophoresis of PGD

12% starch gels were made with a 1 in 10 dilution of 0.1M phosphate buffer pH 7.0 (61ml 0.2M Na_2HPO_4, 39ml 0.2M NaH_2PO_4). Prior to degassing, nicotinamide adenine dinucleotide phosphate (NADP) (1mg/50ml) was added to the gel. 10mg NADP was also added to the bridge buffer reservoir at the cathode. Sample application papers were loaded at the cathode. Electrophoresis was at 3V/cm for 17hrs.

2.2.3.3 Isoelectric Focusing

Isoelectric focusing was performed in 5% polyacrylamide gels 240 x 100 x 0.4mm. The gel mixture consisted of 2ml 87% glycerol, 1.5ml 50% acrylamide/bis-acrylamide 29:1 (Bio-Rad) and 0.9ml Ampholines pH 5-7 (Pharmacia). The volume was made upto 15ml and set with 6μl TEMED and 170μl 3% AMPS. Contact with the electrodes was achieved with strips of No.17 Whatman paper, 10 x 230mm, soaked in 1M H_3PO_4 for the anode and 1M NaOH for the cathode. Prefocusing was for 15mins at 1600V, 25mA and 10W. 5μl of sample was applied at the anode end of the gel on pieces of No.3 Whatman paper 5 x 5mm. Focusing continued at the previous settings. After 1hr, the application papers were removed and focusing continued upto 4800Vhrs.

2.2.3.4 Sodium-Dodecyl-Sulphate Polyacrylamide Gel Electrophoresis

Sodium-dodecyl-sulphate polyacrylamide gel electrophoresis (SDS PAGE) was conducted on 5-15% polyacrylamide gradient gels according to the method of Karlsson et al, (1983), in conjunction with the discontinuous buffer system devised by Laemmli (1970). The gel dimensions were 170 x 150 x 1mm. The 5% solution consisted of 10.6ml dH_20, 2.8ml 30% acrylamide:bis 37.5:1 (Biorad), 3.3ml Tris-HCl pH 8.8, 200µl 10% SDS, 15µl TEMED, and 50µl 10% AMPS, and the 15% solution of 1.8ml glycerol, 3ml dH_20, 8.4ml 30% acrylamide:bis 37.5:1, 3.3ml Tris-HCl pH 8.8, 200µl 10% SDS, 15µl TEMED, and 25µl 10% AMPS. The stacker gel solution was 15ml dH_20, 2.5ml 30% acrylamide:bis 37.5:1, 2.5ml Tris-HCl pH 6.8, 200µl 10% SDS, 20µl TEMED, and 100µl 10% AMPS.

10µl of sample was mixed with an equal volume of SDS-PAGE loading buffer (2% SDS, 10% glycerol, 5% β-mercaptoethanol and a trace amount of bromophenol blue in 0.064M Tris-HCl pH 6.8) and placed in a boiling waterbath for 3mins. 10µl of rainbow coloured protein molecular weight markers (Amersham Life Science) were added to an equal volume of loading buffer, and boiled for 1min. Electrophoresis was conducted in TGS buffer for 4hrs at 180V, 40mA or 17hrs at 40V, 20mA.

2.2.4 PROTEIN DETECTION METHODS

2.2.4.1 PGM Activity Stain

The agar overlay to detect PGM activity following starch gel and IEF electrophoresis, based on the method of Harris & Hopkinson (1976), was as follows; 20ml 0.5M Tris-HCl pH 8.0, 2ml 0.2M $MgCl_2$, 100mg glucose-1-phosphate (Glc-1-P), 5.6U glucose-6-phosphate dehydrogenase (G6PD) in a 1 in 5 dilution with saturated ammonium sulphate, 10mg NADP, 10mg 3-[4,5-dimethylthiaxol-2-yl]-2,5 diphenyltetrazolium bromide (MTT), 5mg Phenazine methosulphate (PMS) and 20ml 2% agar noble (Difco). The stain was developed at 22°C in the dark. Detection of the PGM3 isozymes required 200mg Glc-1-P, 11.2U G6PD and incubation at 37°C to develop the stain.

2.2.4.2 PGD Activity Stain

The agar overlay to detect PGD activity following starch gel electrophoresis was as follows; 10ml 0.5M Tris-HCl pH 8.0 5ml 0.2M $MgCl_2$, 10mg phosphogluconate, 5mg NADP, 5mg MTT, 5mg PMS and 20ml 2% agar noble. The stain was developed at 22°C in the dark.

2.2.4.3 Immunoblot Detection

2.2.4.3.1 Electroblotting of Starch and SDS-PAGE gels

The nitrocellulose membrane (Schleicher & Schuell), Scotchbright pads and sheets of No.3 Whatman paper were soaked in TGM (section 2.1.4) for 10mins. The electroblotting cassette was loaded with the gel and the filter sandwiched between two sheets of Whatman paper and the Scotchbrights, with the gel at the cathode side. Electroblotting conditions for starch gels were 3hrs at 40V, 150mA and for SDS-PAGE gels were 4 hrs at 100V, 400mA. The filter was then blocked in 1X PBS/0.1% Tween 20 (Polyoxyethylene-sorbitan monolaurate) (Sigma) for 30mins.

2.2.4.3.2 Passive Blotting of IEF gels

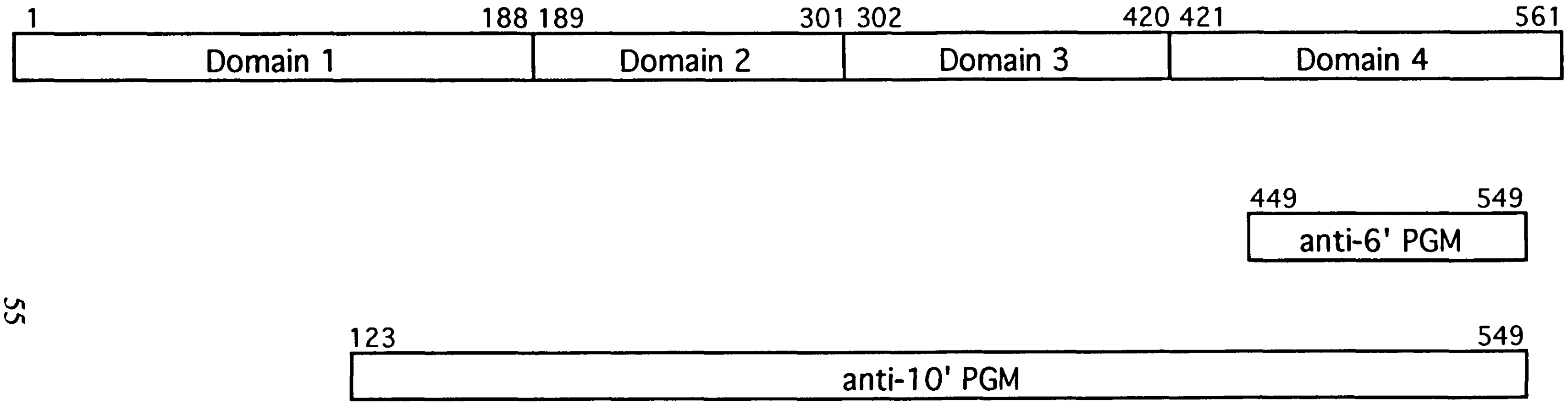
The nitrocellulose membrane was soaked in distilled water and placed on to the gel, followed by two sheets of soaked No.3 Whatman paper and a glass plate. Once wrapped in cling film and a 500g weight placed on top, the blotting assembly was left for 2hrs. The filter was then blocked in 1X PBS/0.1% Tween 20 for 30mins.

2.2.4.3.3 Detection of Antigen

Three polyclonal antibodies were available. The first were anti-rabbit PGM antibodies, raised against commercially available rabbit PGM (Drago, 1992). The other two were anti-human PGM1 polyclonal antibodies; anti-6' PGM antisera was raised against a fusion peptide encoding the majority of domain 4 of PGM1, whilst the anti-10' was raised against a fusion peptide encoding primarily domains 2, 3 and 4 (Figure 2.1).

The primary antiserum was diluted 1 in 500 with 1X PBS/0.1% Tween 20 and incubated with the filter for 17hrs at 22°C. Following 5 X 5min washes in 1X PBS/0.1% Tween 20, the filter was incubated with horseradish peroxidase conjugate rabbit anti-goat immunoglobulins (DAKO) diluted 1 in 500 with 1X PBS/0.1% Tween 20. The rabbit anti-goat IgG's were previously absorbed with human plasma at a ratio of 60µl to 1µl plasma. After 2hrs, the filter was washed for 5 X 5min washes in 1X PBS/0.1% Tween 20. Visualization of bound antigen was by 500µl 3,3'-diaminobenzidine (DAB) (Sigma), 25ml 1X PBS, and 12.5µl 9% hydrogen peroxide (Boots).

Figure 2.1 Diagram to show the domains encoded by the fusion peptides.



2.2.5 PREPARATION OF GENOMIC DNA AND RNA

2.2.5.1 Preparation of K562 Genomic DNA

Pelleted cells were treated with 5ml lysing buffer (10ml TE buffer containing 100mM EDTA, 400µl 10% SDS and 2mg proteinase K), mixed for 1hr at 22°C and then incubated at 37°C for 17hrs. Following two phenol/chloroform extractions and a chloroform extraction, the DNA was precipitated with 0.1vol 5M NaOAc and 2.5vol 100% ethanol. The DNA was hooked out and washed in 70% ethanol. The DNA was dried and resuspended in 50µl of standard TE buffer.

2.2.5.2 Preparation of *E.coli* DNA

Cells were harvested by centrifugation at 1087g at 22°C, and resuspended in 2ml Solution I (50mM Glucose, 10mM EDTA, 25mM Tris, pH 8.0). The bacterial cell wall was disrupted by two cycles of freeze/thawing, and the DNA extracted by a single phenol extraction followed by two chloroform extractions. The DNA was precipitated with 0.1vol 5M NaOAc and 2.5vol 100% ethanol, placed at -20°C for 17hrs. Following centrifugation at 12,000g, for 5mins, the DNA was washed in 70% ethanol, dried and resuspended in distilled water.

2.2.5.2 Preparation of Total RNA

All laboratory equipment (bottles, measuring cylinders, glass pipettes, centrifuge tubes, eppendorfs) was soaked in 0.02% diethylpyrocarbonate (DEPC) overnight, and then autoclaved. All buffers, except 20% SDS, were made with DEPC treated water.

Cell pellets from K562, 6997 and 7014 were resuspended in 4ml of Buffer I (0.01M Tris-HCl pH8.5, 0.0015M $MgCl_2$, 0.14M NaCl), and 80µl of 5% Nonidet P40 (BDH) was added. After vigourous mixing, the samples were centrifuged at 5,876g for 8 mins at 4°C. The supernatant was removed and mixed with 0.2vol of Buffer II (0.05M Tris-HCl pH8.0, 0.01M EDTA pH7.5, 0.01M NaCl, 0.5% SDS). The RNA was extracted by a single phenol/chloroform extraction, followed by 3 chloroform extractions. Precipitation of RNA with 0.05vol 5M NaCl and 3vols of 100% ethanol proceeded at -20°C overnight. The RNA was pelleted by centrifugation at 13,221g for 30mins at room temperature. Following a wash in 70% ethanol, the pellet was resuspended in 50µl of distilled water, and stored at -70°C.

2.2.5.4 Preparation of pA^+ RNA

pA^+RNA was prepared using a Dynabeads mRNA Purification Kit (Dynal). 75µg of K562 total RNA was applied to the oligo$(dT)_{25}$ dynabeads and the resulting pA^+RNA was eluted in 12µl of elution buffer (2mM EDTA pH7.5, Dynal).

2.2.6 ESTIMATION OF NUCLEIC ACID CONCENTRATION

2.2.6.1 Spectrophotometry

The absorbance of the sample was measured at 260nm, and the purity evaluated by scanning between 200nm and 300nm wavelengths. The concentration of the nucleic acid was calculated using the following; 1 OD unit is equivalent to a concentration of 50mg/ml of double stranded DNA and 40 mg/ml of RNA (Sambrook et al, 1989).

2.2.6.2 Molecular Weight Standards

The concentration of PCR products were estimated by direct comparison with known molecular weight standards. 5µl of PCR molecular weight marker (Amersham) contained 50ng of DNA of each band size: 50bp, 100bp, 200bp, 300bp, 400bp, 500bp, 525bp, 700bp and 1000bp.

2.2.7 AGAROSE GEL ELECTROPHORESIS

2.2.7.1 Standard Agarose Gels

DNA was mixed with 0.1 vols of loading buffer (0.25% Bromophenol blue, 40% sucrose). Genomic DNA digested with restriction endonucleases was separated in 1X TBE, 0.8% agarose gels, 240 x 180 x 10mm. Electrophoresis was at 40V, 50mA for 17hrs. PCR products and recombinant plasmids digested with restriction endonucleases were separated in 1X TBE, 2% agarose gels, either 60 x 40 x 8mm or 110 x 140 x 10mm. Electrophoresis was at 55V, 60mA for 1hr or 80V, 100mA for 4hrs respectively. Addition of ethidium bromide at 5µg/l allowed visualization of the DNA by UV light. Estimation of band size was provided by comparison with the 1kb DNA ladder (Gibco BRL).

2.2.7.2 Nusieve Agarose Gels

For the preparation of PCR products for direct sequencing (section 2.2.14.2.1), the amplified DNA was electrophoresed in 1X TAE, 2% nusieve agarose at 4°C.

2.2.7.3 Hybrid Agarose Gels

For greater resolution of DNA fragments, such as digested PCR products, a mixture of nusieve and agarose was used in a ratio of 3:1.

2.2.8 POLYMERASE CHAIN REACTION

2.2.8.1 Genomic DNA PCR

For amplification of genomic DNA, approximately 200ng of DNA was added to a reaction mixture of 1X Promega Original Buffer (10mM Tris-HCl pH8.8, 50mM KCl, 1.5mM $MgCl_2$, 0.1% Triton X-100), 1mM of total dNTPs and 50pmoles of each primer, overlaid with an equal volume of mineral oil. Following heating at 95°C for 5mins, 1U Taq Polymerase (Storage Buffer A [50mM Tris-HCl, pH 8.0, 100mM NaCl, 0.1mM EDTA, 50% Glycerol, 1% Triton X-100], Promega), was added. PCR amplification was carried out for 35 cycles with denaturing at 94°C for 20s, annealing for 45s and elongating at 72°C for 45s. Primer sequences and annealing temperatures are shown in Figure 2.2.

2.2.8.2 Reverse Transcriptase PCR (RT-PCR) of *PGM1*

2.5μg of RNA were added to a mixture of 1X PCR reaction buffer (10mM Tris-HCl pH8.8, 50mM KCl, 1.5mM $MgCl_2$, 0.1% Triton X-100), 1mM of total dNTP's and 250pmoles random hexamer in a total volume of 23μl. After heating at 65°C for 10mins, 200U of Murine moloney-leukaemia virus reverse transcriptase (MMLV-RT) was added to give a final volume of 25μl. This was incubated at 42°C for 60 - 90 mins.

The entire cDNA sample was used for the subsequent PCR reaction, the volumes of 1X PCR reaction buffer and 1mM total dNTP's adjusted to a final volume of 100μl. 50pmoles of each primer was used. As previously, following heating at 95°C for 5mins, 1U Taq Polymerase was added. PCR amplification was carried out for 35 cycles with denaturing at 94°C for 20s, annealing for 45s

Figure 2.2 Oligonucleotide primers used for PCR.

	Forward primer	Reverse primer	Annealing Temperature °C
PGM cDNA primers:			
Pair 1	5' GCCAGCCAAGTCCGCCGCTCTGAC 3'	5' GGGGCCCCCTGGGTTGTGACTGGCT 3'	64
Pair 2	5' GAAAAATCAAAGCCATTGGTGGGAT 3'	5' GGCACCGAGTTCTTCACAGAGGATC 3'	56
Pair 3	5' GCTATGCATGGAGTTGTGGGACCGT 3'	5' TTGTAGCACTAGCCACCCGGTCCAG 3'	57
Pair 4	5' CTTTGCACGGAGCATGCCCACGAGT 3'	5' TGTTTGCGCCCTCAGCTTCCACCTC 3'	58
Pair 5	5' GATCATTGGCAAAAGCATGGCCGGA 3'	5' TCGATGTACAGCCGAATGGTGGCCC 3'	56
Pair 6	5' TCCGACTGAGCGGCACTGGGAGTGC 3'	5' GCCCGCAGGTCCTCTTTCCCTCACA 3'	60
Low stringency PCR primers:			
Ser116F/MgR	5' TCATTCTGACAGCCAGTCACAACCC 3'	5' GATCCCCATCTCCATCAAAGGCAGC 3'	35-55
MgSerF/MgR	5' AATGGAGGTCCTGCTCCAGAAGCAA 3'	5' GATCCCCATCTCCATCAAAGGCAGC 3'	35 - 1 cycle 60 - 35 cycles
Hyhbf primers:			
Hyhbf.F/Hyhbf.R	5' AAGCTGCCGGACGAGATC 3'	5' CTCGACATGGGTCGAACC 3'	55
Hyhbf.F2/Hyhbf.R2	5' GAAGAGTTGCTCGATCAGCC 3'	5' ACCATCATTGATGTTCAAG 3'	55
EST primers:			
EST.F/EST.R	5' CTCTTTTCTGATATAACGGC 3'	5' TCCCAATAGACCTTATAACC 3'	50
EST.F2/EST.R2	5' GACTTGCTGCAACCACATTT 3'	5' GCAGTTATCATGATTCCAGC 3'	50
PMM.F/PMM.R	5' ATCGGTGTGGTGGGCGGCTCT 3'	5' GCAAGCCCACCCAGATGGGG 3'	60

Figure 2.2 cont.

PGM Genomic DNA primers:			
PGM exon 1	5' GACCCAGGCGTACCAGGACC 3'	5' CCCGTTGGCGGCAGCGATGC 3'	61
PGM exon 4	5' GCAGGTTTACAGCAATATAGTCACA 3'	5' TGAAGCATCATGATACACACAGAAG 3'	50
PGM exon 5	5' GTGCCCCTGCGAACTCGGCAGTTA 3'	5' GATCCCCATCTCCATCAAAGGCAGC 3'	57
PGM exon 8	5' GGGATGCAGAGCCAAACCATATCAAG 3'	5' TAAGACAGGAGAGGCTGTGGATGCG 3'	55
PGM exon 11	5' AAGCTTCTCTCTATGTCTTCCTCAG 3'	5' GCCCGCAGGTCCTCTTTCCCTCACA 3'	55
Miscellaneous primers:			
Control PGD cDNA	5' GTCTGTGCTTTTAATAGGAC 3'	5' GATGATGTCACCAGGTATCC 3'	50
pCDM8F/pCDM8R	5' GAACCCACTGCTTAACTGGC 3'	5' CGCAGAACTGGTAGGTATGG 3'	55
pCDM8F2/pCDM8R2	5' GACTCACTATAGGGAGACCC 3'	5' AAGATCCTCTAGAGTCGCGG 3'	55

and elongating at 72°C for 45s. Primer sequences and annealing temperatures are shown in Figure 2.2.

2.2.8.3 Low Stringency RT-PCR

5μg of RNA were added to a mixture of 1X RT reaction buffer (50mM Tris-HCl pH8.3, 75mM KCl, 3mM $MgCl_2$), 10mM DTT, 1mM of total dNTP's, 10U placental RNAse inhibitor, and 250pmoles random hexamer in a total volume of 33μl. After heating at 65°C for 10mins, 400U of MMLV-RT was added to give a final volume of 35μl. This was incubated at 42°C for 60 - 90 mins.

5μl of cDNA was added to a PCR reaction mix of 1X Original Buffer, 1mM of total dNTPs and 50pmoles of each primer. 1U Taq Polymerase was added following a denaturing step of 95°C for 5mins. PCR amplification was carried out for 30 cycles with denaturing at 94°C for 1min, annealing for 2min and elongating at 72°C for 2min. Annealing temperatures of Tm-30°C to Tm-10°C were investigated. Primer sequences are shown in Figure 2.2.

2.2.8.4 Degenerate Primer PCR

First strand cDNA was prepared as detailed in section 2.2.8.3. 5μl was added to a PCR reaction mix of 1X Original Buffer, 1mM of total dNTPs and 10nmoles of each primer. For amplification from genomic DNA, approximately 200ng of DNA was added to the PCR reaction mix. PCR amplification was carried out for 35 cycles with denaturing at 94°C for 30s, annealing for 30s and elongating at 72°C for 45s. All primer sequences and annealing temperatures are provided in Chapter Four, in figures 4.8 and 4.14.

Nested degenerate primer PCR reactions were set up as above, except the primer concentration was reduced to 100pmoles, and each round of amplification was for 25 cycles. 1μl of PCR product from the first round was used as template for the second.

2.2.8.5 'Touchdown' PCR

In some PCR experiments, the specificity of the PCR was increased using a 'touchdown' programme (Don et al, 1994). The PCR begins at an annealing temperature above the melting temperature (Tm) of the primers, and is decreased by 1°C every second cycle, over nine stages to touchdown at between Tm and Tm-5°C for 30 cycles of PCR. Tm values for PCR primers were calculated using the following equation: Tm°C = 69.3 + 0.41(G+C%) - (650 + number of bases in oligomer)

2.2.8.6 Primers

Oligonucleotide primers were obtained from Oswel DNA Service, (University of Edinburgh and University of Southampton). Primers were normally supplied dissoved in 1ml. However, the degenerate primers were supplied as freeze dried DNA, which was then dissolved to provide a 10nmoles working solution. The sequences of genomic, RT-PCR and low stringency PCR primers, along with the annealing temperatures used, are provided in figure 2.2.

2.2.9 RESTRICTION ENZYME DIGESTS

Digests were set up according to suppliers instructions. Following incubation of the DNA at 65°C for 5mins, 1X reaction buffer, 20mM spermidine and restriction endonuclease, at 3U/µg of DNA, were added. Digests were incubated at the recommended temperature overnight. Digests of genomic DNA contained 5 or 10µg of DNA, digests of PCR products generally used 9.5µl of product, in a total reaction volume of 12.5µl, and digests of plasmid DNA, to determine the size of cloned insert, used 300ng DNA.

2.2.10 SOUTHERN BLOT ANALYSIS

2.2.10.1 Transfer of DNA to Nitrocellulose

The gel was denaturated in 1.5M NaCl, 0.5M NaOH, for 30mins, and then neutralized in 1.5M NaCl, 0.5M Tris-HCl, 1mM EDTA, pH 7.2, for 1 - 4hrs on a shaker at 22°C. The DNA was then transferred from the gel onto Hybond N+ nylon membrane (Amersham Life Science), according to supplier instructions, using the method of capillary blotting. After 17hrs, the DNA was fixed to the membrane by baking at 80°C for 2hrs.

2.2.10.2 Preparation of Hybridization Probes

Hybridization probes were either derived from cloned DNA or PCR products. For cloned DNA, 5µg of plasmid DNA was digested. For PCR products, "needle PCR" was performed; the initial PCR product was stabbed with a needle, and the DNA transferred into a second PCR reaction for reamplification. 180µl of this product was used to prepare the probe. Following electrophoresis, the DNA was excised from the gel under long wave UV and the DNA was eluted using one of the methods described below. For the electroelution and spinning through glass wool techniques, normal agarose and 1X TBE were used, whereas for the Wizard DNA PCR prep purification kit (Promega), nusieve agarose and 1X TAE were used.

2.2.10.2.1 Electroelution

The gel slice was placed against the side of a length of dialysis tubing and 1ml of 1X TBE was added. This was then placed in the gel tank and electrophoresis was carried out at 100V, 110mA for 90mins, eluting the DNA from the gel slice. The DNA was precipitated from the TBE by the addition of 0.1vol 2M NaOAc and 2.5vol 100% ethanol. Following 17hrs at -20°C, the DNA was spun down by centrifugation at 12,000g for 15mins at 4°C. After a wash in 70% ethanol, the DNA was resuspended in a final volume of 20µl of distilled water.

2.2.10.2.2 Spinning Through Glass Wool

This technique was based on the method of He et al, (1992). The excised gel slice was placed on siliconized glass wool in a small eppendorf containing a hole in the bottom. This was then placed in a large eppendorf and spun at 12,000g for 45secs at 22°C. The DNA was precipitated with 0.1vol 2M NaOAc and 2.5vol 100% ethanol, either at -20°C overnight or -70°C for 1 hr. Once washed in 70% ethanol, the pellet was resuspended in 10µl distilled water.

2.2.10.2.3 Wizard DNA PCR Prep Purification Kit

The DNA was eluted from the gel slice according to the protocol provided with one exception: following elution of the DNA in 50µl, a further 25µl of sterile water was added and spun through the column to ensure complete elution of the DNA.

2.2.10.3 Prehybridization of Filter

The filter was pretreated with hybridization buffer (5X SSC, 5X Denhardts, 0.5% SDS), at 65°C, for at least 1hr prior to addition of the probe. 100µg Herring sperm DNA (DNA Type XIV from Herring Testes) was boiled for 5mins and then added to preheated hybridization buffer immediately before application to the membrane.

2.2.10.4 Labelling of Probe

The HPGM1 probe was labelled either using the Multiprime DNA Labelling System or the Rediprime DNA Labelling System (both Amersham Life Science), according to the manufacturers rapid protocol. 25µg of DNA was labelled with 3-5µl ^{32}P dCTP, depending on the reference date. The reaction was either incubated at 37°C for 20mins or at 22°C for 2-4hrs. The unincorporated nucleotides were removed from the labelled probe by centrifugation of the labelling mix through a column containing G50 sephadex (Pharmacia) in TE buffer.

2.2.10.5 Hybridization

The labelled probe was boiled for 5mins before application to the filter, to denature the DNA. The hybridization proceeded for 17hrs at 65°C.

2.2.10.6 Stringency Washes

Filters were washed down twice in Wash I (2X SSC, 0.1% SDS) for 10mins at room temperature, once in Wash II (1X SSC, 0.1% SDS) for 15mins at 65°C, and finally in Wash III (0.1X SSC, 0.1% SDS) for 10mins at 65°C. Filters were monitored after each wash and put down when the counts were approximately 2cps, even though not all of the washes may have been carried out. Autoradiography was at -70°C.

2.2.11 DETERMINATION OF THE PGM1 POLYMORPHSIM

Following PCR of genomic DNA to amplify exon 4, encompassing the site coding the 2/1 protein allele(s) and exon 8, encompassing the site coding the +/- protein allele(s), the PCR products were used to determine the PGM1 polymorphism, either by single strand conformation polymorphism (SSCP) analysis (Orita et al, 1989a; Orita et al, 1989b; March et al, 1993b) or by diagnostic restriction endonuclease digestion (March et al, 1993a).

2.2.11.1 SSCP Analysis

SSCP analysis was carried out using the Phastsystem (Pharmacia). An equal volume of PCR product and SSCP loading buffer (95% formamide, 0.02M EDTA pH 8.0, 0.05% bromophenol blue) were heated at 95°C for 10mins. The samples were electrophoresed on Homogeneous 20 Phastgels (Pharmacia), with Phastgel Native Buffer Strips (Pharmacia) used to form a contact between gel and electrode. The programme consisted of three stages; pre-run of 400V, 20mA, 2W, 10Vhr, sample application of 400V, 5mA, 2W, 2Vhr and separation of 400V, 10mA, 2W, 175Vhr. For exon 4 PCR products, the temperature of the run was 5°C, whilst for exon 8 PCR products it was 10 °C. Following separation, the gel was silver stained in the developing chamber of the Phastsystem according to the manufacturers protocol.

2.2.11.2 Restriction Endonuclease Analysis

In both exon 4 and exon 8, the nucleotide substitutions which underlie the common polymorphisms alter restriction endonuclease recognition sites. In exon 4 PCR products, *BglII* cleaves the PGM1*2 allele (AGATC**T**) but not the PGM1*1 allele (AGATC**C**). However, a reciprocal digest with *AlwI* cleaves the PGM1*1 allele, (**G**GATCN$_4$), but not the PGM1*2 allele (**A**GATCN$_4$). In exon 8 PCR products, NlaIII cleaves the PGM1*- allele (**C**ATG), but not the PGM1*+ allele (**T**ATG). Prior to digestion, PCR products were concentrated two-fold by ethanol precipitation. Digestion was carried out according to manufacturers recommendations, and the resulting DNA fragments were electrophoresed on 6% hybrid agarose gels (section 2.2.7.3).

2.2.12 CLONING OF DEGENERATE PRIMER PCR PRODUCTS

Following amplification of cDNA using degenerate primers, the PCR products were ligated into the plasmid vector pCRII, and transformed into *E.coli* INVαF', according to the protocol supplied with the TA Cloning Kit (Invitrogen, R&D Systems).

2.2.13 PREPARATION OF CLONED DNA

Recombinant plasmids containing degenerate primer generated inserts were grown up in L-broth supplemented with 50μg/ml of ampicillin. To analyze the size of the cloned inserts, the "quick miniprep" technique was used to prepare the DNA. To obtain DNA for sequencing, the Wizard Maxipreps Purification System (Promega) was used.

2.2.13.1 The "Quick Mini-Prep"

This technique is based on the method of Jones and Schofield, (1990). Briefly, 4mls of culture were spun down and the cells resuspended in 150μl of Solution I

(50mM Glucose, 10mM EDTA, 25mM Tris, pH 8.0) and 300µl of Solution II (0.2M NaOH, 1% SDS). Following 5mins incubation on ice, 225µl of 3M KOAc, pH 4.8, was added. Following a further 5mins incubation on ice, the samples were centifuged and the DNA precipitated by the addition of 100% ethanol to the supernatant. The DNA was pelleted, washed in 70% ethanol, dried and resuspended in 20µl of sterile water containing 12.5µg/ml RNase.

2.2.13.2 Preparation of DNA for Sequencing

To ensure high quality DNA for double-stranded sequencing, the recombinant plasmid DNA was prepared using the Wizard Maxipreps Purification System (Promega), according to the protocol supplied by the manufacturers. The single exception was that the resuspended DNA pellet, once mixed with the purification resin, was left for 10mins, rather than preceeding on with the next step immediately.

2.2.14 SEQUENCING OF PCR PRODUCTS

Double stranded sequencing of cloned PCR products and direct sequencing of PCR products utilized the Sequenase DNA Sequencing Kit (USB, Amersham), based on the principle of dideoxy sequencing (Sanger et al, 1977).

2.2.14.1 Sequencing of Cloned DNA

5µg of plasmid DNA was denatured by NaOH and then ethanol precipitated to provide a sample of single stranded DNA in 7µl of sterile water. The primers used for sequencing were M13 (-24) reverse primer 5' AACAGCTATGACCATG 3', which sequenced the forward strand and the M13 (-40) forward primer 5' GTTTTCCCAGTCACGAC 3', which sequenced the reverse strand. 1.0pmole of primer was used in each sequencing reaction. The reactions were carried out following the protocol supplied with the Sequenase kit.

2.2.14.2 Direct Sequencing of PCR Products

The PCR products for sequencing were prepared by "needle PCR" (section 2.2.10.2) and 180µl of product was electrophoresed in nusieve agarose and 1X TAE. The DNA was purified using the Wizard DNA PCR Prep Purification Kit (Promega) (section 2.2.10.2.3).

The sequencing reactions were performed according to the protocol, with three exceptions: i) dimethylsulfoxide (DMSO) was added in a ratio of 1:9 to each of the termination mixes, ii) 1µl of DMSO was added to the labelling reaction, which contained 300ng of DNA and 300ng of PCR primer, and iii) the labelling reaction was heated at 99°C for 2mins and immediately placed on ice.

2.2.14.3 Polyacrylamide Gel Electrophoresis

6% polyacrylamide gels, 210 x 500 x 0.4-1.2mm, were made using 80ml ready-to-use sequencing gel solution containing a ratio of 19:1 acrylamide:bis. They were set using 140µl of TEMED and 140µl of 25% AMPS. The gels were pre-warmed for approximately 1hr at 2500V, 43mA and 70W, until they reached 55°C. The samples were incubated at 72°C for 2 mins and 3µl of sequencing reaction was loaded. Electrophoresis then continued at between 45-48W, to maintain the temperature of the gel at 50°C.

2.2.15 COMPUTER ANALYSIS

2.2.15.1 GCG Wisconsin Package

The Genetics Computer Group (GCG) Wisconsin Package was available via the Human Genome Mapping Project Resource Centre. The software programmes used include:

bestfit	makes an optimal alignment of the best segment of similarity between two sequences
blast (and derivatives)	searches for sequences, either peptide or nucleic acids, similar to a query sequence
fasta (and derivatives)	searches for similarities to a query sequence; it is more sensitive than blast
map	displays both strands of a DNA sequence with restriction sites and possible protein translations
mfold	predicts optimal and suboptimal secondary structures for an RNA molecule using the most recent energy minimization method of Zucker
peptidesort	gives the digested peptide fragments of an amino acid sequence and summarizes the composition of the whole protein
pileup	creates a multiple sequence alignment from a group of related sequences using progressive pairwise alignments

seqed	is an interactive editor for entering and modifying sequences
stringsearch	identifies sequences by searching for keywords in the sequence information
translate	translates nucleotide sequence into peptide sequence

2.2.15.2 Phylogenetic Analysis

Phylogenies were constructed using the software package PAUP - Phylogenetic Analysis Using Parsimony (Swofford, 1990) and neighbour-joining distance method (Satiou & Nei, 1987).

CHAPTER THREE:

CHARACTERIZATION OF THE CELL LINE K562

K562 is an erythroleukaemic cell line derived from a pleural effusion of a patient with chronic myelogenous leukaemia in terminal blast crisis (Lozzio & Lozzio, 1975, Andersson et al, 1979). Undifferentiated K562 cells show markers characteristic of erythropoiesis, such as spectrin, and of granulopoiesis, such as My-1 (Marie et al, 1981), indicating abnormal gene expression. In addition, the cell line can be induced to differentiate along erythroid and megakaryocytic lineages, with the subsequent expression of haemoglobin and acetylcholinesterase (Ajmar et al, 1983) and integrin (Fong & Santoro, 1994)

Investigations into the genetic stability of human cell lines, by comparing the electrophoretic patterns of a variety of cytosolic enzymes, revealed an abnormal PGM pattern in K562 (Povey et al, 1980). The PGM1 isozymes were absent, whilst there was an increase in activity of the PGM2 and PGM3 isozymes. No abnormal protein phenotypes in the glycolytic enzymes glucose phosphate isomerase and lactate dehydrogenase, nor in any of the other metabolic enzymes assayed, were observed in K562. Of the other human cell lines investigated, all showed normal PGM1 phenotypes. Thus, the absence of PGM1 activity in K562 is a unique characteristic of the cell line.

PGM1 is the major isozyme of phosphoglucomutase activity. Therefore, if the enzyme deficiency in K562 is due to rearrangements of the PGM1 gene(s), this cell line might be an ideal resource for cloning other members of the PGM gene family using cDNA strategies. Therefore, the molecular basis of this deficiency was investigated and the cell line was characterized, with respect to PGM1, at the level of the protein, the gene and the RNA.

In parallel with these studies, two anti-human PGM1 polyclonal antibodies were investigated to determine their cross-reactivity to PGM2 and PGM3, and thus assess their usefulness as tools for screening cDNA expression libraries.

3.1 CHARACTERIZATION OF K562 AT THE PROTEIN LEVEL

3.1.1 DETECTION OF PGM ACTIVITY

The K562 cell line used in this study was examined to verify the PGM phenotype observed by activity staining.

3.1.1.1 Starch Gel Electrophoresis and Isoelectric Focusing

Extracts of K562 cells and human placentae were electrophoresed on starch gels and stained for PGM activity. The unusual phenotype, an absence of PGM1 and increased activity of PGM2, was observed in K562 (Figure 3.1a). Prolonged incubation of the activity stain allowed the detection of the PGM3 isozymes (Figure 3.1b), which also showed increased activity in K562 compared to the placentae.

Starch gel electrophoresis separates proteins by net charge and the molecular seiving affect of the starch, producing diffuse areas of activity staining where the proteins have migrated to. In comparison, isoelectric focusing (IEF) separates proteins according to net charge alone and due to the pH gradient set up across the gel, well defined bands are produced at the isoelectric point of the protein following activity staining. Due to the greater concentration of protein at a single point, K562 was electrophoresed on IEF gels to determine if PGM1 activity could be detected. As can be seen from figure 3.2, no PGM1 activity was observed in K562. The single well defined band is PGM2.

3.1.1.2 Estimation of the Sensitivity of the PGM Activity Stain

The sensitivity of the PGM activity stain was studied to give an estimation of the minimum level of enzyme activity that has to be present to be detected. A serial dilution of a PGM1*1+ placental extract was carried out and electrophoresed by IEF. The activity stain detected the PGM1 isozyme at a 1 in 64 dilution (Figure 3.3a). Previous studies indicated that PGM1 activity could be increased with the addition of of an equal volume of 0.5% haemoglobin (Drago, 1992). In this case, the PGM1 activity of the placental extract could be detected at 1 in 512, an eight fold enhancement (Figure 3.3b). K562 extracts mixed with 0.5% haemoglobin did not show any PGM1 activity (Figure 3.3c).

3.1.2 DETECTION OF PGM ANTIGEN

Having confirmed the absence of PGM1 activity in K562, the possibility of an enzymically inactive form was considered.

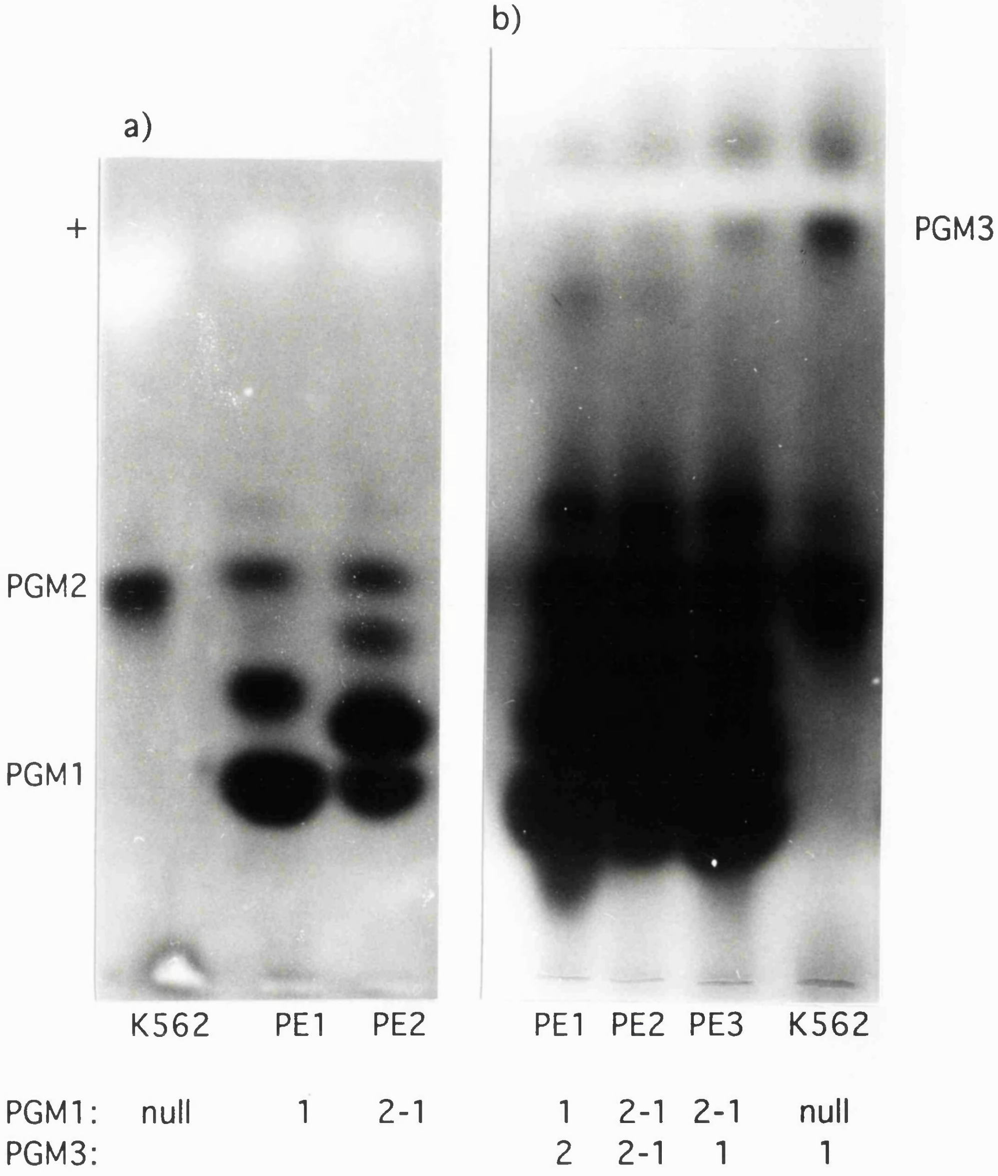


Figure 3.1 Detection of PGM isozymes by enzyme activity staining following starch gel electrophoresis of K562 and placental extracts a)Gel stained for PGM1 and PGM2 isozymes. b) Gel over-stained to detect PGM3 isozymes.

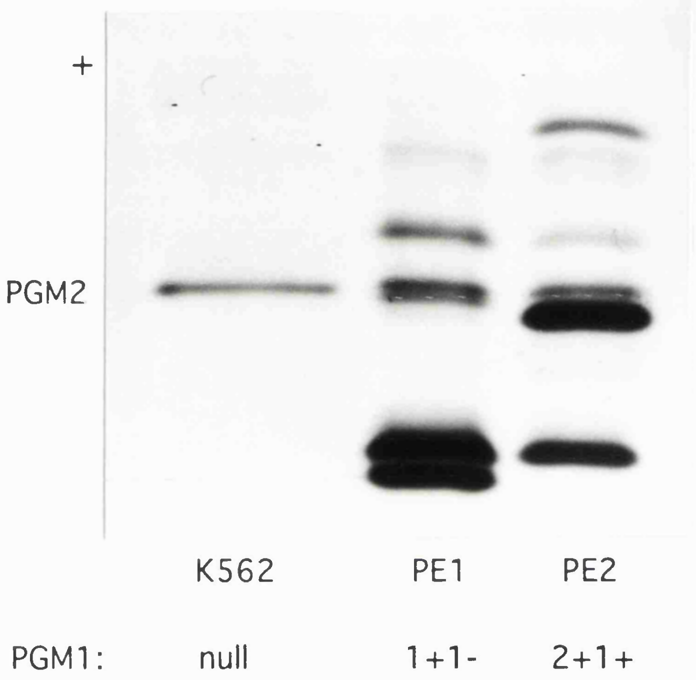


Figure 3.2 Detection of PGM isozymes by enzyme activity staining following polyacrylamide gel isoelectric focusing of K562 and placental extracts.

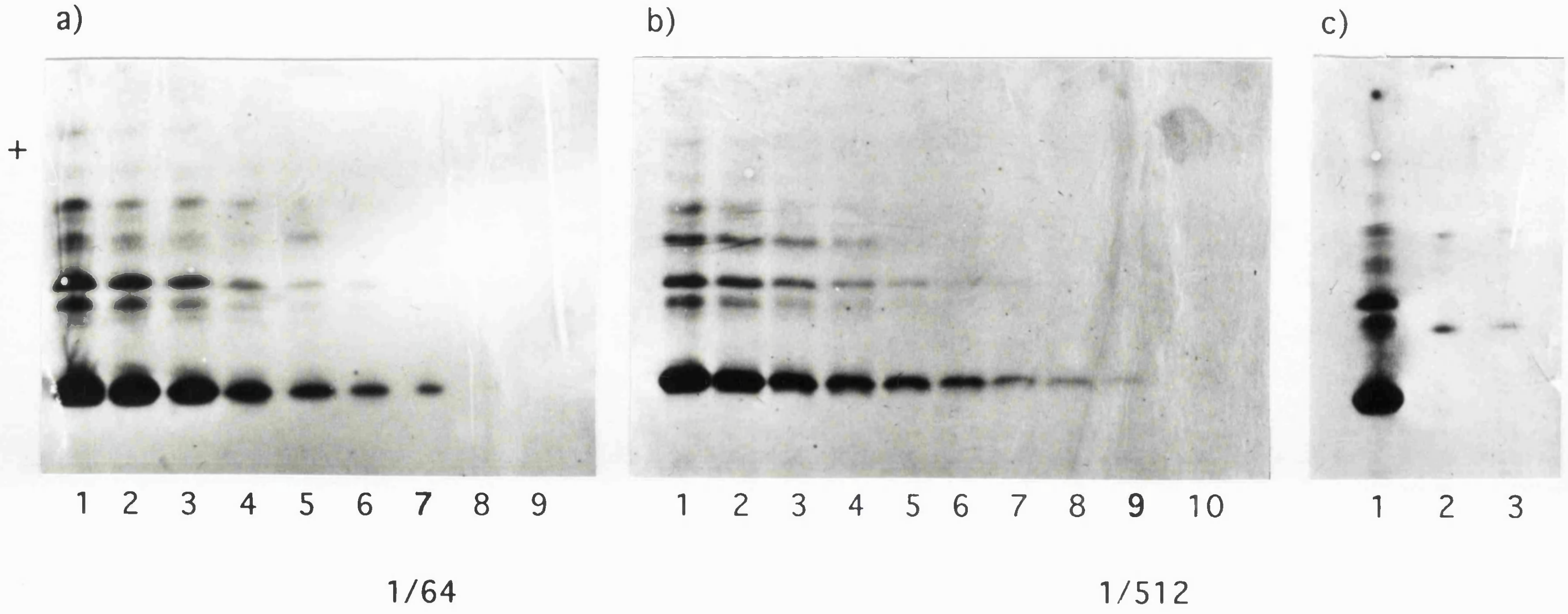


Figure 3.3 Estimation of sensitivity of the PGM activity stain following isoelectric focusing.
a) Serial dilutions of placental extract PGM1*1+ from undiluted in lane 1 to 1 in 256 in lane 9.
b) Serial dilutions of placental extract PGM1*1+ from undiluted to 1 in 256 mixed with an equal volume of 0.5% haemoglobin; lane 10 is 0.5% haemoglobin.
c) K562 extract undiluted in lane 2 and mixed with an equal volume of 0.5% haemoglobin in lane 3. Lane 1 is undiluted placental extract PGM1*1+.

3.1.2.1 Immunoblot Detection Using Anti-Rabbit PGM Antibodies

A method based on standard Western blot techniques was devised to electrophoretically transfer proteins from starch gels onto nitrocellulose membrane. The PGM1 in the placental extracts was detected by the anti-rabbit PGM polyclonal antibodies. However, nothing was observed in K562, suggesting the PGM1 protein was absent from these cells (Figure 3.4a). Immunodetection was also carried out following IEF. The anti-rabbit PGM detected the PGM1 isozymes in placentae, but again, no immunoreactivity was seen in K562 (Figure 3.5a).

In order to determine if any abnormally sized forms of PGM1 antigen were present, that had not previously been detected following starch gel electrophoresis and IEF, SDS-PAGE was carried out. PGM1 was identified as expected as a band of approximately 62,000mw in the placental extracts but not in K562 (Figure 3.6a). There was also no evidence for an abnormally sized PGM1 in K562.

3.1.2.2 Estimation of the Sensitivity of the Anti-Rabbit PGM

The sensitivity of immunoblot detection was investigated on IEF gels, using the procedure previously used to determine the sensitivity of the activity stain. The placental PGM1 isozymes could be detected at a dilution of 1 in 32 (Figure 3.7a). In this case, addition of 0.5% haemoglobin did not improve the sensitivity; the level of detection remaining at 1 in 32 (Figure 3.7b). K562 extracts did not show any regions of immunoreactivity with or without the addition of 0.5% haemoglobin (Figure 3.7c).

3.1.2.3 Immunoblot Detection Using Anti-Human PGM1 Antibodies

Anti-human PGM1 polyclonal antibodies, anti-6' PGM and anti-10' PGM, were also used for immunoblot detection of PGM1. In addition, the immunoreactivity of these antibodies towards PGM2 and PGM3 was analyzed, as K562 expresses higher levels of PGM2 and PGM3.

Immunoblot detection following starch gel electrophoresis identified no regions of antigen binding in K562 corresponding to the PGM1 isozymes. In addition, no region of immunoreactivity corresponding to the PGM2 or PGM3 isozymes was evident in either the placenta or K562 (Figure 3.8). Following IEF, again no specific immunoreactivity was observed with the K562 extract, and in the

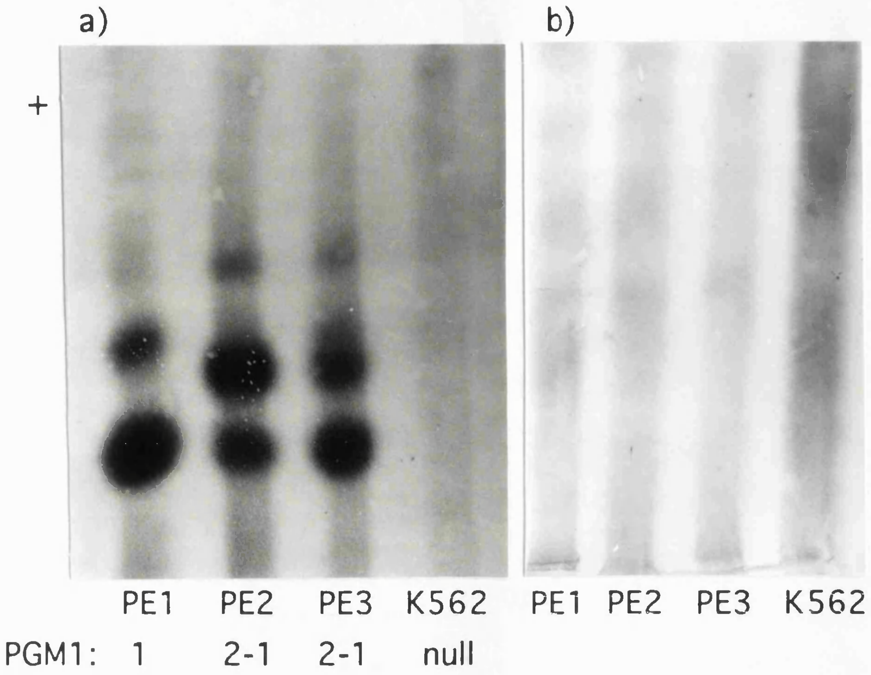


Figure 3.4 Detection of PGM1 isozymes by immunoblot analysis following starch gel electrophoresis of K562 and placental extracts.
a) Detection with anti-rabbit PGM polyclonal antibodies.
b) Detection with pre-immune serum.

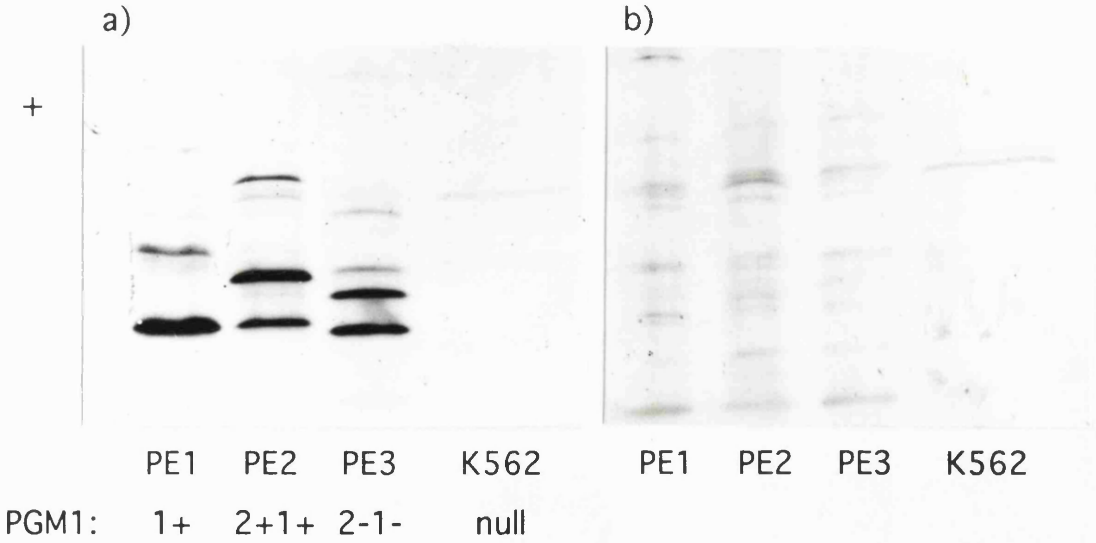


Figure 3.5 Detection of PGM1 isozymes by immunoblot analysis following isoelectric focusing of K562 and placental extracts.
a) Detection with anti-rabbit PGM polyclonal antibodies.
b) Detection with pre-immune serum.

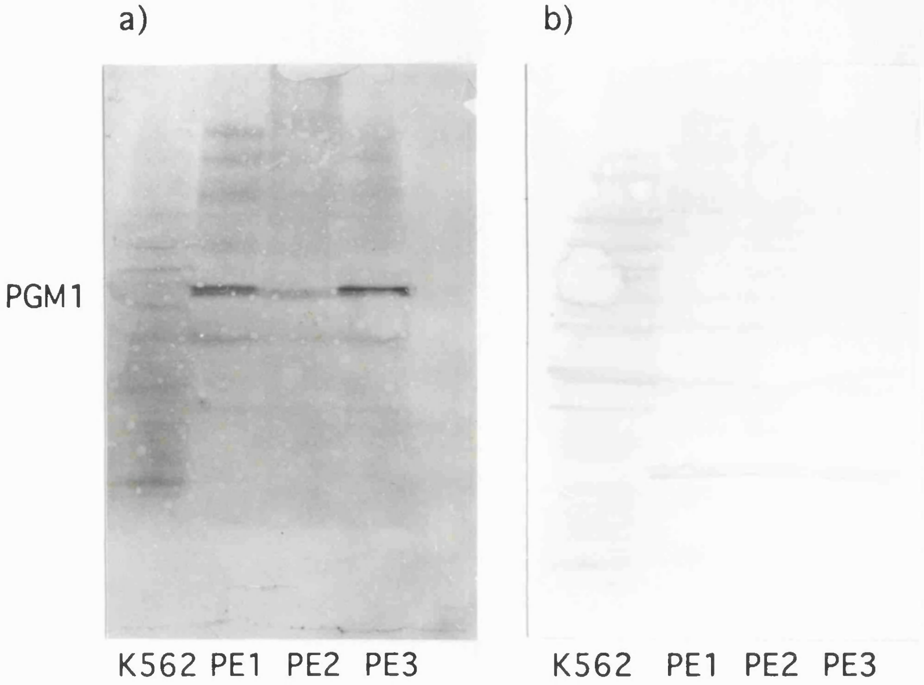


Figure 3.6 Detection of PGM1 isozymes by immunoblot analysis following SDS-PAGE of K562 and placental extracts.
a) Detection with anti-rabbit PGM polyclonal antibodies.
b) Detection with pre-immune serum.

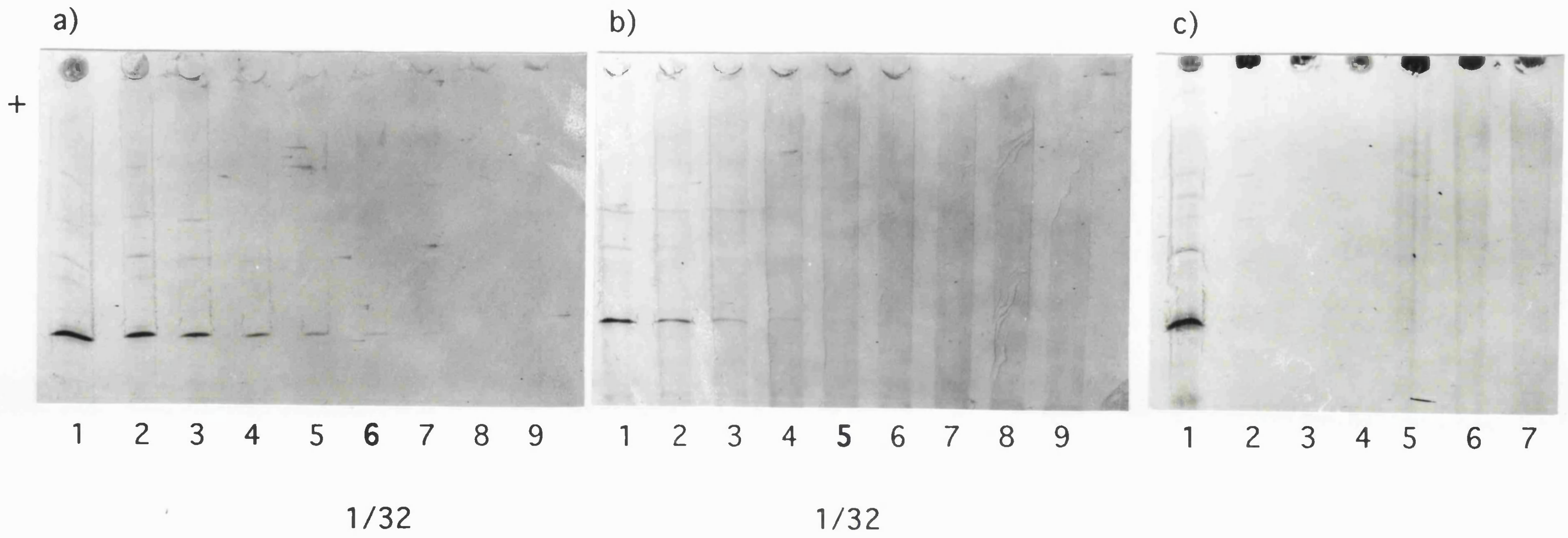


Figure 3.7 Estimation of sensitivity of the anti-rabbit PGM polyclonal antibodies following isoelectric focusing.

a) Serial dilutions of placental extract PGM1*1+ from undiluted in lane 1 to 1 in 256 in lane 9.

b) Serial dilutions of placental extract PGM1*1+ from undiluted to 1 in 256 mixed with an equal volume of 0.5% haemoglobin; lane 10 is 0.5% haemoglobin.

c) Serial dilutions of K562 extract; undiluted in lane 2 to 1 in 4 in lane 4, and mixed with an equal volume of 0.5% haemoglobin in lanes 5, 6 and 7. Lane 1 is undiluted placental extract PGM1*1+.

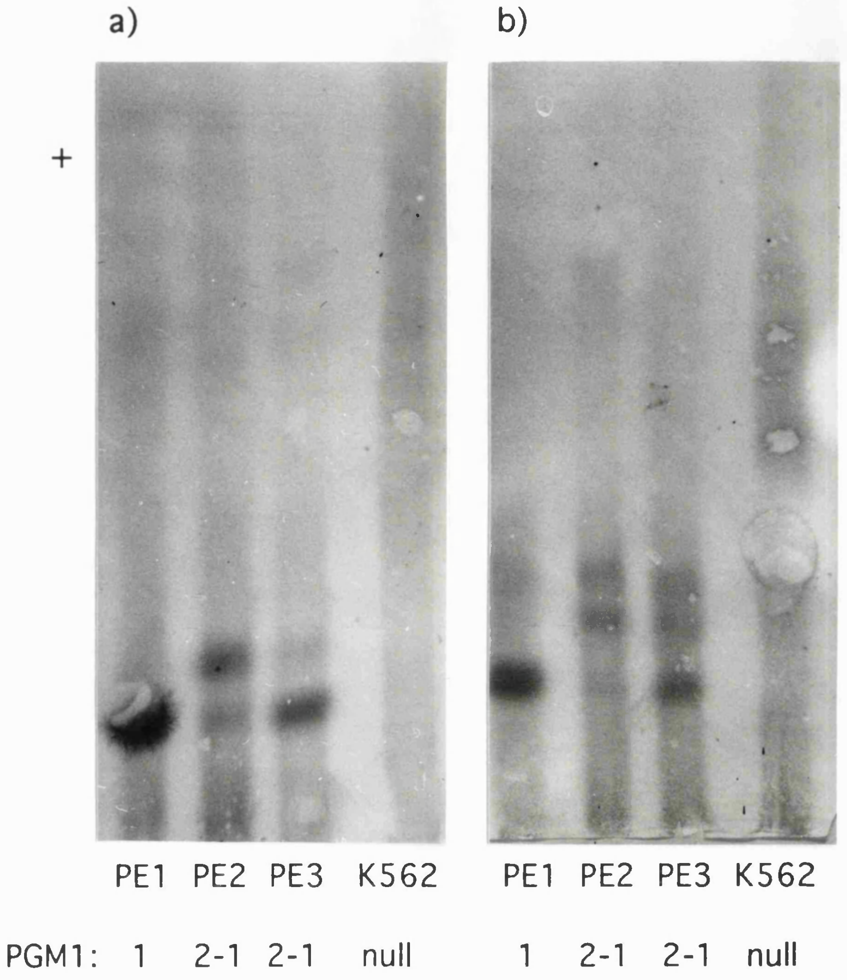


Figure 3.8 Determination of immunoreactivity of anti-human PGM1 polyclonal antibodies to PGM2 and PGM3 by immunoblot analysis following starch gel electrophoresis of K562 and placental extracts. a) Detection with anti-6' PGM polyclonal antibodies. b) Detection with anti-10' PGM polyclonal antibodies. Detection with pre-immune serum is shown in Figure 3.4.

placental extracts, only PGM1 was detected (Figure 3.9). The faint band observed in K562 in the PGM2 position is non-specific since it is present when the pre-immune sheep serum is used instead of the polyclonal antibodies (refer to figure 3.5b).

Finally, the anti-human PGM1 antibodies were used to detect antigen following SDS-PAGE. The PGM1 protein was detected in placental extracts by the anti-6' PGM antibodies, but the anti-10' PGM showed very little specific antigen binding (Figure 3.10). The high background staining is non-specific and is seen in the pre-immune sheep serum (refer back to figure 3.6b), despite preadsorption of the antibodies with both K562 and placental extracts. No specific immunoreactivity was observed towards proteins of higher molecular weight, as we would expect if they reacted with PGM2 or PGM3 (estimated to be 73,000mw and 64,000mw respectively) in either the placental or K562 extracts. Therefore, the anti-human PGM1 polyclonal antibodies, anti-6' PGM and anti-10' PGM, do not cross-react with PGM2 or PGM3.

In summary, there is no evidence of enzymically inactive PGM1 isozymes in K562, nor of abnormally sized PGM1 proteins. The basis of this apparent absence of PGM1 protein was investigated further by first looking at the gene and then the mRNA.

3.2 CHARACTERIZATION OF THE *PGM1* GENE

3.2.1 FLUORESCENCE *IN-SITU* HYBRIDIZATION

The K562 karyotype is known to be triploid in nature, and shows three apparent chromosomes 1. Cytogenetic analysis of this cell line using chromosome specific paints was performed by Dr. Jenny Parrington and Dr Margaret Fox. They determined that there are two normal chromosomes 1, one derivative chromosome 1, der(1)t(1;11)(p32;q24), and three additional chromosomes which contain some part of chromosome 1, der(18)t(1;18)(p32;q23), der(21)t(1;21)(q31:p13) and der(1)t(1;6;20)(p21 q12;q25;q11) (Figure 3.11a, b & c) (Fox et al, 1996). Using the PGM1 genomic clone (Lo HPGM1) in fluorescence *in-situ* hybridization (FISH) experiments, they located the *PGM1* gene on the two normal chromosomes and the first derivative chromosome 1 (Figure 3.11d). All the signals were localized to 1p31. Thus, despite the complex karyotype of this cell line, three *PGM1* genes appeared to be present, unaffected by gross rearrangements.

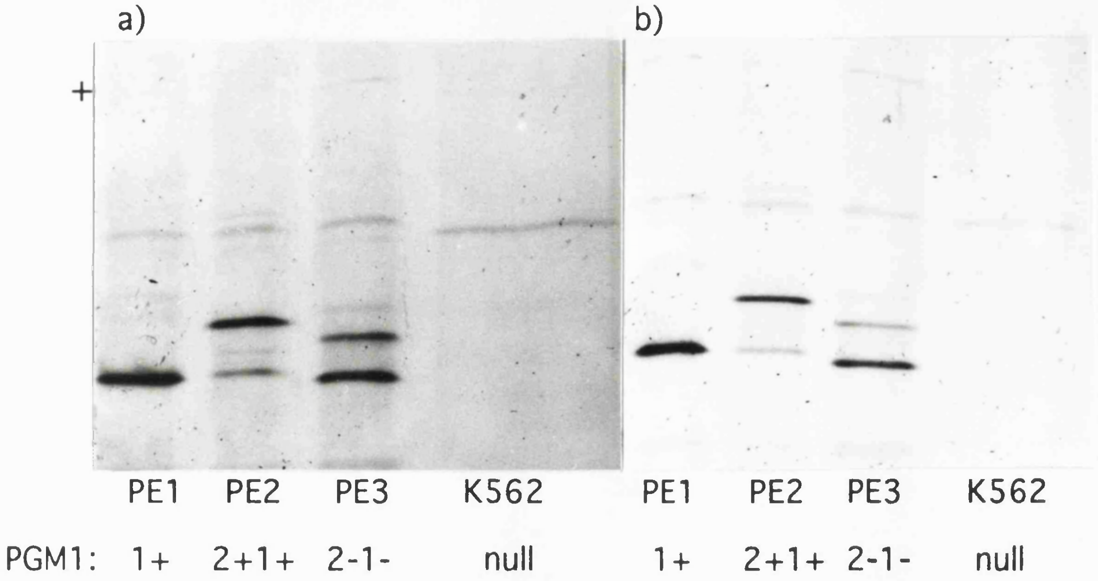


Figure 3.9 Determination of immunoreactivity of anti-human PGM1 polyclonal antibodies by immunoblot detection following isoelectric focusing of K562 and placental extracts. a) Detection with anti-6' PGM polyclonal antibodies. b) Detection with anti-10' PGM polyclonal antibodies. Detection with pre-immune serum is shown in Figure 3.5.

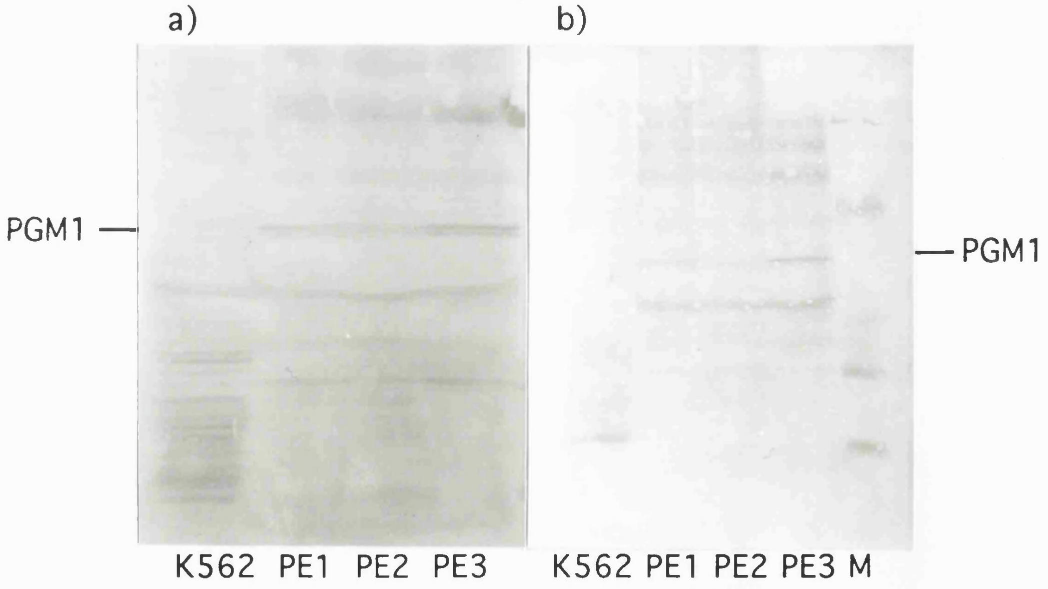


Figure 3.10 Determination of immunoreactivity of anti-human PGM1 polyclonal antibodies by immunoblot detection following SDS-PAGE of K562 and placental extracts. a) Detection with anti-6' PGM polyclonal antibodies. b) Detection with anti-10' PGM polyclonal antibodies. Detection with pre-immune serum is shown in Figure 3.6. M = protein molecular weight marker.

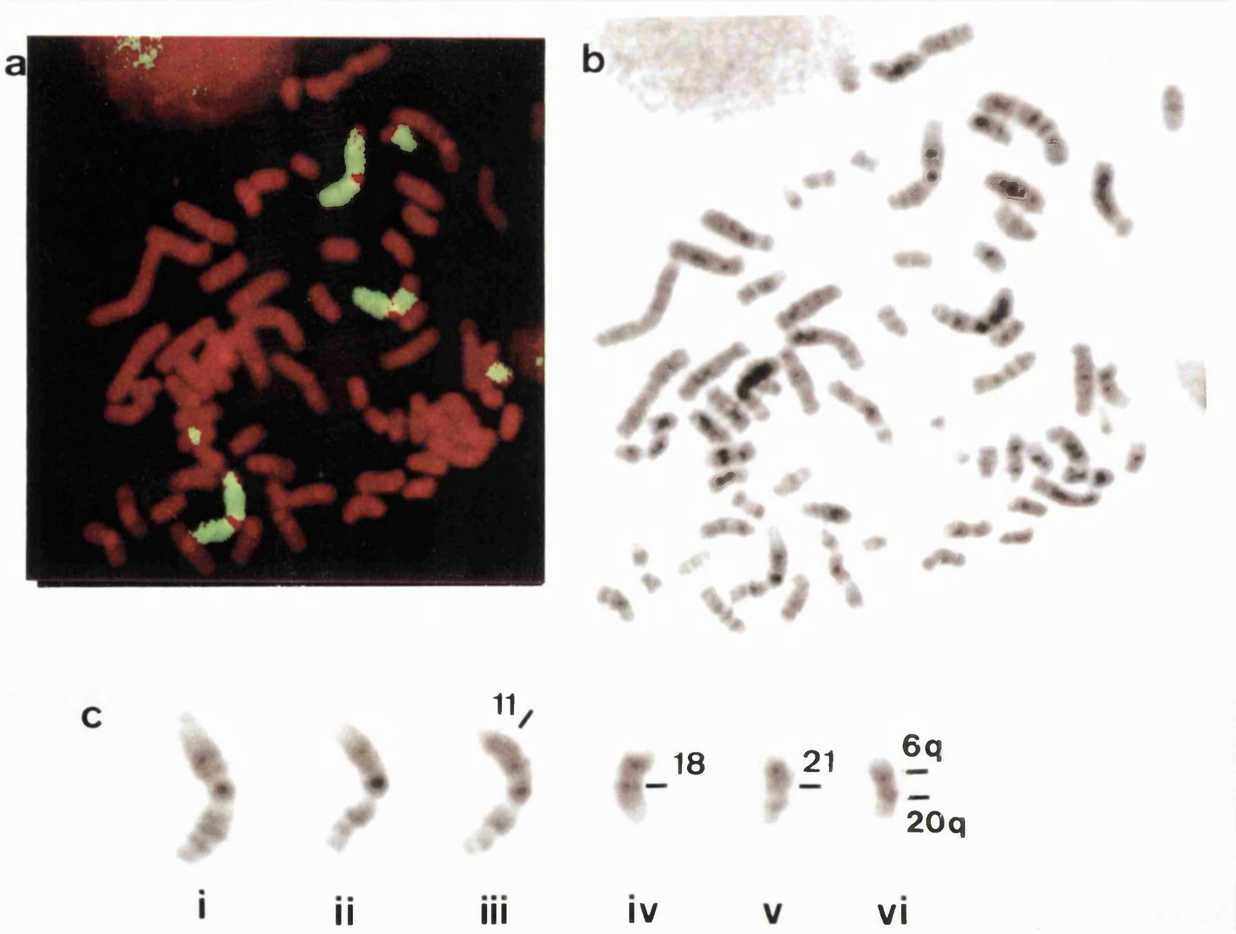


Figure 3.11 Cytogenetic analysis of the K562 cell line. a) FISH using a chromosome 1 specific paint. b) G-banding of the same cell. c) The chromosomes identified by the chromosome 1 specific paint: i) and ii) normal chromosomes 1, iii) der (1) t (1;11)(p32;q24), iv) der (18) t (1;18)(p32;q23), v) der (21) t (1;21)(q31;p13), vi) der (1) t (1;6;20)(p21q12; q25;q11).

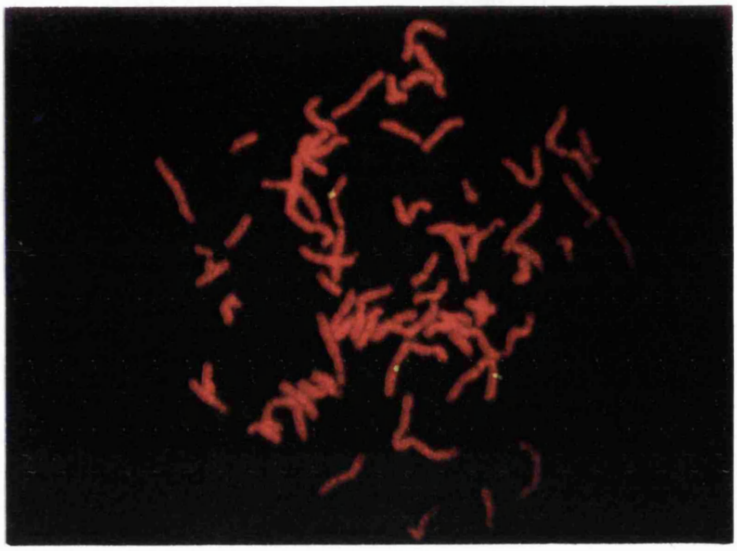


Figure 3.11d Fluorescence *in-situ* hybridization of a K562 cell using the LoHPGM1 cosmid as probe. Three signals are evident, mapping to 1p31.

3.2.2 SOUTHERN BLOT ANALYSIS

The structure of the three *PGM1* genes in K562 was investigated by Southern blot analysis to determine if the enzyme deficiency may have resulted from rearrangements of these genes; perhaps involving non-reciprocal crossovers between the two recombinogenic regions. Genomic DNA was digested with *EcoRI* and *MspI*. The restriction fragment lengths identified by hybridization with the HPGM1 probe were identical in K562 and control leucocyte DNA samples (Figure 3.12). The bands obtained with DNA digested with *TaqI* did vary within the four samples (Figure 3.13). This reflects the two diallelic RFLPs in the PGM1 gene involving *TaqI* sites, A and B (Hollyoake et al, 1992). K562 was identified as homozygous at each of the two sites, with the phenotype A2-B1. The autoradiography data for K562 digested with *TaqI* and *EcoRI* is shown in figure 3.14. The variation in the intensity of the hybridization signals from K562 is equivalent to that observed in the control leucocyte DNA. This suggests that all three copies of the *PGM1* gene contain all of the exons.

In conclusion, these results indicate that no gross rearrrangements of the exons encoding PGM1 have occurred at any of the three gene loci.

3.3 CHARACTERIZATION OF PGM1 mRNA

The RNA studies were carried out by reverse transcriptase (RT) PCR on three control cell lines of normal PGM phenotypes and K562 (Figure 3.15). Six pairs of primers spanning the entire coding region of *PGM1*, from nt 18 to 1935, were used. The primers were designed to flank at least one intron to ensure unambiguous identification of cDNA amplification (Figure 3.16). To check the quality of K562 RNA, control cDNA primers for the amplification of 6-phosphogluconate dehydrogenase (*PGD*) were used. In this case, no significant difference was seen between K562 and the controls.

For each pair of primers, the RT-PCR products from K562 RNA were of the expected size, but very faint compared to those of the control cell lines (Figure 3.17). In each case, an equivalent amount of PCR product was produced, indicating the presence of full length transcripts. The identity of these products as *PGM1* was confirmed by hybridization with ^{32}P labelled HPGM1. The intensity of the bands was estimated to be eight fold less in K562, compared with the control samples (Figure 3.18). Therefore it appears that the characteristic PGM1 enzyme deficiency of the K562 cell line is associated with very low levels of the PGM1 mRNA transcript.

Figure 3.12 Restriction fragment lengths from K562 and control leucocyte DNA digested with *EcoRI* and *MspI*

Enzyme	*Eco*RI		*Msp*I	
DNA Sample	K562	Control DNA	K562	Control DNA
Restriction	9.3	9.3	6.1	6.1
Fragment	6.3	6.3	3.3	3.3
Lengths	4.9	4.9	2.2	2.2
(kb)	4.2	4.2	1.3	1.3
	2.8	2.8		

Figure 3.13 Restriction fragment lengths from K562 and control leucocyte DNA digested with *TaqI*

Bands	Constant	Polymorphic			
DNA	All DNAs	K562	N7	N9	N18
		8.4	8.4	8.4	8.4
Restriction					5.7
Fragment	5.2				
Lengths		4.3	4.3	4.3	
(kb)	3.0				
	2.7				
			2.6		2.6
	1.9				
	1.7				
Phenotype		A2-B1	A2-B1B2	A2-B1	A1-B1B2

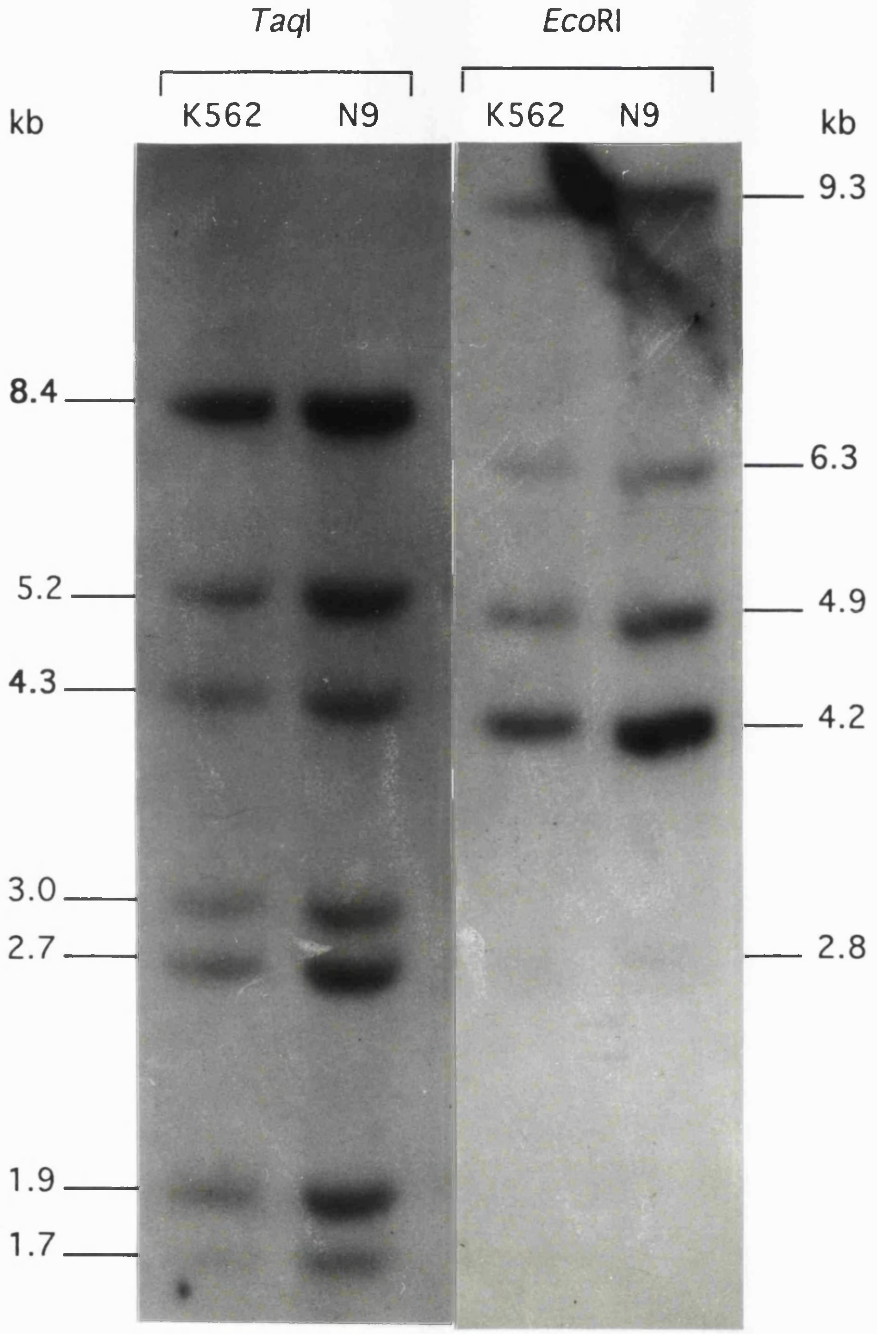


Figure 3.14 Autoradiography results from Southern blot analysis of K562 and control leucocyte DNA digested with *Taq*I and *Eco*RI and hybridized with the HPGM1 probe. The *Taq*I polymorphic restriction fragments are shown in bold.

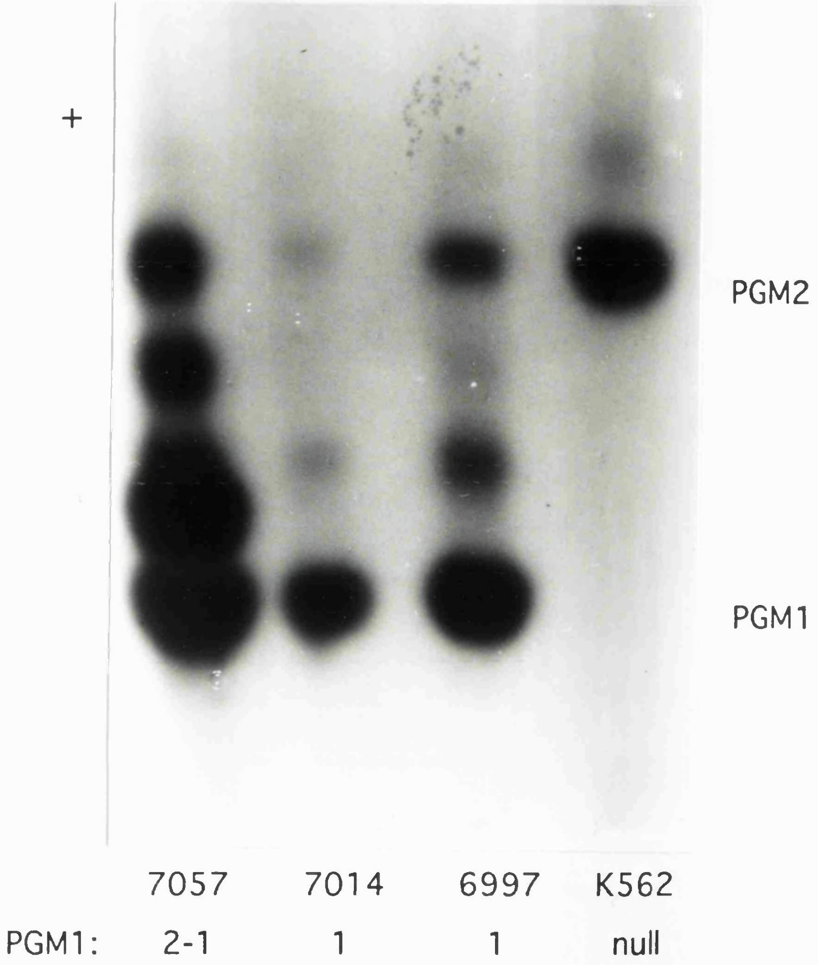
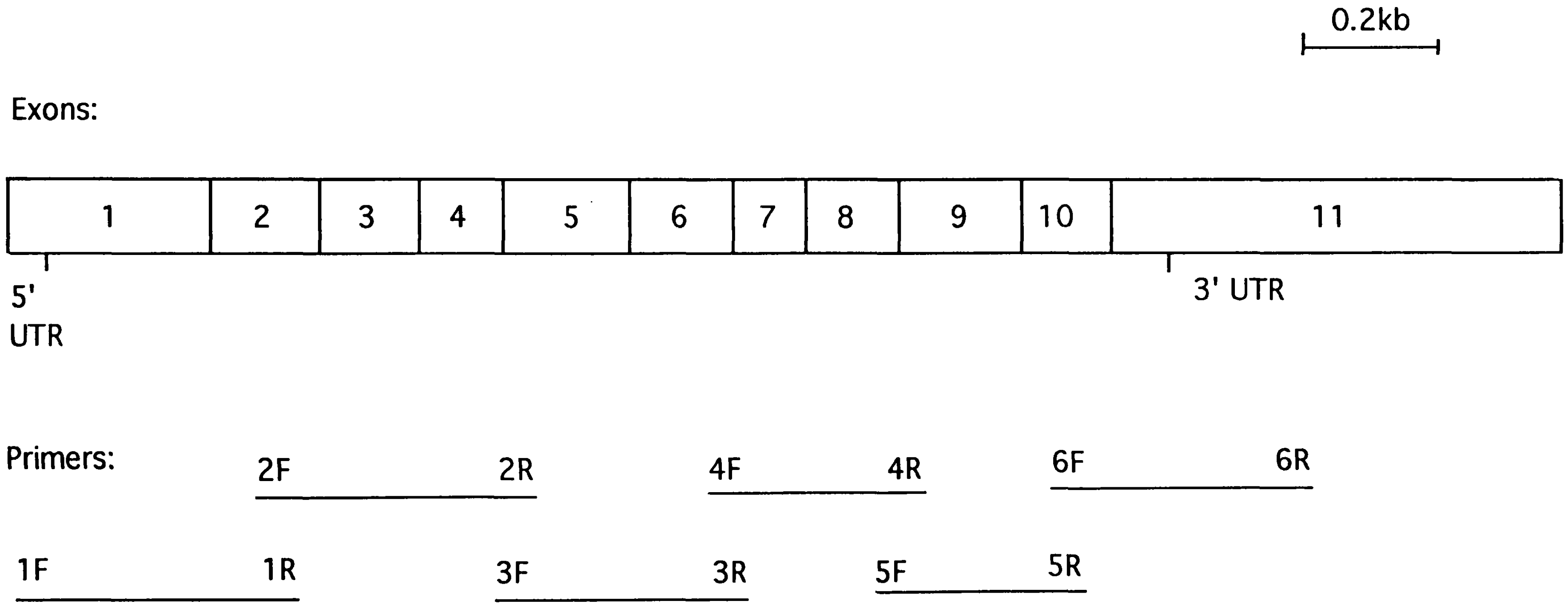


Figure 3.15 Detection of PGM isozymes in control lymphoblastoid cell lines and K562 by enzyme activity staining following starch gel electrophoresis.

Figure 3.16 Location of cDNA primer pairs spanning the entire coding region of *PGM1*.



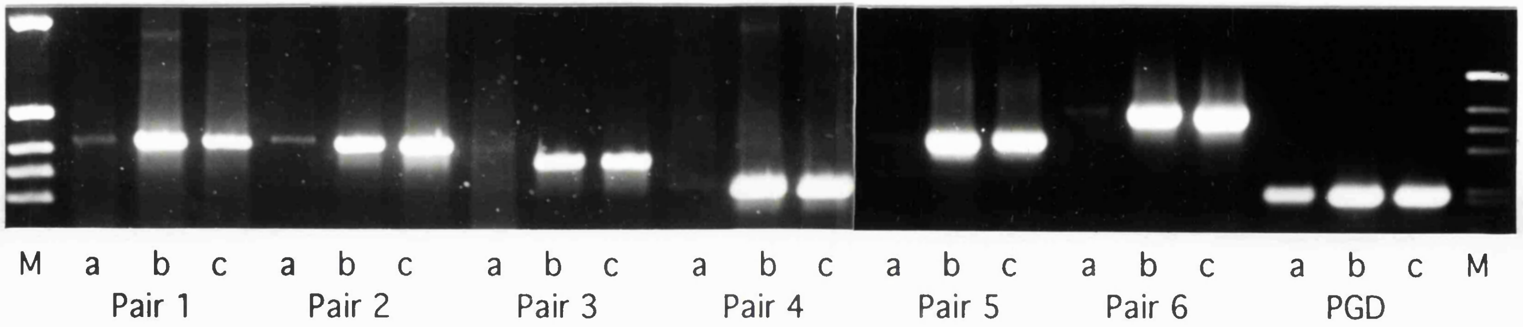


Figure 3.17 RT-PCR products from K562 and control lymphoblastoid cell lines amplified by the PGM1 cDNA primers and control PGD cDNA primers. a = K562; b = 6997; c = 7014; M = molecular weight size marker.

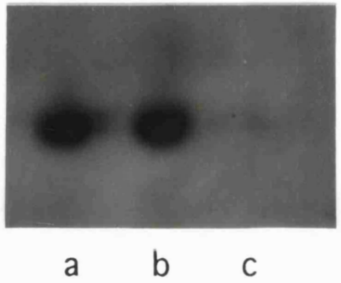


Figure 3.18 Autoradiography results of Pair 6 RT-PCR products from K562 and control lymphoblastoid cell lines hybridized with HPGM1. a = K562; b = 6997; c = 7014

3.4 DETERMINATION OF THE PGM1 PHENOTYPE IN K562

3.4.1 SSCP ANALYSIS

The single base changes which underlie the protein alleles can be detected by SSCP analysis of PCR products amplified from exons 4 and 8. The exon 4 PCR products encompassing the site encoding the 1/2 protein allele(s), show a faster migrating band if they carry the _PGM1_*1 allele, and a slower migrating band if they carry the _PGM1_*2. K562 clearly shows the two bands, indicating it is heterozygous (Figure 3.19a, lane 5). The exon 8 PCR products, encompassing the site encoding the +/- protein allele(s), are a little more difficult to distinguish. However, the faster migrating band corresponds to the _PGM1_*- allele, and the slower band to the _PGM1_*+ allele. K562 appears to be homozygous for the _PGM1_*+ allele (Figure 3.19b, lane 5). The deduced _PGM1_ phenotype for K562, therefore, is 2+1+.

3.4.2 RESTRICTION ENZYME ANALYSIS

Confirmation of the SSCP results were obtained by restriction enzyme analysis, as the base changes which underlie the PGM1 protein polymorphism lead to changes in restriction endonuclease recognition sites. Exon 4 PCR products from K562 and white blood cell DNA controls were digested with _Bgl_II. This enzyme cleaves the _PGM1_*2 allele (AGATC**T**$^{\mathbf{723}}$), but not the _PGM1_*1(AGATC**C**$^{\mathbf{723}}$). Figure 3.20a shows that K562 possesses both alleles, with the heterozygote restriction pattern clearly evident (lane 5). The reciprocal digest with _AlwI_, which cleaves the _PGM1_*1 allele ($^{\mathbf{723}}$**G**GATCN$_4$), but not the _PGM1_*2 ($^{\mathbf{723}}$**A**GATCN$_4$), supports this data (Figure 3.20b, lane 1). The exon 8 PCR products were digested with _Nla_III, which cuts the _PGM1_*- allele ($^{\mathbf{1320}}$**C**ATG) but not the _PGM1_*+ ($^{\mathbf{1320}}$**T**ATG). K562 is clearly homozygous for the _PGM1_*+ allele (Figure 3.20c, lane 1).

3.5 ANALYSIS OF PGM1 ALLELES K562 mRNA

We have determined that the K562 PGM1 phenotype is 2+1+ and that there are three intact copies of _PGM1_ gene. The heterozygosity exhibited at the exon 4 polymorphic site allowed us to investigate whether the third chromosome carried the _PGM1_*1 or the _PGM1_*2 allele, to determine if the genotype of K562 was 2+2+1+ or 2+1+1+.

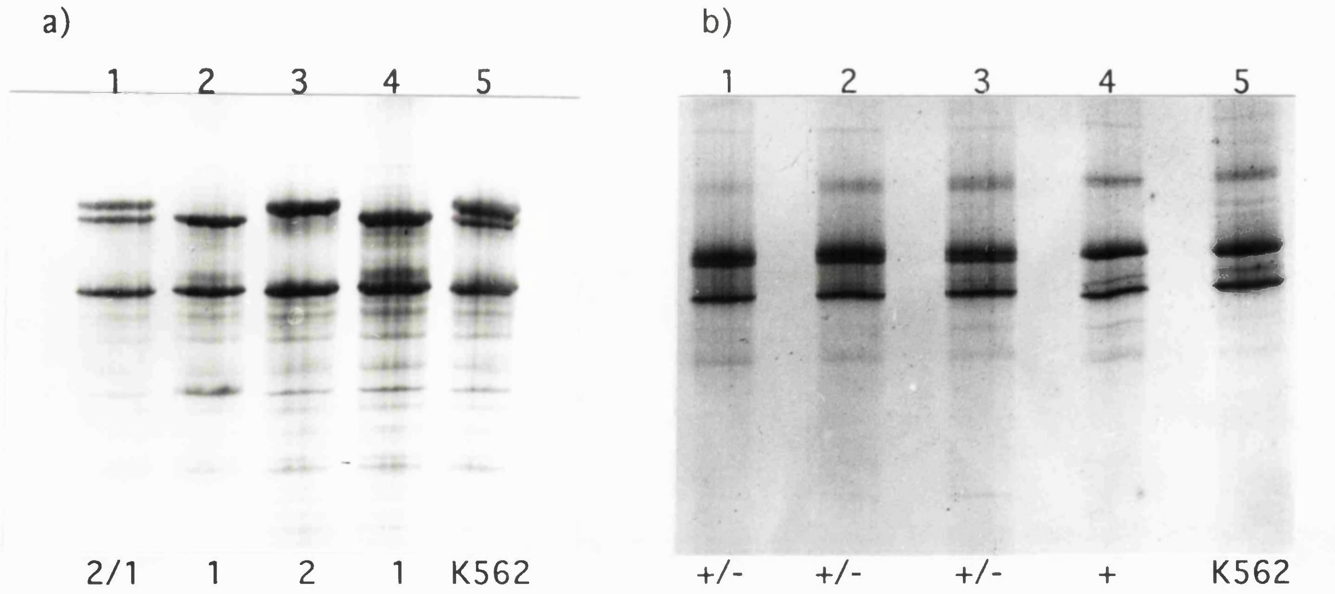


Figure 3.19 Silver stained SSCP gels of a) exon 4 PCR products demonstrating the 2/1 polymorphism and b) exon 8 PCR products demonstrating the +/- polymorphism of PGM1 in K562 and control leucocyte DNA samples. Lane 1 control DNA PGM1*2+1-; lane 2 control DNA PGM1*1+1-; lane 3 control DNA PGM1*2+2-; lane 4 control DNA PGM1*1+; lane 5 K562.

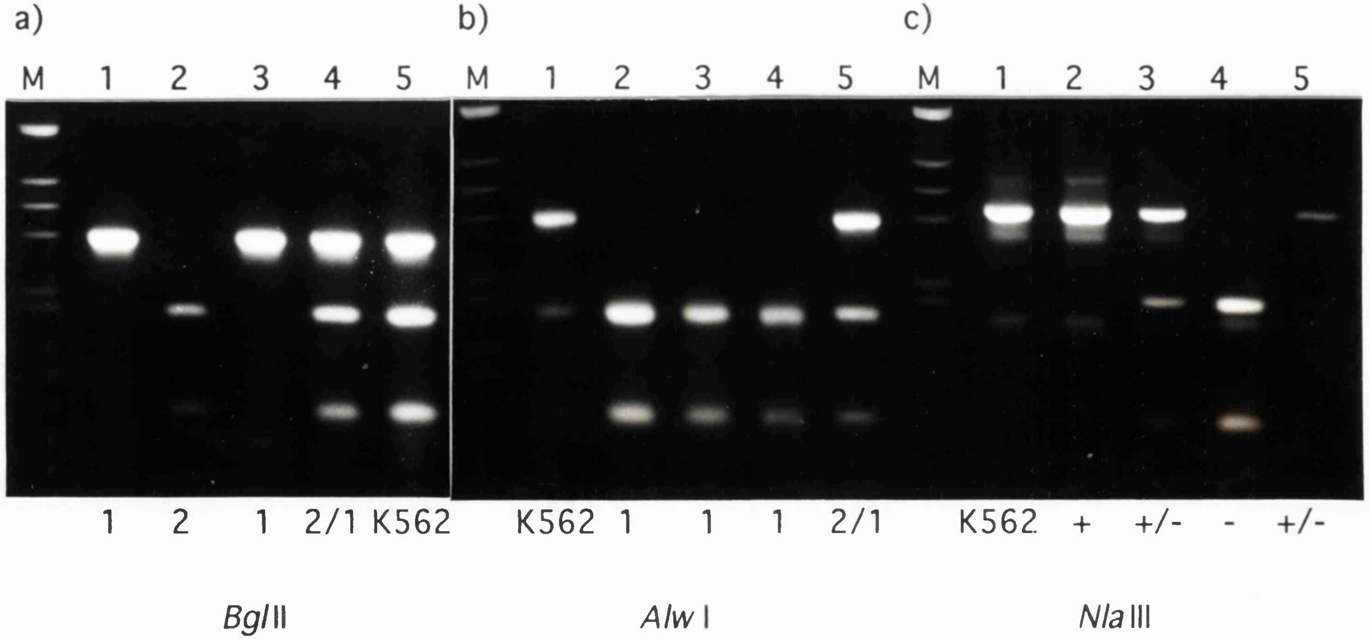


Figure 3.20 RFLP analysis of PCR products from K562 and control leucocyte DNA samples demonstrating the PGM1 polymorphism. a) Exon 4 PCR products digested with *Bgl* II, which cleaves the 2 allele but not the 1; lane 1 control DNA PGM1*1+; lane 2 control DNA PGM1*2+2-; lane 3 control DNA PGM1*1+1-; lane 4 control DNA PGM1*2+1-; lane 5 K562 DNA. b) Exon 4 PCR products digested with *Alw* I, which cleaves the 1 allele but not the 2; lane 1 K562 DNA; lane 2 control DNA PGM1*1+; lane 3 control DNA PGM1*1+1-; lane 4 control DNA PGM1*1-; lane 5 control DNA PGM1*2+2-. c) Exon 8 PCR products digested with NlaIII, which cleaves the - allele but not the +; samples as for b). M = molecular weight size marker.

RT-PCR products encompassing exon 4 were reamplified and digested with *BglI*. The digest shows that the majority of the 419bp fragment from K562 is digested to give the expected 346bp band characteristic of the *PGM1**2 allele (Figure 3.21). However, there is still a faint band remaining which is uncut. In the reciprocal digest, with *AlwI*, a very faint band was evident in K562 corresponding to the *PGM1**1 allele. This data suggests that the genotype of K562 is actually 2+2+1+. However, the level of *PGM1**2 transcripts compared to *PGM1**1 appears to be far greater than 2:1 expected from the genotype proposed.

3.6 SUMMARY

i) No PGM1 activity was detected in K562, either after starch gel electrophoresis or IEF, verifying the observation reported by Povey et al, (1980). Immunoblot detection using anti-PGM polyclonal antibodies showed that the PGM1 enzyme deficiency is associated with an absence of protein. Further, two anti-human PGM1 polyclonal antibodies did not show immunoreactivity with PGM2 or PGM3. Therefore they appear to be unsuitable tools for screening cDNA expression libraries for these members of the PGM gene family.

ii) Cytogenetic and FISH analysis suggested three intact copies of the *PGM1* gene occur in K562; two on the normal and one on the derivative chromosome 1s. In each case, they localized to 1p31. The Southern blot analysis supported this data; genomic DNA digested with three restriction endonucleases failed to detect any abnormal restriction fragments.

iii) Analysis of the mRNA indicates a low level of full length PGM1 transcripts. This suggests that the molecular basis of the PGM1 deficiency in K562 is abnormal regulation. However, it is not possible to distinguish from this data if this is due to a mutation affecting transcription or mRNA stability.

iv) The deduced protein phenotype of K562 is 2+1+ determined by both SSCP and restriction enzyme analysis. Analysis of mRNA indicated a bias in the level of the *PGM1**2 allele transcript, suggesting the genotype is more accurately designated 2+2+1+.

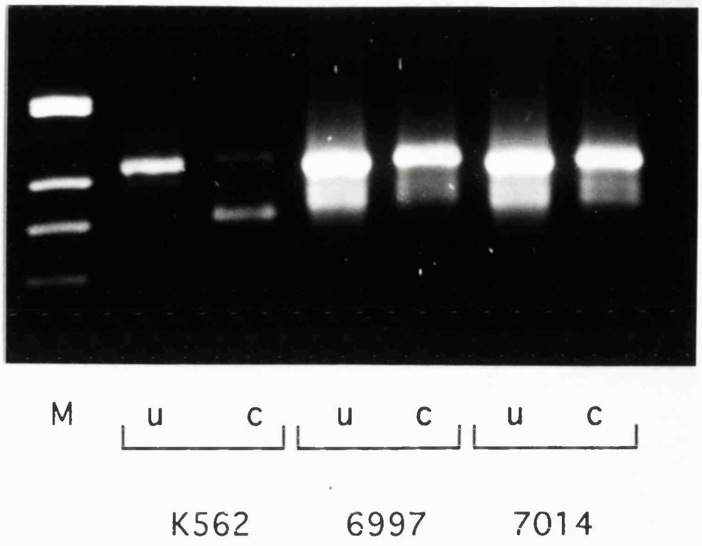


Figure 3.21 Digestion of RT-PCR products from K562 and control lymphoblastoid cell lines with *Bgl*II. The level of the *PGM1**2 allele PCR product, digested by *Bgl*II, appears to be more than double the *PGM1**1 allele PCR product, which is undigested .

3.7 CONCLUSIONS

This investigation was principally carrried out in order to assess the usefulness of K562 as a resource for cloning other members of the PGM gene family. Although three copies of the structural gene are present, and do not appear to be rearranged, the eight fold reduction in the levels of PGM1 mRNA transcript suggest that K562 will be a very useful resource. cDNA primed from K562 is likely to contain a lower pool of *PGM1* cDNA than most other sources. Since the primers used for low stringency and degenerate PCR (considered in Chapter Four) are based upon conserved regions of the PGM1 protein, the ratio of amplified *PGM1* to PGM-related sequences is likely to be less.

The molecular basis for the abnormal regulation of the K562 PGM1 transcripts has not been established. However it is unlikely that it is simply attributable to the presence of three copies of the gene. The K562 cell line is known to be triploid in nature and the karyotype shows considerable chromosomal abnormalities (Fox et al, 1996). During these studies, analysis of *PGD*, which is also localized to the short arm of chromosome 1 close to *PGM1* and is therefore probably present in three copies, showed normal levels of PGD mRNA and activity in K562 compared to other cell lines. Thus deficiency of PGM1 protein appears to be a specific and unique feature of K562.

The mRNA studies indicated a bias towards the expression of the *PGM1**2 allele rather than *PGM1**1. This may be taken as evidence that the genotype of K562 was 2+2+1+. However, the amount of PCR product amplified from the *PGM1**2 allele compared to the *PGM1**1 allele was much greater than the 2:1 ratio which would be expected. Therefore, the molecular basis of the PGM1 deficiency in K562 may be due to a trans-acting element affecting all three *PGM1* genes, yet its effect may not be equivalent on each of the genes. Thus, the disproportionate expression of alleles may represent a greater inhibitory effect on the *PGM1**1 allele than the *PGM1**2, such that a genotype of either 2+2+1+ or 2+1+1+ shows greater levels of the *PGM1**2 allele. However, retrospective analysis of the ethidium bromide stained gels, demonstrating the *PGM1* polymorphism, appears to also show a greater level of *PGM1**2 allele product in comparison with the control DNA heterozygote. Therefore, the PGM1 genotype for K562 is suggested to be 2+2+1+.

CHAPTER FOUR:

PCR-BASED SEARCH FOR MEMBERS OF THE PGM GENE FAMILY

Phosphoglucomutase (PGM) and phosphomannomutase (PMM) proteins from numerous species of eukaryotes and prokaryotes show a high level of amino acid conservation at regions essential for catalytic activity (Section 1.3.2; Figure 1.13). Two PCR-based strategies utilizing primers designed to these conserved regions were investigated to identify other members of the PGM gene family. The first approach was low stringency PCR, which identifies closely related sequences, whereas the second approach of degenerate primer PCR allows for a greater level of divergence between the sequences. Both strategies required the use of cDNA as template for the PCR.

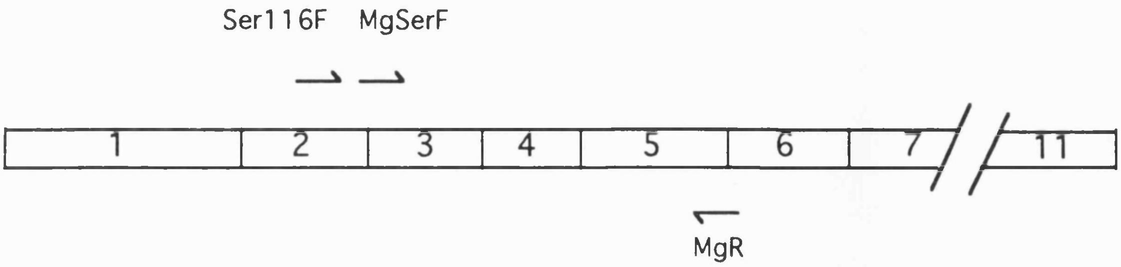
4.1 LOW STRINGENCY PCR

The principle of low stringency PCR is based upon using a much lower annealing temperature than in a standard PCR. This allows for the presence of mismatches between the bases of the primer and the template DNA such that amplification of nucleotide sequences showing similarity to the target sequence can occur. For example, using an annealing temperature of Tm-26^{o}C, β-globin specific primers were able to amplify a corresponding region in the δ-globin gene (Scharf et al, 1986). The authors used the same approach to amplify allelic variants in the HLA DQα locus.

The proposed approach for low stringency PCR with PGM primers was, following amplification, separation of the products by electrophoresis and transfer to Hybond N^{+} (Amersham) for analysis using HPGM1 as probe. The filters were to be washed at low stringency to compensate for the expected nucleotide divergence. Any bands of hybridization identified following autoradiography, would be investigated by cloning and nucleotide sequencing.

Primers were designed to regions of the PGM1 protein that are completely conserved at the nucleotide level between rabbit and human. Two forward primers were designed: Ser116F, sited in exon 2 over the active site of the protein and MgSerF, sited over the exon 2 and exon 3 boundary. The single reverse primer used with these primers was MgR, sited over the exon 5 and exon 6 boundary, covering the magnesium binding loop (Figure 4.1).

Figure 4.1 Location of low stringency RT-PCR primers



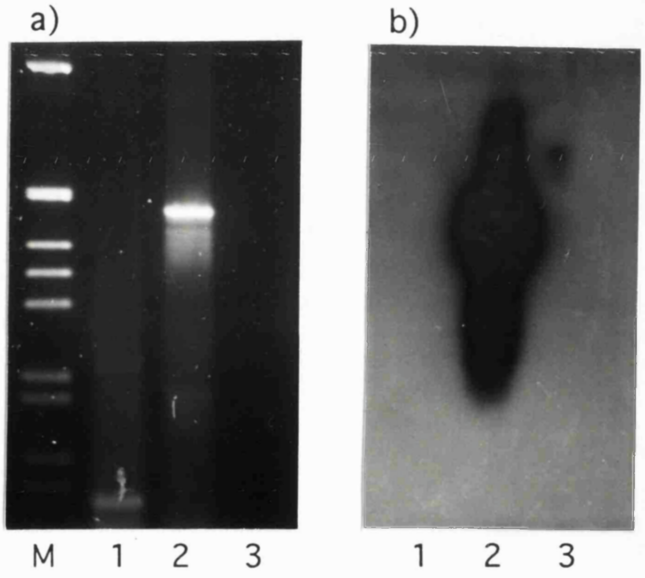


Figure 4.2 Low stringency PCR of K562 and control cell line 6997 using MgSerF and MgR primers. a) Ethidium bromide stained gel of PCR products. b) Southern blot analysis of PCR products probed with HPGM1 following 3 days autoradiography. Lane 1 K562; lane 2 6997; lane 3 dH_2O control; M = molecular weight size marker.

The K562 cell line which expresses low levels of PGM1 transcripts, but relatively high levels of PGM2 and PGM3 isozymes, was judged to be a useful resource for the RT-PCR experiments. In addition, RNA from two lymphoblastoid cell lines, 6997 and 7014, was used. These two cell lines express the three loci, PGM1, PGM2 and PGM3 (see figure 3.12).

4.1.1 OPTIMIZATION AND RESULTS OF LOW STRINGENCY RT-PCR

In the reverse transcription reaction, a mixture of random hexamer primers provided a better cDNA pool than either the MgR or an oligo dT primer. The conditions for amplification of the cDNA were investigated. A two step strategy, with a single cycle at 35°C annealing for 5mins followed by 35 cycles at Tm-30°C for 30secs produced a band of 475bp from the cell line 6997. This is the size expected from *PGM1* using the MgSerF and MgR primers. In K562, no corresponding sized band was evident, only one of 120bp (Figure 4.2a). This was thought to be too small to represent a PGM-related sequence in which the structure and function of the protein could be maintained. This was supported by Southern blot analysis; the HPGM1 probe hybridized strongly to the 475bp product from 6997, but no hybridization was evident with the 120bp product in K562 with low stringency washing after three days of autoradiography (Figure 4.2b). Additionally, if the sequence was PGM-related, following low stringency PCR of K562 with the Ser116F and MgR primer pair one would expect to observe a band of approximately 188bp, yet no products were obtained.

The PCR conditions were then altered to decrease the specificity of the PCR by using low annealing temperatures for all 30 cycles of the reaction. Annealing temperatures of 35°C, 40°C, 45°C, 50°C and 55°C were investigated, corresponding to Tm-30°C, Tm-25°C, Tm-20°C, Tm-15°C and Tm-10°C for the primer pair Ser116F and MgR. Times for each stage of the cycle were also increased: denaturing at 94°C for 1min, annealing for 2mins and elongation at 72°C for 2mins. In each case, the primers produced smears of DNA; no distinct bands were amplified from cDNA reverse transcribed from the lymphoblastoid cell lines 6997 and 7014 (Figure 4.3a). PCR products amplified with annealing temperatures of 50°C and 55°C only showed hybridization to the HPGM1 probe at 543bp, the expected size of amplification products from *PGM1* (Figure 4.3b).

Thus, low stringency PCR appears to only amplify *PGM1*; no closely related sequence of similar size was detected from K562 using the HPGM1 probe, nor are there any different sized PCR products with nucleotide similarity to *PGM1*

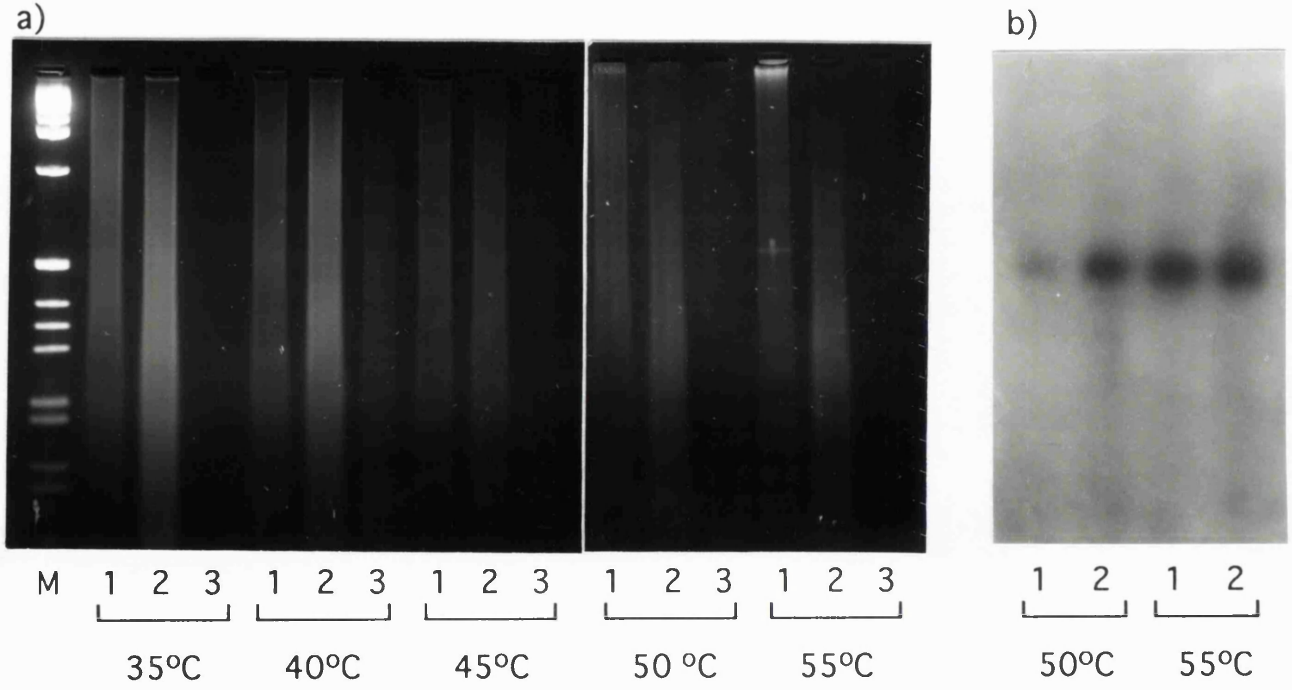


Figure 4.3 Low stringency PCR of control cell lines 6997 and 7014 using the primers Ser116F and MgR at Increasing annealing temperatures. a) Ethidium bromide stained gel of PCR products. b) Southern blot analysis of PCR products annealed at 50°C and 55°C with the HPGM1 probe. Lane1 6997; lane 2 7014; lane 3 dH_2O control; M = size marker

evident from the other two cell lines. This suggests that the divergence at the nucleotide level between PGM1 and the other PGM isozymes is too great for low stringency PCR to be effective in cloning the genes encoding PGM2 and PGM3.

4.2 DEGENERATE PRIMER PCR

Degenerate primer PCR involves the use of primers which allow for codon changes whilst conserving the amino acid sequence of recognized protein motifs. This strategy has been used to identify genes in which only a portion of the amino acid sequence is known. This is exemplified by an early report from Lee et al, in which degenerate oligonucleotide primers were used to amplify a porcine urate oxidase cDNA sequence (Lee et al, 1988). This was subsequently used as a probe to identify the full-length cDNA. Each of the primers had a redundancy of 32 fold, that is, the total number of different primer sequences in each of the primer mixes. As with the low stringency PCR, a low annealing temperature (eg Tm-25°C) was used during amplification. (Calculation of the Tm when using the degenerate primers is based upon the most AT-rich primer sequence - see Chapter 2, section 2.2.8.5.) Subsequent investigations showed that primers of a much higher degeneracy could be used equally successfully: primer mixes with 256 and 1024 fold degeneracy were used to amplify bovine cardiac muscle hexokinase cDNA sequences (Griffin et al, 1988).

The strategy has also become an established approach for the identification of both paralogous and orthologous genes. Degenerate primers designed to the catalytic domain of protein-tyrosine kinases have identified two novel sequences expressed in murine haemopoietic cells, which appear to be members of the protein-tyrosine kinase gene family (Wilks, 1989). More recently, conserved cellulase family-specific sequences were used to clone cellulase homologues from *Fusarium oxysporum* (Sheppard et al, 1994). The broad application of this approach is illustrated by the use of amplified degenerate primer PCR products as hybridization probes on Drosophila polytene chromosomes to determine the genomic position of kinesin-related proteins (Endow & Hatsumi, 1991).

4.2.1 DEGENERATE PRIMER PCR STRATEGY

The degenerate primers incorporated restriction endonuclease recognition sites at the 5' end to simplify cloning of the PCR products. However, an alternative approach in which the PCR products were cloned directly into a T-vector (pCRII) was ultimately employed. T-vector cloning is dependent on the property of most Taq polymerases of a non-template dependent activity which adds a single A nucleotide to the 3' end of the PCR product. Digestion of a plasmid vector with a restriction endonuclease which leaves blunt ends, such as *Sma*I, followed by the addition of a single T residue using Taq polymerase, allows the amplified product to be ligated into the vector without enzymatic modification (Marchuk et al, 1991).

The strategy for the identification of PGM-related sequences using degenerate primers is shown diagramatically in figure 4.4. Following degenerate primer PCR of template DNA (see section 4.2.2.2), the products were ligated into the pCRII vector (Figure 4.5), and used to transform *E.coli* INVαF' (TA Cloning Kit, Invitrogen). The cloning site of the pCRII vector lies in the *lacZ* gene, enabling blue/white selection of transformants. Disruption of *lacZ* by ligation of a PCR product produces a white colony when grown on agar plates in the presence of X-gal. Bacterial colonies containing recombinant plasmids were picked and small cultures grown overnight. Following preparation of plasmid DNA, the size of the ligated insert was determined by digestion with *Eco*RI. Recombinant plasmids containing PCR products of between 250bp and 750bp were identified and cultures set up to provide plasmid DNA for sequencing.

All nucleotide sequence obtained was compared with the Genbank, EMBL and Swissprot databases to determine if it was a known sequence, a sequence with similarity to a known sequence or a completely novel sequence. The completely novel sequences were analyzed for open reading frames (ORF) in the same frame as the primers. Newly identifed ORFs which showed similarity to PGM1 could then be mapped, and characterized further by Northern blot analysis and isolation of a cDNA clone.

4.2.2 PGM AND PMM SEQUENCE BASED DEGENERATE PRIMER PCR

The investigation of degenerate primer PCR as an approach for the identification of other members of the PGM gene family centred upon the use of primers designed to amplify cDNA following the comparison of numerous PGM and PMM protein sequences.

Figure 4.4 Strategy for the identification of PGM-related sequences using degenerate primer PCR.

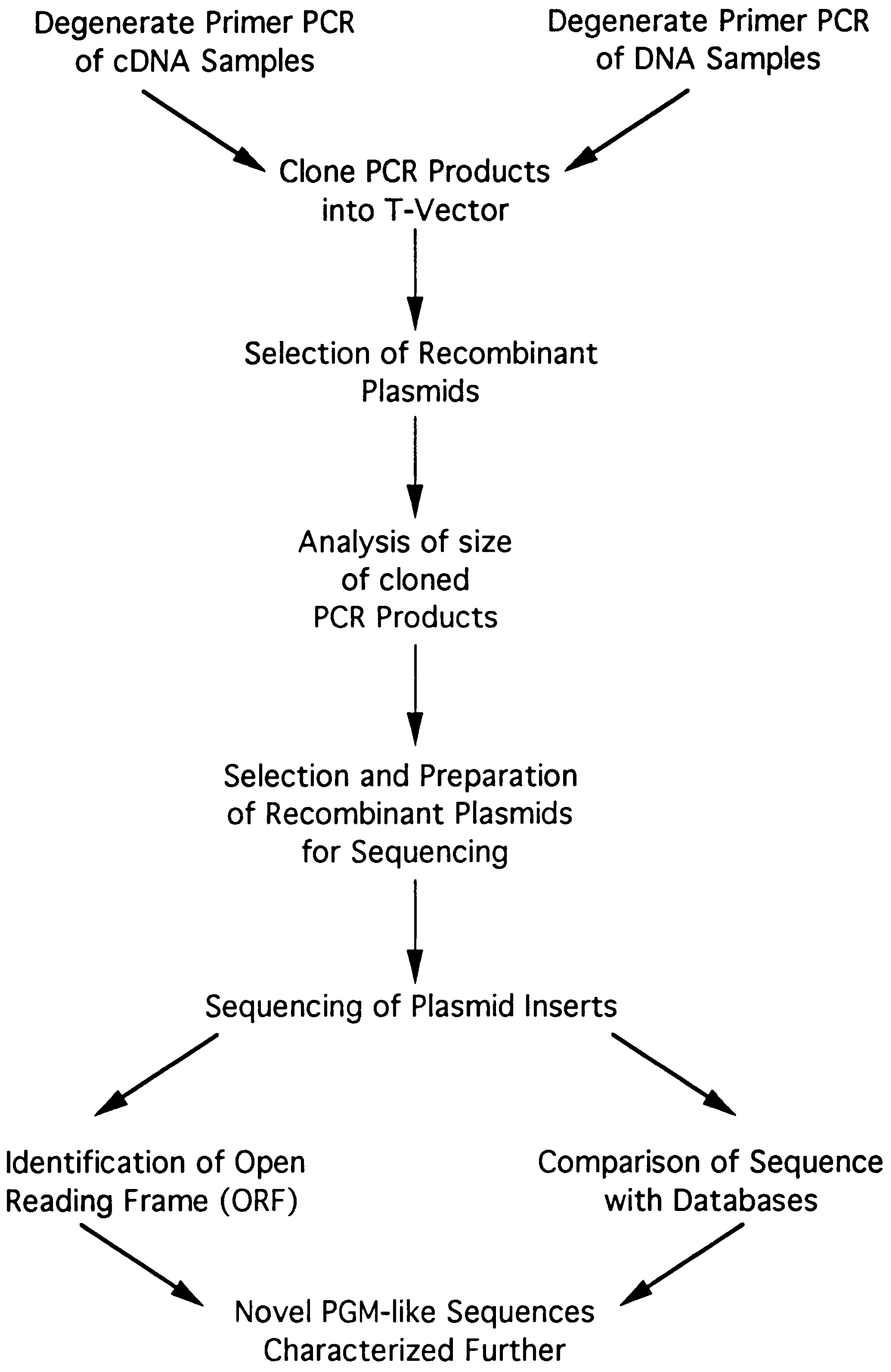
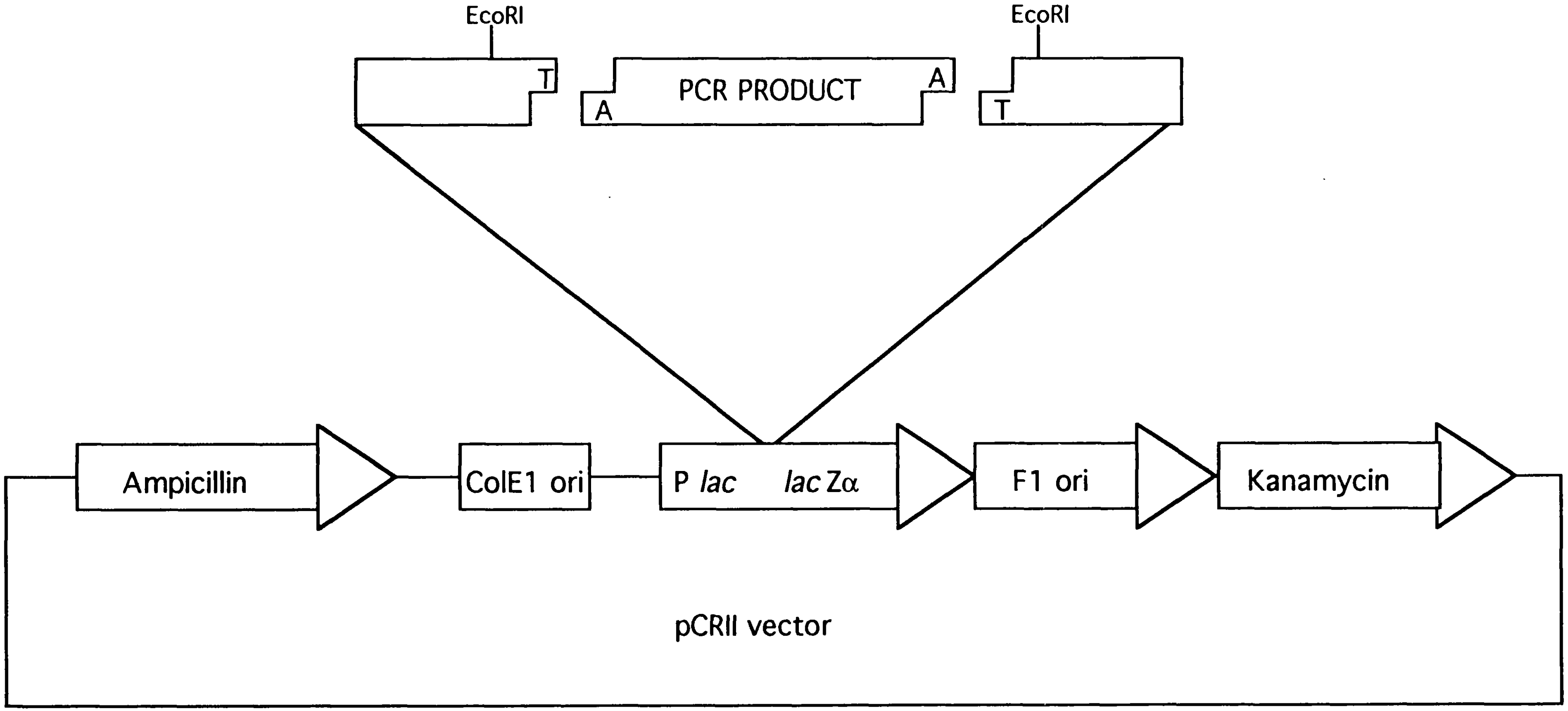

Degenerate Primer PCR of cDNA Samples
Degenerate Primer PCR of DNA Samples
Clone PCR Products into T-Vector
Selection of Recombinant Plasmids
Analysis of size of cloned PCR Products
Selection and Preparation of Recombinant Plasmids for Sequencing
Sequencing of Plasmid Inserts
Identification of Open Reading Frame (ORF)
Comparison of Sequence with Databases
Novel PGM-like Sequences Characterized Further

Figure 4.5 Diagram of the pCRII T-vector



4.2.2.1 The Degenerate Primers

The degenerate primers were based on the highly conserved active site and magnesium binding loop protein motifs, TASHNP and AFDGDGDR, found in mammalian PGM1. Redundancy was kept to a minimum by positioning the primers as shown in figure 4.6. DegSer116F is designed to contain all possible nucleotide primers encoding the amino acids ASHNP, whilst DegMgR contains all nucleotide sequences encoding AFDGDG. In addition to amplifying sequences encoding these amino acids, the primers will also anneal to sequences encoding amino acids in the other five frames. However, DNA sequencing will determine if the ORF is in the same frame as the primer. The restriction endonuclease recognition sites at the 5' end of the primers appeared not to interfere with the performance of the initial round of PCR, and served to raise the melting temperature of the primers in subsequent rounds of amplification.

The degenerate primers were subsequently redesigned to take into account the newly published information on PGM and PMM sequences. DegSer116F2 covered the amino acids (G/A)SHNP and DegMgR2 the amino acids GD(G/F/A)DR. These primers differed from the first pair by showing redundancy in the amino acid sequences as well as in the nucleotides.

In an attempt to improve the specificity of the degenerate primer PCR strategy, nested primers were designed, to use in combination with DegSer116F2 and DegMgR2. The relative positions of these primers, with respect to PGM1, is shown in figure 4.7. DegSer116F2 was used as an outer forward primer in combination with two degenerate reverse primers, DegPGMR and DegPMMR. DegPGMR was sited over the putative glucose binding loop GEESFG which is highly conserved in phosphoglucomutase proteins. DegPMMR was sited over a corresponding region in the phosphomannomutases where the amino acid sequence GEMSAG is conserved. The nested PCR used DegMgR2 as the reverse primer, and two degenerate forward primers DegPGMF and DegPMMF. DegPGMF was based on a conserved region, D(N/F)GIK, nine residues downstream of the active site of the phosphoglucomutases. A corresponding region in the phosphomannomutases, DYNGMK, was the site of the DegPMMF primer. The conservation of these sites was established by multiple sequence alignments of all the available PGM and PMM sequences.

A summary of the primer sequences and annealing temperatures for the PGM and PMM sequence based degenerate primers is shown in figure 4.8.

Figure 4.6 Degenerate primers DegSer116F and DegMgR based on the active site and magnesium binding loop protein motifs.

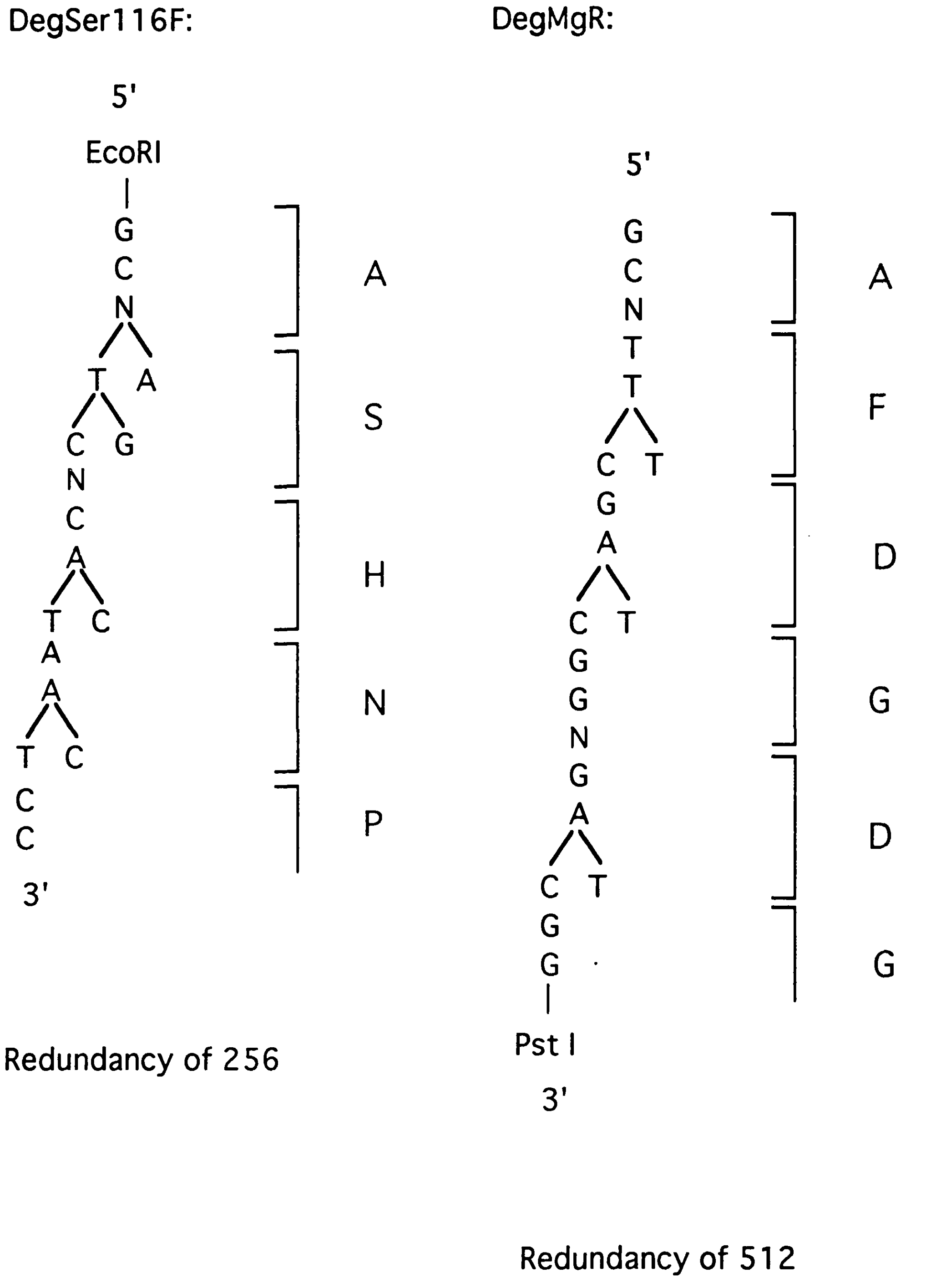


Figure 4.7 Location of the PGM cDNA degenerate primers

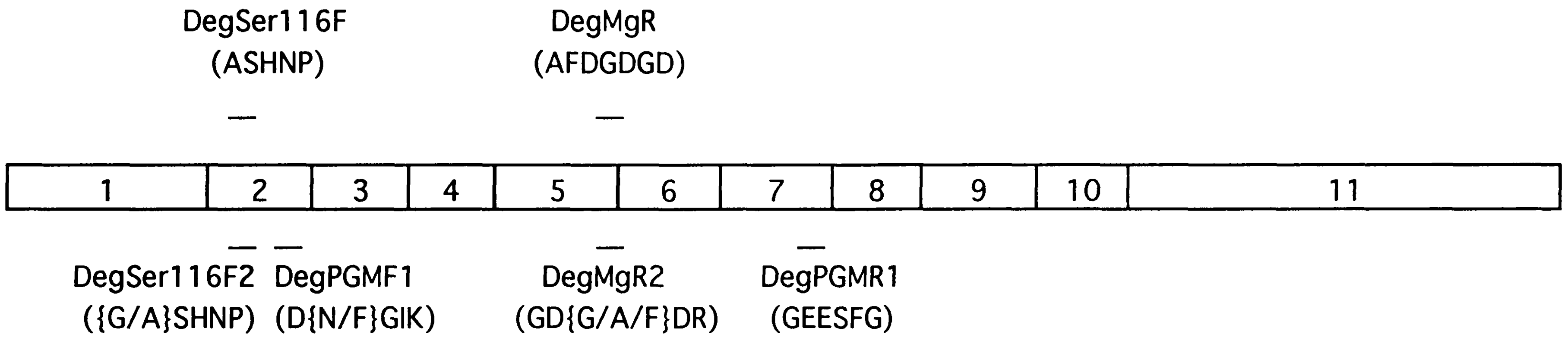


Figure 4.8 Nucleotide sequences of PGM cDNA degenerate primers and annealing temperatures.

Forward Primer	Reverse Primer	ANNEALING TEMPERATURE
DegSer116F 5' GGAATTCCGCNWSNCAYAAYCC 3'	DegMgR 5' CCTGCAGGNCCRTCNCCRTCRAANGC 3'	48°C / Tm-10°C
DegSer116F2 5' GAAGCTTCGSNWSNAYAAYCC 3'	DegMgR2 5' CTCTAGAGCGRTCNVMRTCNCC 3'	55°C / Tm-3°C
DegSer116F2	DegPGMR 5' CTCTAGAGCCRAANSWYTCYTCNCC 3' DegPMMR 5' CTCTAGAGGCNSYCATYTCNCC 3'	55°C / Tm-3°C
DegPGMF 5' GAAGCTTCGAYWWYGGNATYAA 3' DegPMMF 5' GAAGCTTCGAYTAYAAYGGNATGAA 3'	DegMgR2	53°C / Tm°C

4.2.2.2 Selection of Template DNA Samples

RNA was obtained from several human B cell lines, JG, TC and ED, in addition to K562. In each case, the reverse transcription stage of the RT-PCR was primed using random hexamer primers.

In a pilot study to investigate the flexibility of degenerate primers for cloning *PGM1* homologues from other species, pA^+ RNA from rat skeletal muscle provided a further source of cDNA, along with DNA samples from *E.coli*, (prepared by repeated freeze/thawing), and *Trypanosome cruzi*, (provided by the London School of Hygiene and Tropical Medicine (LSHTM)). Rat *PGM1* and numerous PGM-related sequences from different serotype groups of *E.coli* have been cloned, and therefore provide suitable positive controls for the degenerate primer strategy. PGM is a well characterized isozyme marker in *T.cruzi*, although the cDNA has not been isolated.

An additional source of template DNA for degenerate primer PCR was two cDNA libraries, K562 (H5) and placenta, full term (H9), which were obtained from the Human Genome Mapping Project Resource Centre.

4.2.2.3 Optimization of PCR Conditions

Experiments, based on the method reported by Lee and Caskey, (1990), were conducted on control human cell lines to determine the optimum PCR conditions for the DegSer116F and DegMgR primers. The recommended times of the 35 cycle PCR programme were 30secs for denaturing at 95°C, 30secs for annealing and 60secs for elongation at 72°C. To reduce the length of the complete PCR, the elongation time was reduced to 45secs, with no difference observed in the PCR products. An annealing temperature of 48°C (Tm-10°C of the most AT rich primer) produced the strongest distinct bands within the background smear of DNA when using the DegSer116F and DegMgR primers. However, for the subsequent degenerate primers, Tm-3°C or Tm°C of the most AT rich primer was used (figure 4.8)

The concentration of primer in the reaction mix was increased 200 fold in comparison with a normal PCR, 10nmoles instead of 50pmoles, to compensate for the number of primer sequences present. This led to smears of DNA appearing in the no-DNA controls in some cases. Hybridization of the PCR products with the HPGM1 probe, however, identified *PGM1* only in those samples with JG and ED cDNA added and not in the no-DNA control, suggesting the smears were due to primer/primer interactions.

Nested degenerate primer PCR was performed using the same PCR programme: denaturing at 95°C for 30secs, annealing for 30 secs and elongating at 72°C for 45s, but each round of PCR consisted of 25 cycles. The concentration of primers was also reduced in the nested PCR reactions. 100pmoles of forward primer and 100 pmoles of reverse primer were used (i.e. 50pmoles of each primer when the two degenerate primers DegPGMF and DegPMMF or DegPGMR and DegPMMR were used).

4.2.2.4 Results of Degenerate Primer PCR

Amplification of the cDNA samples produced a copious amount of low molecular weight DNA in all samples with some faint specific bands in JG and rat (Figure 4.9). PCR results from *E.coli* gave a strong 475bp band, whilst the libraries showed only a weak band of approximately 420bp. The subsequent transformation results are shown in figure 4.10. In total, of 114 recombinant plasmids analysed, only 25 contained insert sequences of between 250bp and 750bp. Of these, 9 were vector rearrangements. Of the remainder, in addition to sequences primed by the same primer at each end, and sequences with no ORF, a few interesting sequences were cloned. It appears that the specific bands of 520bp, 475bp and 450bp amplified from rat, *E.coli* and JG respectively are PGM-related sequences. The two faint bands amplified from the cDNA libraries appear to be *E.coli* chromosomal DNA.

RAT *PGM1*: The clone obtained from the rat PCR product was sequenced and identified as rat *PGM1.* This is evidence that the degenerate primer PCR strategy has the capability, given sufficient conservation of nucleotides, to identity homologous sequences from other species. Five nucleotide differences between the sequence of the reverse primer and the published sequence were observed. Despite the primer not being identical to the target sequence, it was clearly able to anneal sufficiently to drive the first round of PCR. The rat RT-PCR *PGM1* sequence also differs from the published sequence at nucleotide 749, showing an A-G transition. This leads to a change in codon 236 of ATC to GTC resulting in replacement of Ile by Val. It is not clear, however, whether this is a polymorphic site in the rat or an artefact from the PCR, cloning or sequencing techniques.

Figure 4.9 PCR results from experiments using degenerate PGM cDNA primers

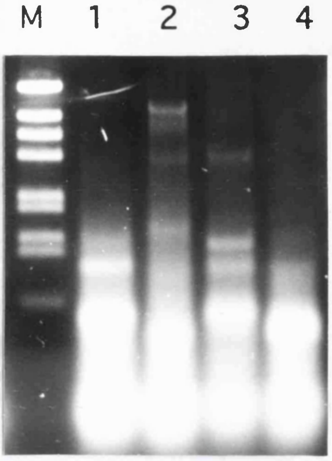
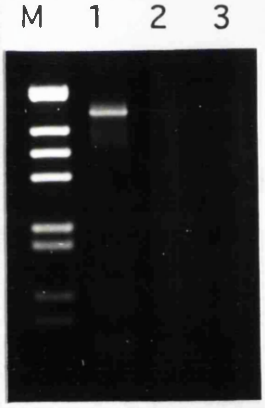
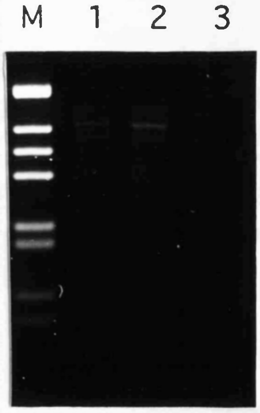
Primers	Template DNA	Results	
DegSer116F/ DegMgR	cDNA samples:		
	K562 pA$^+$ RNA	Lane 1: No distinct bands	
	JG total RNA	Lane 2: Smear, band of 450bp	
	Rat pA$^+$ RNA	Lane 3: Bands of 300 & 520bp	
	Negative control	Lane 4: dH$_2$O	
	DNA sample:		
	E.coli	Lane 1: band of 475bp	
	RT-PCR control	Lane 2: 7014	
	Negative control	Lane 3: dH$_2$O	
DegSer116F2/ DegMgR2	cDNA libraries:		
	K562 (H5)	Lane 1: Smear, band of 420bp	
	Placenta (H9)	Lane 2: Smear, band of 420bp	
	Negative control	Lane 3: dH$_2$O	

Figure 4.10 Transformation results of PGM cDNA degenerate primer PCR products.

Primers	Template	No. white colonies/ total	No. colonies selected	No. inserts	No. vector rearr.	Inserts
DegSer116F/ DegMgR	RNA samples:					
		116/145	6	2	2	-
	K562 pA^+ RNA	257/362	12	4	2	1 x DegSer116F F&R 1 x novel hnRNP protein, L protein
	Rat pA^+ RNA	20/51	18	1	-	1 x Rat PGM1
	JG total RNA	25/60	18	7	4	1 x primers not in frame 2 x putative human homoloogue of yhbf
	DNA sample:					
	E.coli	20/51	12	4	-	4 x *yhbf* gene of *E.coli*
DegSer116F2/ DegMgR2	cDNA libraries:					
	K562 (H5)	108/256	24	3	-	1 x no data - failed to sequence 1 x 47 to 48 centrisome, *E.coli* K15 1 x similar to *E.coli* K12 chr. region
	Placenta (H9)	68/199	24	4	1	1 x no ORF 2 x *E.coli* K12 chr. region, 92.8-0.01 min

E.COLI YHBF: Four clones were isolated from *E.coli* and each contained an identical nucleotide sequence, which was identified as *yhbf* from *E.coli* (Genbank Acc. No. L12968). This is a hypothetical protein of unknown function but contains both the active site and the magnesium binding loop protein motifs. It shows 52.4% similarity/25.8% identity at the amino acid level with human PGM1 (Bestfit, GCG).

HUMAN *YHBF?*: Two independently isolated clones from JG were shown to contain inserts of identical nucleotide sequence. A preliminary database search indicated that the 210bp sequence was 61% identical to the *yhbf* gene of *E.coli.* Sequencing of the insert in both directions revealed an ORF in the same frame as the peptides on which the degenerate primers are based. A detailed characterization of this novel PGM-like sequence is provided in Chapter Five.

4.2.2.5 Results of Nested Degenerate Primer PCR

The use of nested degenerate primers appeared to improve the specificity of the PCR (Figure 4.11). In the cDNA samples, faint but distinct bands were obtained. A clean 475bp band was amplified from the cDNA libraries, whilst a ladder of bands was amplified from *T.cruzi.* The subsequent transformation results are shown in figure 4.12. In total, 84 recombinant plasmids were analyzed, with 37 showing inserts of between 250bp and 750bp. No novel PGM-like sequences were identified from the cDNA samples. However, the identity of the distinct bands amplified from the cDNA libraries and the isolation of a novel non-PGM-related sequence from *T.cruzi* were obtained.

E.COLI PGM: Nested degenerate primer PCR carried out on the K562 and placenta cDNA libraries resulted in the amplification of a 475bp band. All of the 27 clones isolated were shown to have originated from *E.coli.* One sequence, originating from the placental cDNA library, was identified as the *E.coli putA* gene encoding the multifunctional enzyme proline dehydrogenase/δ-1-pyrroline-5-carboxylate dehydrogenase (Genbank Acc. No. U05212). Although the sequence was primed with the DegSer116F2 primer, it was not PGM-related. The remaining sequences, 16 from the K562 cDNA library and 10 from the placental cDNA library, were all identified as *E.coli pgm* (Genbank Acc. No. M77127). Thus, it appears that the aliquots of cDNA library contain contaminating bacterial DNA, and therefore, they are not ideal resources for this strategy.

T.CRUZI ASAT: Although the *PGM* gene from *T.cruzi* was not identified by nested degenerate primer PCR, a novel non-PGM-related sequence was isolated from

Figure 4.11 PCR results from experiments using nested degenerate PGM cDNA primers

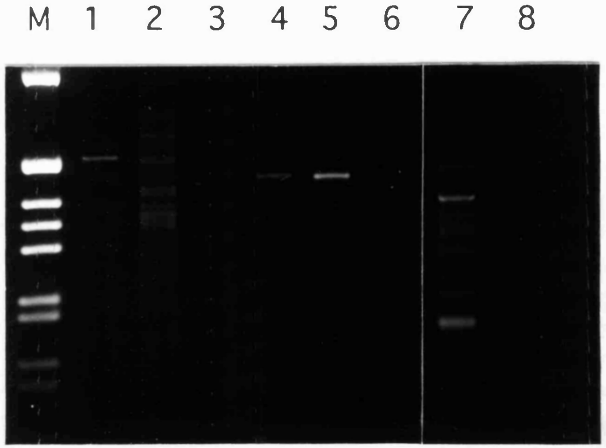
Primers	Template DNA	Results	
DegSer116F2/ DegPGMR & DegPMMR then DegPGMF & DegPMMF/ DegMgR2	cDNA samples:		
	ED total RNA	Lane 1: Ladder of bands including 520bp	
	TC total RNA	Lane 2: Ladder of bands including 520bp	
	Negative control	Lane 3: dH_2O	
	cDNA libraries:		
	K562 (H5)	Lane 4: 475bp band	
	Placenta (H9)	Lane 5: 475bp band	
	Negative control	Lane 6: dH_2O	
	DNA sample:		
	T.cruzi	Lane 7: Ladder of distinct bands, including 550bp	
	Negative control	Lane 8: dH_2O	

Figure 4.12 Transformation results of PGM cDNA nested degenerate primer PCR products.

Primer	Template	No. white colonies / total	No. selected colonies	No. inserts	Inserts
DegSer116F2/ DegPGMR & DegPMMR then DegPGMF & DegPMMF/ DegMgR	RNA sample:				
	ED total RNA	54/135	12	3	1 x no ORF 1 x similarity to 18s rRNA 1 x DegMgR F&R
	TC total RNA	72/144	12	3	2 x DegMgR F&R 1 x no data - mixed clone
	cDNA libaries:				
	K562 (H5)	65/233	18	16	16 x *E.coli pgm*
	Placenta (H9)	27/83	18	11	10 x *E.coli pgm* 1 x *E.coli* proline dehydrogenase (*putA*)
	DNA sample:				
	T.cruzi	220/370	24	4	1 x aspartate aminotransferase (*ASAT*) 1 x *T.cruzi* 82kDa surface antigen 1 x no ORF 1 x DegMgR F&R

T.cruzi which showed homology to a variety of eukaryotic cytosolic aspartate aminotransferase genes (*ASAT*). The nucleotide sequence was 68% identical to mouse, 67% identical to chicken and 65% identical to arabidopsis *ASAT* over 137 nucleotides. The outer set of nested primers amplified the DNA, with DegSer116F2 acting as the reverse primer and DegPGMR as the forward primer. ASAT, like PGM, is an isozyme marker for distinguishing between different isolates of *T.cruzi*. Therefore, the LSHTM were provided with the clone in order to identify and characterize the *ASAT* gene.

4.2.3 AGM SEQUENCE BASED DEGENERATE PRIMER PCR

N-acetylglucosamine phosphomutase (AGM) has been identified in *S.cerevisiae*, and shown to possess phosphoglucomutase activity. The yeast protein is of similar molecular weight to human PGM1 and the contains the active site and magnesium binding loop motifs characteristic of the other phosphohexomutases. The distance between these two motifs in AGM is 228 amino acids, which is greater than the 168 amino acids seen in human PGM1. Degenerate primers specific to the AGM protein were designed to investigate the presence of a homologue in the human genome. The strategy employed was the same as used previously (refer to figure 4.4).

4.2.3.1 AGM Degenerate Primers, Template DNA and PCR Conditions

The problem with designing AGM specific degenerate primers was deciding the amino acid sequence upon which they should be based. Without any homologous sequences for comparison, or information on the secondary structure of the protein, identification of probable conserved regions is more difficult. However, proline, glycine and hydrophobic residues are found to be highly conserved betweeen homologues, as they are generally important for maintaining the secondary structure. Thus, the forward degenerate primer, DegAGMF1 covered the magnesium binding loop residues FDGDADR and the reverse primer, DegAGMR1, was based on a hydrophobic region of the protein, DMLAVL (Figure 4.13).

Nested degenerate primer PCR was carried out in an attempt to increase the specificity of the PCR. Since AGM contains both the active site and magnesium binding site motifs, the degenerate primers DegSer116F2 and DegMgR2 were used along with two new AGM specific primers (Figure 4.13). DegAGMF2 was based on a hydrophobic region, GILAV, at the carboxyl end of the protein and was used as an outer primer with DegMgR2. DegAGMR2,

Figure 4.13 Location of the AGM cDNA degenerate primers.

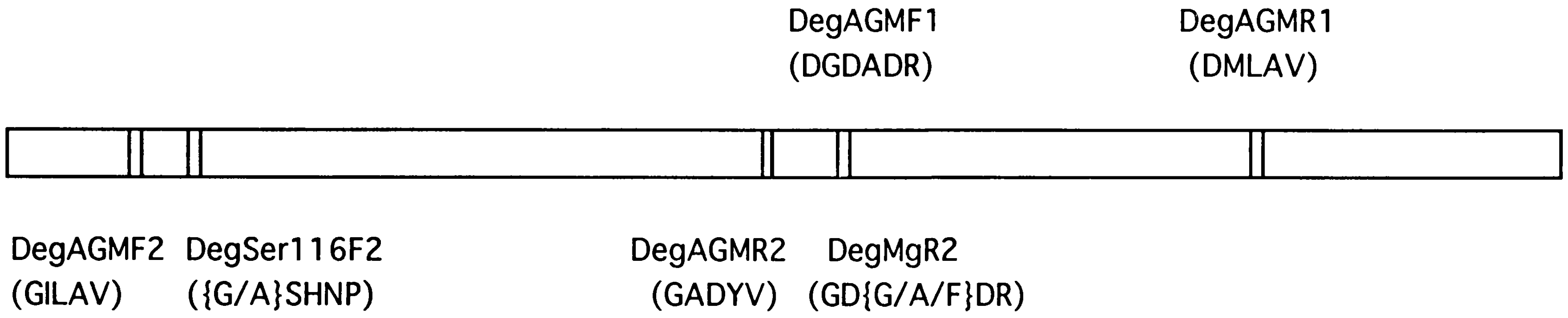


Figure 4.14 Nucleotide sequences of AGM cDNA degenerate primers and annealing temperatures.

Forward Primer	Reverse Primer	ANNEALING TEMPERATURE
DegAGMF1 5' TTYGAYGGNGAYGCNGAYAG 3'	DegAGMR1 5' ARNACNGCNAGCATRTC 3'	45°C / Tm°C
DegAGMF2 5' GAAGCTTCGGNATYYTNGCNGT 3'	DegMgR2 5' CTCTAGAGCGRTCNVMRTCNCC 3'	55°C / Tm°C
DegSer116F2 5' GAAGCTTCGSNWSNAYAAYCC 3'	DegAGMR2 5' CTCTAGAGACRTARTCNGCNCC 3'	58°C / Tm°C

based on hydrophobic residues, GADYV, located 22 amino acids upstream of the magnesium binding loop, was used for the nested PCR with DegSer116F2. A summary of the data showing the nucleotide sequences of the AGM primers and the annealing temperatures used is seen in figure 4.14.

Total RNA from the cell lines K562, JG, ED and TC was reverse transcribed using random hexamer primers. In addition, degenerate PCR was carried out on the K562 and placenta cDNA libraries. The standard degenerate primer PCR programme was used: denaturing at 95°C for 30secs, annealing for 30 secs and elongation at 72°C for 45secs. Degenerate primer PCR consisted of 35 cycles and the reaction mix contained 10nmoles of each primer. Nested degenerate primer PCR consisted of two rounds of 25 cycles, with the reaction mix containing 100pmoles of each primer.

4.2.3.2 Results

Amplification with DegAGMF1 and DegAGMR1 of K562 cDNA produced a faint doublet of 420 and 450bp. No bands, however, were produced from the JG and ED cDNA samples (Figure 4.15). A faint band of 325bp was amplified from both of the cDNA libraries. The subsequent transformation results are shown in figure 4.16. In total, 93 recombinant plasmids were analyzed, and of these only 19 appeared to contain inserts of between 250bp and 750bp. Of these 16 were identified as vector rearrangements. None of the remaining three were novel PGM-like sequences; the single clone from K562 was identified as human 18s rRNA, and as found previously, the recombinant plasmids transformed with PCR products from the cDNA libraries contained *E.coli* sequences.

Nested degenerate primer PCR with the AGM specific primers on the cDNA samples did not increase the specificity of the PCR. None of the bands amplified by RT-PCR in the first round were enhanced by nested PCR primers in the second round. After both the first and second round of PCR, no products were detected from the cDNA libraries.

4.3 SUMMARY

i) Low stringency PCR was investigated in an attempt to identify *PGM2* and *PGM3* cDNA sequences. Under a variety of cycling parameters, only *PGM1* was amplified from the control lymphoblastoid cell lines. Hybridization of the PCR products with the HPGM probe identified no related sequences amplified from K562 or the control cell lines.

Figure 4.15 PCR results from experiments using degenerate AGM cDNA primers.

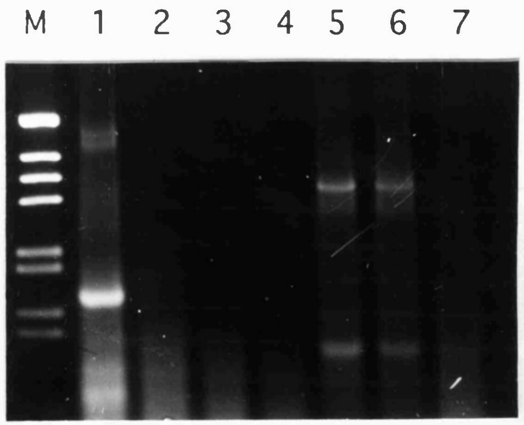
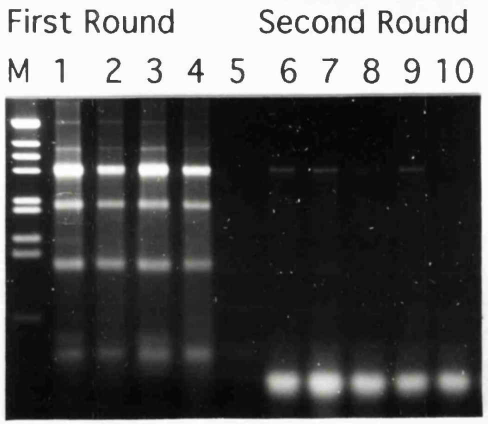
Primers	Template DNA	Results	
DegAGMF1/DegAGMR1	cDNA samples:		
	K562 pA$^+$ RNA	Lane 1: Faint doublet of 420 &450bp, strong band 175bp	
	JG total RNA	Lane 2: No bands, low mol. wgt. smear	
	ED total RNA	Lane 3: No bands, low mol. wgt. smear	
	Negative control	Lane 4: dH$_2$O	
	cDNA libraries:		
	K562 (H5)	Lane 5: Faint 325bp band	
	Placenta (H9)	Lane 6: Faint 325bp band	
	Negative control	Lane 7: dH$_2$O	
DegAGMF2/DegMgR2 then DegSer116F2/DegAGMR2	cDNA samples:		 First Round Second Round M 1 2 3 4 5 6 7 8 9 10
	K562 total RNA	Lanes 1 & 6	
	JG total RNA	Lanes 2 & 7	
	ED total RNA	Lanes 3 & 8	
	TC total RNA	Lanes 4 & 9	
	Negative control	Lanes 5 & 10: dH$_2$O	
	cDNA libraries:	No bands or smears were produced from either library	
	K562 (H5)		
	Placenta (H9)		

Figure 4.16 Transformation results of AGM cDNA degenerate primer PCR products.

Template	Primers	No. white colonies/ total	No. selected colonies	No. inserts	No. vector rearr.	Inserts
RNA sample:	DegAGMF1/ DegAGMR1					
K562 total RNA		31/76	18	1		1 x human 18s rRNA
		30/56	27	0		
cDNA libraries:						
K562 (H5)		25/67	24	12	11	1 x *E.coli* chr. region 76.0-81.5 min
Placenta (H9)		63/163	24	6	5	1 x *E.coli* transposable element IS21

ii) Degenerate primer PCR was investigated, with redundancies in the nucleotide sequence of the primers allowing less conserved sequences to be amplified. Primers were based upon conserved motifs found in the PGM and PMM proteins and also more specifically in the yeast AGM. In both cases, nested degenerate primers were used to try to improve the specificity of the PCR. Transformation with the degenerate primer PCR products produced 1201 recombinant plasmids, of which 291 (24%) were analyzed. Of these, only 85 (29%) contained inserts (of non-vector origin) and 2 (0.7%) were identified as novel PGM-like sequences.

iii) RT-PCR using the PGM and PMM sequence based degenerate primers illustrated the technique was applicable for the identification of homologous genes, with the cloning of a cDNA sequence from rat *PGM1*, and identification of more divergent PGM-related sequences from other species, with the cloning of *yhbf* from *E.coli*. Most importantly, RT-PCR of human RNA led to the identification of two recombinant plasmids each containing an identical novel PGM-related sequence. A detailed characterization of this sequence is presented in Chapter Five.

iv) The use of nested PGM and PMM sequence based degenerate primers on the cDNA samples did not appear to improve the specificity of the PCR, with respect to cloning novel members of the PGM gene family. However, an improvement in the specificity of the PCR was obtained using the cDNA libraries as template, such that the *E.coli PGM* gene from contaminating bacterial DNA was amplified, cloned and sequenced. The nested degenerate primers also amplified a partial coding sequence for the aspartate amino transferase (*ASAT*) gene in *T.cruzi*.

v) RT-PCR using the AGM-based degenerate primers did not identify any novel PGM-related sequence. The use of nested primers did not appear to improve the specificty of the PCR. However, this may be due to the choice of primers rather than absence of an *AGM* homologue in man.

4.4 CONCLUSIONS

Low stringency PCR suggests that the nucleotide sequence of conserved regions of the PGM1 protein has diverged sufficiently between PGM1 and the other PGM isozymes that the primers are unable to anneal to the template DNA and initiate amplification. Therefore this strategy is unsuitable for the identification of PGM2 and PGM3 cDNA sequences.

Degenerate primer PCR has been successful, with the identification of a novel PGM-related sequence from human RNA. However, this is not a highly efficient technique. The degenerate primers are capable of producing PCR products which are subsequently not identified by random selection of recombinant plasmids; for example, *PGM1* was amplified from the human B cell lines, as shown by hybridization with the HPGM1 probe, yet no recombinant plasmids containing human *PGM1* cDNA sequences were identified. Many selected recombinant plasmids were identified as false positives, and true positives, on further analysis, were revealed as vector rearrangements. In addition, the primers amplify sequences which are not PGM-related, such as *T.cruzi ASAT*. In this sequence, the DegSer116F2 primer acted as a reverse primer and was found to encode the peptide PSRIACG, whilst the DegPGMR primer was the forward primer, encoding the peptide SRAERLL. This serves to demonstrate the numerous peptides encoded by the primers, in addition to those on which they are based.

As a source of template DNA, K562 was judged to be a useful resource for these experiments. Due to the low levels of PGM1 mRNA transcript, cDNA transcribed from K562 was thought to contain a lower pool of *PGM1* cDNA than most other sources, and therefore, both low stringency and degenerate primer PCR would produce a higher ratio of amplified PGM-related sequences to *PGM1* than most other sources. However, the use of K562 did not lead to the identification of any novel PGM-related sequences, perhaps due to the low efficiency of the degenerate primer strategy, discussed above.

In addition, contaminating bacterial DNA was identified in the ligation mixes of the cDNA libraries. Due to the ability of the degenerate primers to amplify the diverged bacterial sequences, these libraries were therefore not ideal resources for this degenerate primer PCR strategy. This problem could be overcome, however, by using primers specific to the multiple cloning site of the plasmid vector in combination with the degenerate primers. For example, a primer designed to the multiple cloning site downstream of the region of insertion could be used in combination with the degenerate primer DegSer116F2. Only cloned cDNA with sequences complementary to the DegSer116F2 primer, and in the correct orientation, would be amplified. A reciprocal PCR experiment using a primer designed to the multiple cloning site upstream of the region of insertion could then be used to amplify cloned cDNA in the opposite orientation.

CHAPTER FIVE:

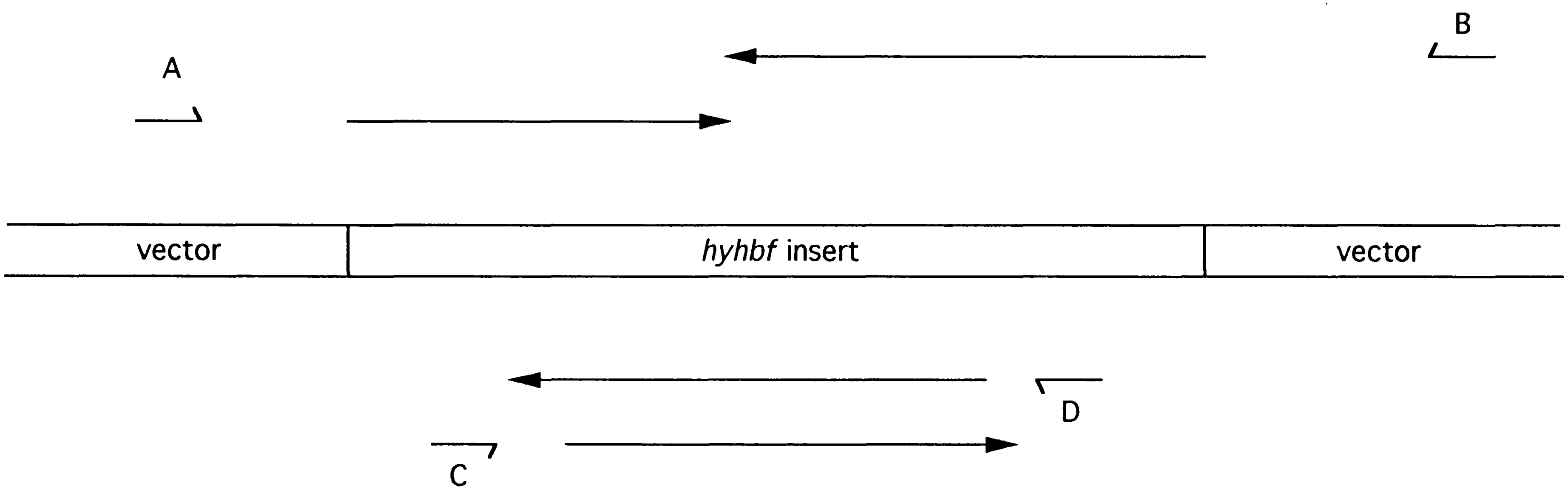
CHARACTERIZATION OF HUMAN YHBF SEQUENCE

A novel PGM-like sequence was cloned using the degenerate primer PCR strategy described in Chapter Four. In summary, cDNA prepared from the JG lymphoblastoid cell line was amplified using the DegSer116F and DegMgR primers and the PCR products were cloned. Partial nucleotide sequences of inserts from two clones were determined in both forward and reverse directions. The clones were identical and an open reading frame (ORF) was identified. The nucleotide sequence was used to screen the Genbank and Swissprot databases and a good match was found with the *yhbf* gene of *E.coli*. The *yhbf* gene encodes a hypothetical protein of unknown function but which contains both the active site and magnesium binding loop characteristic of the PGM gene family. This chapter describes the characterization of the human partial cDNA clone, including further nucleotide sequence analysis, RT-PCR and genomic DNA PCR experiments, and an attempted chromosomal localization. Due to the high level of similarity and identity of the sequence to *yhbf* at both the nucleotide and amino acid level, it is referred to as *hyhbf* (for human *yhbf*).

5.1 NUCLEOTIDE SEQUENCE ANALYSIS

The entire nucleotide sequence of the *hyhbf* insert was obtained using plasmid vector based forward and reverse primers for sequencing (Figure 5.1). There was a 5bp overlap. The complete sequence of both strands was verified using internal forward and reverse primers. The initial comparison of the deduced amino acid sequences of hyhbf and yhbf revealed that the *hyhbf* sequence was out of frame in relation to *yhbf*. The position in the protein sequence where the misalignment had occurred corresponded exactly to a GC-rich region in the *hyhbf* nucleotide sequence. Such GC-rich regions are prone to form secondary structures and are a well known source of sequencing artefacts, such as compressions. Re-examination of the autoradiographs revealed a very dark staining band in the C track within the GC-rich region, indicating a possible sequence compression. This type of artefact can sometimes be prevented by using the purine analogue inosine (dITP), in place of guanine. In the case of *hyhbf*, the use of dITP was successful and confirmed that there was a compression in this GC-rich region; the dark staining band in the C track divided into two bands and this led to the restoration of the expected reading frame.

Figure 5.1 Sequencing strategy to obtain the full insert sequence of *hyhbf*



Sequencing Primers

A: M13 (-24) Reverse Primer	5' AACAGCTATGACCATG 3'	C: SEQ.F	5' AATTCTTCTCGGGCCAGG 3'
B: M13 (-40) Forward Primer	5' GTTTTCCCAGTCACGAC 3'	D: SEQ.R	5' CTCGACATGGGTCGAACC 3'

The nucleotide sequence of the *hyhbf* insert consists of 448bp and encodes a polypeptide of 149 amino acids (Figure 5.2). It has a high G+C content, of 0.61, compared with 0.54 G+C for *E.coli yhbf* and 0.47 G+C for human *PGM1*, over the corresponding regions. Restriction endonuclease maps of *hyhbf* and *yhbf* are compared in figure 5.3. The high G+C content of the *hyhbf* sequence is reflected by the number of recognition sites for rare cutter restriction endonucleases. Of the five restriction endonucleases shown which have recognition sites in *hyhbf*, only a single *Bss*HI site is present in *yhbf* and this site is not conserved between the two sequences. The deduced amino acid sequence of *hyhbf* contains 24% charged residues (K,R,D,E,H), of which there is an excess of acidic residues (D+E); the predicted isoelectric point of hyhbf is pI 4.58 (Peptidesort, GCG).

5.2 COMPUTER BASED ANALYSIS

The *hyhbf* nucleotide and amino acid sequences were used to search the Genbank and Swissprot databases; with both databases the best matches were again with *yhbf* from *E.coli.*

A comparison of the *hyhbf* and *yhbf* nucleotide sequences (bestfit, GCG) shows an identity of 64.6% (Figure 5.4). This is remarkably high for DNA sequences from such diverse origins. However, it is not without precedent, as the nucleotide sequence of the *A.tumefaciens PGM* gene has a 60.6% sequence identity with that of human *PGM1.* Not unexpectedly, the high level of conservation between *yhbf* and *hyhbf* nucleotide sequence is reflected in the deduced amino acid sequences, which are 73.6% similar and 61.1% identical (Figure 5.5).

In contrast, amino acid sequence comparison of hyhbf with human PGM1 (Figure 5.6) shows a much poorer match, with a sequence similarity of 49.0% and identity of 29.6%. Nevertheless, in addition to the active site and magnesium binding loop peptides, the conserved amino acids include a number of the glycines and prolines and several hydrophobic residues. These residues are believed to be important for protein secondary structure. For example, comparison with the human PGM1 3-D structure indicates that the conserved glycines and prolines occur in regions of β-sheet, or at the beginning and end of α-helices, where they serve an important structural role. The hydrophobic residues, on the other hand, are located on internal surfaces of the protein, and therefore, are more likely to be conserved than surface residues (Creighton, 1993).

<u>Figure 5.2</u> Complete nucleotide sequence and amino acid translation of the *hyhbf* insert. The degenerate primer sequences which amplified *hyhbf* are shown in bold.

```
        DegSer116F
    ggaattccgcgaggcacaatccgcatcacgacaacggcatcaaattcttctcgggccagg
  1 ---------+---------+---------+---------+---------+---------+ 60
    ccttaaggcgctccgtgttaggcgtagtgctgttgccgtagtttaagaagagcccggtcc

      N  S  A  R  H  N  P  H  H  D  N  G  I  K  F  F  S  G  Q  G

    gcaccaagctgccggacgagatcgaattgatgattgaagagttgctcgatcagccgatga
 61 ---------+---------+---------+---------+---------+---------+ 120
    cgtggttcgacggcctgctctagcttaactactaacttctcaacgagctagtcggctact

      T  K  L  P  D  E  I  E  L  M  I  E  E  L  L  D  Q  P  M  T

    cggtggtcgagtccgagcagctgggcaaggtgtcgcgaatcaacgacgccgccggccgct
121 ---------+---------+---------+---------+---------+---------+ 180
    gccaccagctcaggctcgtcgacccgttccacagcgcttagttgctgcggcggccggcga

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

    acatcgaattctgtaagagcagcgtgccgagcagcaccgactttgccgggctgaagatcg
181 ---------+---------+---------+---------+---------+---------+ 240
    tgtagcttaagacattctcgtcgcacggctcgtcgtggctgaaacggcccgacttctagc

      I  E  F  C  K  S  S  V  P  S  S  T  D  F  A  G  L  K  I  V

    ttgtcgactgtgctcacggtgcggcctacaaggttgcgccgagtgtattccgcgaattgg
241 ---------+---------+---------+---------+---------+---------+ 300
    aacagctgacacgagtgccacgccggatgttccaacgcggctcacataaggcgcttaacc

      V  D  C  A  H  G  A  A  Y  K  V  A  P  S  V  F  R  E  L  G

    gcgcgcaggtggcggtgctctcggcgcagcccaatggcttgaacatcaatgatggttgcg
301 ---------+---------+---------+---------+---------+---------+ 360
    cgcgcgtccaccgccacgagagccgcgtcgggttaccgaacttgtagttactaccaacgc

      A  Q  V  A  V  L  S  A  Q  P  N  G  L  N  I  N  D  G  C  G

    gttcgacccatgtcgaggcgttgcaggccgaggtgctggcgcagcaggcggatctgggta
361 ---------+---------+---------+---------+---------+---------+ 420
    caagctgggtacagctccgcaacgtccggctccacgaccgcgtcgtccgcctagacccat

      S  T  H  V  E  A  L  Q  A  E  V  L  A  Q  Q  A  D  L  G  I

    ttgccttcgacggcgacgggcctgcagg
421 ---------+---------+-------- 448
    aacggaagctgccgctgcccggacgtcc
            DegMgR

      A  F  D  G  D  G  P  A
```

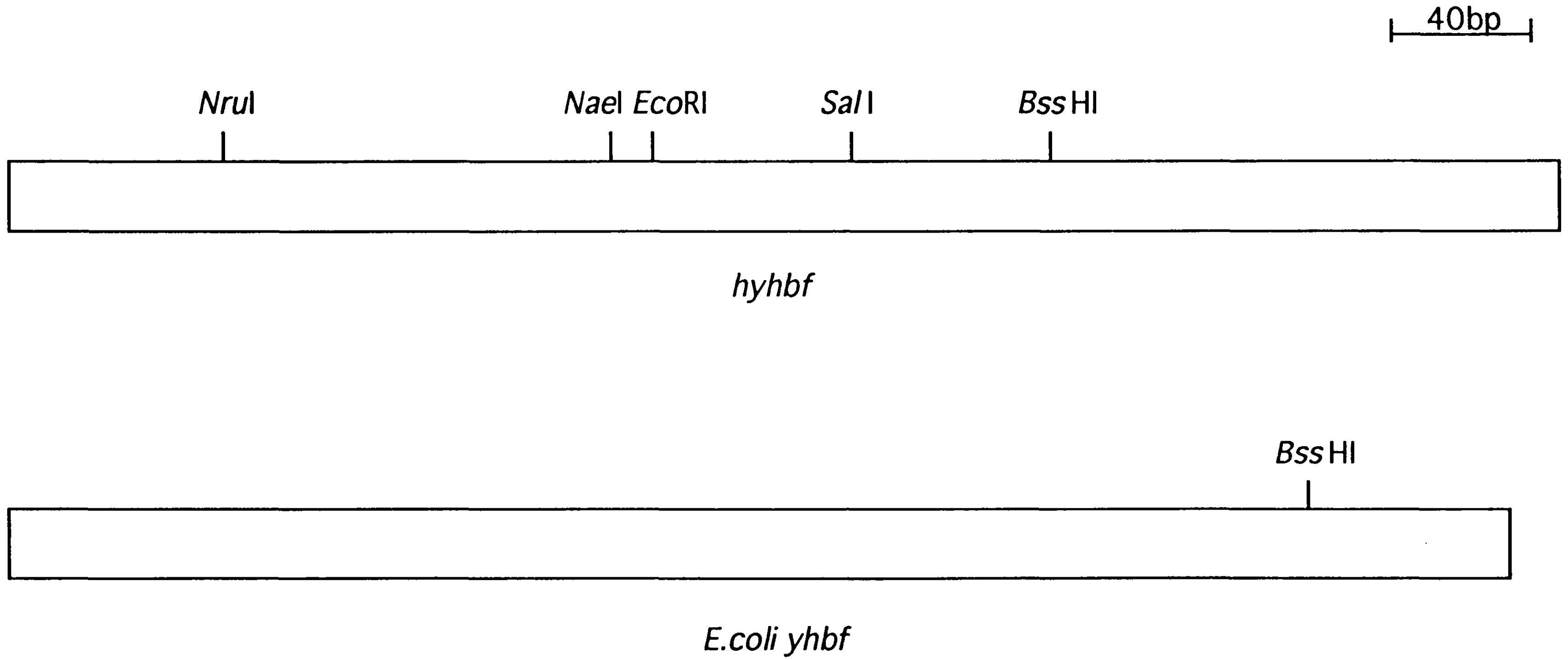


Figure 5.3 Restriction map of *hyhbf* and *E.coli yhbf* . *Eco*RI and *Sal* I were used to confirm the identity of amplified hyhbf sequences. *Nru*I, *Nae*I, and *Bss*HI are restriction endonucleases whose recognition sites include CpG.

Figure 5.4 Nucleotide sequence comparison of *hyhbf* and *E.coli yhbf*. The degenerate primer sequences which amplified *hyhbf* are shown in bold.

```
                         .         .         .         .         .
Human       6 TCCGCGAGGCACAATCCGCATCACGACAACGGCATCAAATTCTTCTCGGG 55
              || ||   ||| || |||    |||| || ||||| |||||||||||
E.coli   2082 TCTGCATCGCATAACCCGTTCTACGATAATGGCATTAAATTCTTCTCTAT 2033

                         .         .         .         .         .
Human      56 CCAGGGCACCAAGCTGCCGGACGAGATCGAATTGATGATTGAAGAGTTGC 105
              | | |||||||| |||||||| | | | |||  |   || |||| |
E.coli   2032 CGACGGCACCAAACTGCCGGATGCGGTAGAAGAGGCCATCGAAGCGGAAA 1983

                         .         .         .         .         .
Human     106 TCGATCAGCCGATGACGGTGGTCGAGTCCGAGCAGCTGGGCAAGGTGTCG 155
              | ||  ||  ||| |     || || || |   | ||||| || |
E.coli   1982 TGGAAAAGGAGATCAGCTGCGTTGATTCGGCAGAACTGGGTAAAGCCAGC 1933

                         .         .         .         .         .
Human     156 CGAATCAACGACGCCGCCGGCCGCTACATCGAATTCTGTAAGAGCAGCGT 205
              || |||   || ||||| || ||||| ||||| || || ||   ||   |
E.coli   1932 CGTATCGTTGATGCCGCGGGTCGCTATATCGAGTTTTGCAAAGCCACGTT 1883

                     .         .         .         .         .
Human     206 GCCGAGCAGCACCGACTTTGCC.....GGGCTGAAGATCGTTGTCGACTG 250
               ||||     ||  |||| |||     |  |||||||| || || || ||
E.coli   1882 CCCGA.....ACGAACTTAGCCTCAGTGAACTGAAGATTGTGGTGGATTG 1838

                         .         .         .         .         .
Human     251 TGCTCACGGTGCGGCCTACAAGGTTGCGCCGAGTGTATTCCGCGAATTGG 300
              |||  |||||||| | ||  |  | |||||||  ||  | |||||| |||
E.coli   1837 TGCAAACGGTGCGACTTATCACATCGCGCCGAACGTGCTGCGCGAACTGG 1788

                         .         .         .         .         .
Human     301 GCGCGCAGGTGGCGGTGCTCTCGGCGCAGCCCAATGGCTTGAACATCAAT 350
              | ||| | ||    |   ||       |||| || ||  | |||||||||
E.coli   1787 GGGCGAACGTTATCGCTATCGGTTGTGAGCCAAACGGTGTAAACATCAAT 1738

                         .         .         .         .         .
Human     351 GATGGTTGCGGTTCGACCCATGTCGAGGCGTTGCAGGCCGAGGTGCTGGC 400
              |  |     ||  | ||| | ||    ||| | |||||    ||||||||
E.coli   1737 GCCGAAGTGGGGGCTACCGACGTTCGCGCGCTCCAGGCTCGTGTGCTGGC 1688

                         .         .         .
Human     401 GCAGCAGGCGGATCTGGGTATTGCCTTCGACGGCGACGG 439
                |  | |||||||| |||||||||||||||||||| ||
E.coli   1687 TGAAAAAGCGGATCTCGGTATTGCCTTCGACGGCGATGG 1649
```

Identity: 64.6%

<u>Figure 5.5</u> Amino acid sequence comparison of hyhbf and *E.coli* yhbf. Amino acids encoded by the degenerate primers are shown in bold.

```
                      .         .         .         .         .
Human      2 ARHNPHHDNGIKFFSGQGTKLPDEIELMIEELLDQPMTVVESEQLGKVSR 51
             |.||| .|||||||| :||||||.:|  ||. ::..:..|:|.:|||.||
E.coli   101 ASHNPFYDNGIKFFSIDGTKLPDAVEEAIEAEMEKEISCVDSAELGKASR 150

                      .         .         .         .         .
Human     52 INDAAGRYIEFCKSSVPSSTDFAGLKIVVDCAHGAAYKVAPSVFRELGAQ 101
             | |||||||||||...|.. .:.:||||||||:||.|.:||.|:|||||.
E.coli   151 IVDAAGRYIEFCKATFPNELSLSELKIVVDCANGATYHIAPNVLRELGAN 200

                      .         .         .         .
Human    102 VAVLSAQPNGLNINDGCGSTHVEALQAEVLAQQADLGIAFDGDG 145
             | .::.:|||:|||.:.|.|.| |||| |||:.|||||||||||
E.coli   201 VIAIGCEPNGVNINAEVGATDVRALQARVLAEKADLGIAFDGDG 244
```

Similarity: 73.6%
Identity: 61.1%

<u>Figure 5.6</u> Amino acid sequence comparison of hyhbf and human PGM1. Amino acids encoded by the degenerate primers are shown in bold.

```
                      .         .         .         .         .
yhbf       1 SARHNP...HHDNGIKFFSGQGTKLPDEIELMIEEL............LD 35
             .|.|||   : | ||||  :.|.. |:.|. .| ::            |.
PGM1     114 TASHNPGGPNGDFGIKFNISNGGPAPEAITDKIFQISKTIEEYAVCPDLK 163

                      .         .         .         .         .
yhbf      36 QPMTVVESEQLG...KVSRINDAAGRYIEFCKSSVPSSTDFAGL...... 76
              .:.|::.:|::   |...:. . .  :| : . :.|  ||.:|
PGM1     164 VDLGVLGKQQFDLENKFKPFTVEIVDSVEAYATMLRSIFDFSALKELLSG 213

                      .         .         .         .         .
yhbf      77 ....KIVVDCAHGAAYKVAPSVF.RELGAQV..AVLSAQPNGLNIND.GC 118
                 ||.:|. ||.. . ...::  ||||..  || :.. :::. :. :.
PGM1     214 PNRLKICIDAMHGVVGPYVKKILCEELGAPANSAVNCVPLEDFGGHHPDP 263

                      .         .
yhbf     119 GSTHVEALQAEVLAQQADLGIAFDGDG 145
             . |....| ..: ..: |:| ||||||
PGM1     264 NLTYAADLVETMKSGEHDFGAAFDGDG 290
```

Similarity: 49.0%
Identity: 29.6%

Amino acid sequence comparisons between hyhbf and a selection of prokaryotic and eukaryotic PGM and PMMs indicated that these key residues are widely conserved within the phosphohexomutases suggesting conservation of protein secondary and tertiary structure (Figure 5.7); it appears, therefore, that the *hyhbf* sequence may be a highly diverged member of the PGM gene family.

5.3 PCR ANALYSIS OF THE HYHBF SEQUENCE

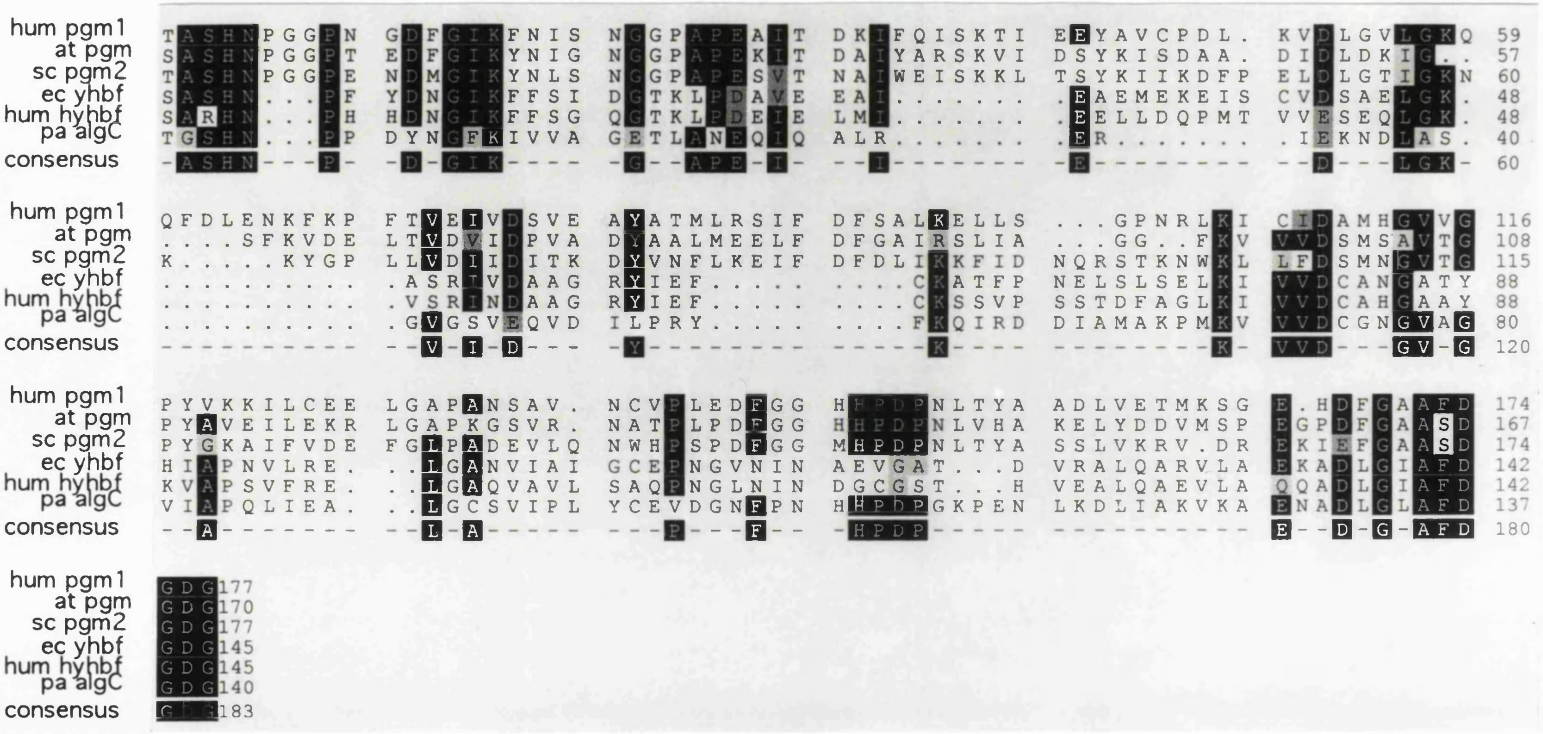
5.3.1 RT-PCR WITH HYHBF.F AND HYHBF.R

In order to investigate the expression of *hyhbf,* RT-PCR was carried out using the HYHBF.F and HYHBF.R primers as described (Figure 5.8). A panel of cDNAs was prepared from K562 pA$^+$RNA and total RNA from K562, JG, ED, Goodwin, Molt-4, 6997 and Storey (all cell lines) as well as human liver total RNA. In the first experiment, RT-PCR was carried out on the full panel except for K562 total RNA. The expected 312bp band was amplified from the JG and K562 (pA+) samples, indicating the presence of an *hyhbf* transcript, but not from the remainder where a ladder of non-specific bands were observed (Figure 5.9).

5.3.2 NESTED RT-PCR

To overcome the problem of weak amplification found in K562, and to increase the specificity of the PCR, a pair of internally nested primers, HYHBF.F2 and HYHBF.R2, were designed (Figure 5.8). Successful PCR between these nested primers should produce a band of 261bp. The full panel of total RNA samples was amplified, except Goodwin, and the expected 261bp band was seen in all cases except from K562 and 6997 total RNA (Figure 5.10, lanes 1-8). However, the expected product was amplified from K562 pA$^+$ RNA (Figure 5.10, lane 9). In two of the samples, JG and K562 pA$^+$ RNA, an extra PCR product of 290bp was observed. This is consistent with amplification having occurred from one internal and one outer primer. It is not clear why the 290bp product was only amplified from these two samples. All the nested PCR products, including the 290bp bands, were confirmed as *hyhbf* by diagnostic restriction endonuclease digests using *Eco*RI and *Sal*I.

In addition, plasmid DNA preparations from two cDNA libraries (K562 and human placenta) were investigated using nested primer PCR to determine if the



```
hum pgm1   TASHNPGGPN GDFGIKFNIS NGGPAPEAIT DKIFQISKTI EEYAVCPDL. KVDLGVLGKQ  59
at pgm     SASHNPGGPT EDFGIKYNIG NGGPAPEKIT DAIYARSKVI DSYKISDAA. DIDLDKIG..  57
sc pgm2    TASHNPGGPE NDMGIKYNLS NGGPAPESVT NAIWEISKKL TSYKIIKDFP ELDLGTIGKN  60
ec yhbf    SASHN...PF YDNGIKFFSI DGTKLPDAVE EAI....... .EAEMEKEIS CVDSAELGK.  48
hum hyhbf  SARHN...PH HDNGIKFFSG QGTKLPDEIE LMI....... .EELLDQPMT VVESEQLGK.  48
pa algC    TGSHN...PP DYNGFKIVVA GETLANEQIQ ALR....... .ER....... .IEKNDLAS.  40
consensus  -ASHN---P- -D-GIK---- -G--APE-I- --I------- -E-------- --D---LGK-  60

hum pgm1   QFDLENKFKP FTVEIVDSVE AYATMLRSIF DFSALKELLS ...GPNRLKI CIDAMHGVVG 116
at pgm     ....SFKVDE LTVDVIDPVA DYAALMEELF DFGAIRSLIA ...GG..FKV VVDSMSAVTG 108
sc pgm2    K.....KYGP LLVDIIDITK DYVNFLKEIF DFDLIKKFID NQRSTKNWKL LFDSMNGVTG 115
ec yhbf    .......... .ASRIVDAAG RYIEF..... ....CKATFP NELSLSELKI VVDCANGATY  88
hum hyhbf  .......... .VSRINDAAG RYIEF..... ....CKSSVP SSTDFAGLKI VVDCAHGAAY  88
pa algC    .......... .GVGSVEQVD ILPRY..... ....FKQIRD DIAMAKPMKV VVDCGNGVAG  80
consensus  ---------- --V-I-D--- -Y-------- -----K---- --------K- VVD---GV-G 120

hum pgm1   PYVKKILCEE LGAPANSAV. NCVPLEDFGG HHPDPNLTYA ADLVETMKSG E.HDFGAAFD 174
at pgm     PYAVEILEKR LGAPKGSVR. NATPLPDFGG HHPDPNLVHA KELYDDVMSP EGPDFGAASD 167
sc pgm2    PYGKAIFVDE FGLPADEVLQ NWHPSPDFGG MHPDPNLTYA SSLVKRV.DR EKIEFGAASD 174
ec yhbf    HIAPNVLRE. ..LGANVIAI GCEPNGVNIN AEVGAT...D VRALQARVLA EKADLGIAFD 142
hum hyhbf  KVAPSVFRE. ..LGAQVAVL SAQPNGLNIN DGCGST...H VEALQAEVLA QQADLGIAFD 142
pa algC    VIAPQLIEA. ..LGCSVIPL YCEVDGNFPN HHPDPGKPEN LKDLIAKVKA ENADLGLAFD 137
consensus  --A------- --L-A----- ---P---F-- -HPDP----- ---------- E--D-G-AFD 180

hum pgm1   GDG177
at pgm     GDG170
sc pgm2    GDG177
ec yhbf    GDG145
hum hyhbf  GDG145
pa algC    GDG140
consensus  GDG183
```

Figure 5.7 Amino acid sequence comparison of hyhbf and selected prokaryotic and eukaryotic PGMs and PMMs. hum pgm1 = human PGM1; at pgm = *Agrobacterium tumefaciens* PGM; sc pgm2 = *Saccharomyces cerevisiae* PGM2; ec yhbf = *Escherichia coli* yhbf; hum hyhbf = human hyhbf; pa algC = *Pseudomonas aeroginosa* algC (PMM).

Figure 5.8 Location of the *hyhbf* primers.

```
    ggaattccgcgaggcacaatccgcatcacgacaacggcatcaaattcttctcgggccagg
  1 ---------+---------+---------+---------+---------+---------+ 60
    ccttaaggcgctccgtgttaggcgtagtgctgttgccgtagtttaagaagagcccggtcc

              HYHBF.F                        HYHBF.F2
    gcaccaagctgccggacgagatcgaattgatgattgaagagttgctcgatcagccgatga
 61 ---------+---------+---------+---------+---------+---------+ 120
    cgtggttcgacggcctgctctagcttaactactaacttctcaacgagctagtcggctact

    cggtggtcgagtccgagcagctgggcaaggtgtcgcgaatcaacgacgccgccggccgct
121 ---------+---------+---------+---------+---------+---------+ 180
    gccaccagctcaggctcgtcgacccgttccacagcgcttagttgctgcggcggccggcga

    acatcgaattctgtaagagcagcgtgccgagcagcaccgactttgccgggctgaagatcg
181 ---------+---------+---------+---------+---------+---------+ 240
    tgtagcttaagacattctcgtcgcacggctcgtcgtggctgaaacggcccgacttctagc

    ttgtcgactgtgctcacggtgcggcctacaaggttgcgccgagtgtattccgcgaattgg
241 ---------+---------+---------+---------+---------+---------+ 300
    aacagctgacacgagtgccacgccggatgttccaacgcggctcacataaggcgcttaacc

    gcgcgcaggtggcggtgctctcggcgcagcccaatggcttgaacatcaatgatggttgcg
301 ---------+---------+---------+---------+---------+---------+ 360
    cgcgcgtccaccgccacgagagccgcgtcgggttaccgaacttgtagttactaccaacgc
                                             HYHBF.R2

    gttcgacccatgtcgaggcgttgcaggccgaggtgctggcgcagcaggcggatctgggta
361 ---------+---------+---------+---------+---------+---------+ 420
    caagctgggtacagctccgcaacgtccggctccacgaccgcgtcgtccgcctagacccat
         HYHBF.R

    ttgccttcgacggcgacgggcctgcagg
421 ---------+---------+-------- 448
    aacggaagctgccgctgcccggacgtcc
```

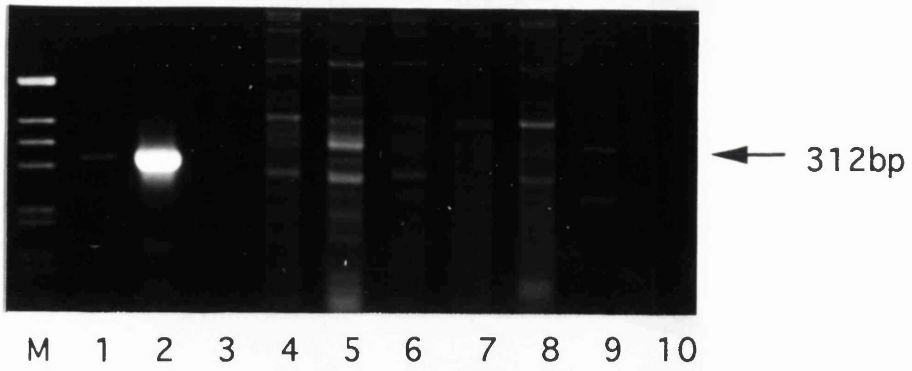


Figure 5.9 RT-PCR of cDNA samples with HYHBF.F and HYHBF.R. Lane 1 K562 (pA$^+$ RNA); lane 2 JG cell line; lane 3 dH_2O control; lane 4 Storey cell line; lane 5 ED cell line; lane 6 6997 cell line; lane 7 human liver; lane 8 Molt 4 cell line; lane 9 Goodman cell line; lane 10 dH_2O control; M = molecular weight size marker.

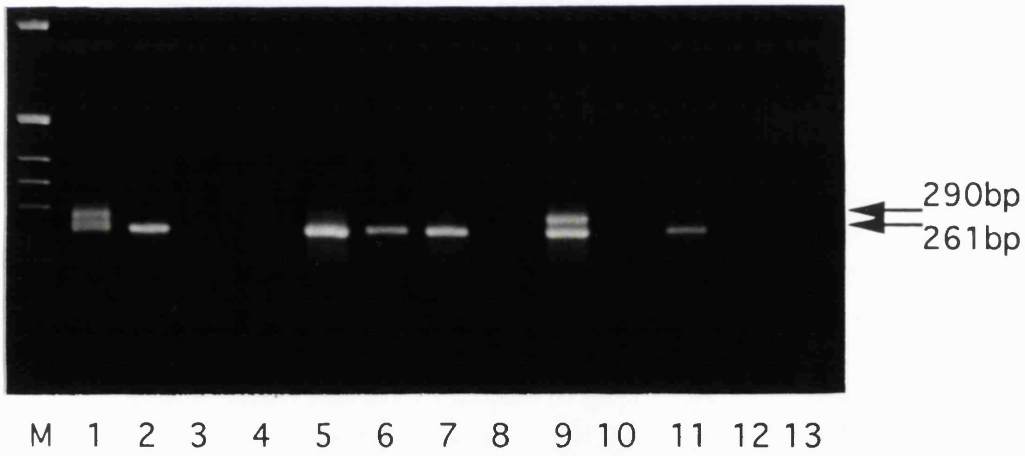


Figure 5.10 Nested RT-PCR of cDNA samples with HYHBF.F and HYHBF.R followed by HYHBF.F2 and HYHBF.R2. Lane 1 JG cell line; lane 2 ED cell line; lane 3 K562 cell line (total RNA);lane 4 6997 cell line; lane 5 Storey cell line; lane 6 human liver; lane 7 Molt4 cell line; lane 8 dH_2O control; lane 9 K562 cell line (pA$^+$ RNA); lane 10 dH_2O control; lane 11 K562 cDNA library; lane 12 human placental cDNA library; lane 13 dH_2O control; M = molecular weight size marker.

hyhbf sequence was present, and thus to assess the potential of using these libraries for the isolation of an *hyhbf* cDNA clone. The expected 261bp band was amplified from the K562 library but not the placental library (Figure 5.10, lanes 11 &12).

5.3.3 GENOMIC DNA PCR

Although the intron/exon structure of the putative *hyhbf* gene is unknown, an attempt was made to amplify *hyhbf* related sequences from genomic DNA. A positive result could occur if the primers were sited in a large exon, or spanned a small intron. In the case of the latter, this might have provided DNA of a length suitable for fluorescence in-situ hybridization, and to allow chromosomal localization of the sequence. Five genomic DNA samples (all prepared from leucocytes) were tested using the HYHBF.F and HYHBF.R primers. In each case a distinct pattern comprising of the same nine PCR products was generated (Figure 5.11). The sizes ranged between approximately 275bp and 900bp. From this data it is plausible that a small *hyhbf* intron has been amplified.

The specificity of the reaction was improved by using a touchdown PCR programme (Don et al, 1994). This technique was devised to circumvent the amplification of spurious bands due to mispriming by one or both of the primers. Generally, any smaller misprimed products have an advantage over a longer correct product during amplification. The touchdown PCR begins at an annealing temperature greater than Tm, and is decreased by 1°C every second cycle, over nine stages to touchdown for 30 cycles of PCR. For genomic DNA PCR of *hyhbf*, the optimal conditions were established by experiment. A touchdown PCR programme with a final annealing temperature of Tm+2°C led to the amplification of a 312bp band in most samples, whilst supressing the appearance of the larger bands seen previously (Figure 5.12). Despite many attempts to improve the PCR conditions, the genomic DNA experiments were difficult to reproduce. However, restriction endonuclease digests with *Eco*RI and *SalI* confirmed that the genomic DNA 312bp PCR products were from the *hyhbf* sequence (Figure 5.13). Thus it appears that the primers amplify the same 312bp band in DNA and RNA.

5.3.4 ORIGIN OF HYHBF SEQUENCE

In view of the genomic DNA PCR results, the possibility was investigated that the cloned *hyhbf* PCR product resulted from amplification of genomic DNA

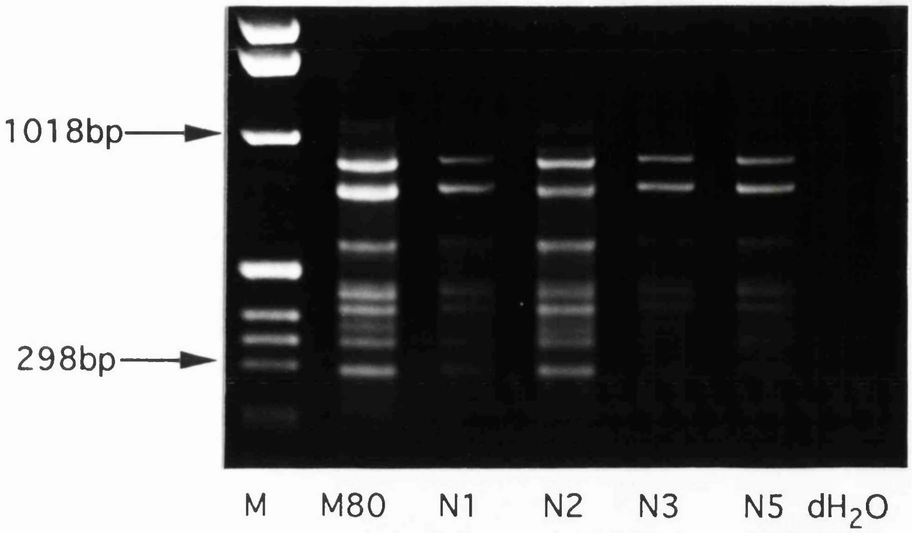


Figure 5.11 Standard genomic DNA PCR of leucocyte DNA samples using HYHBF.F and HYHBF.R. M = molecular weight size marker.

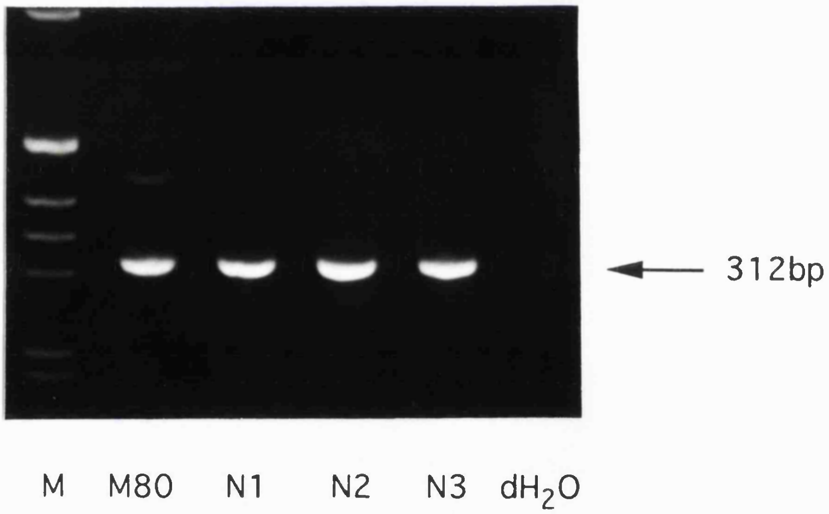


Figure 5.12 Touchdown PCR of leucocyte DNA samples using HYHBF.F and HYHBF.R. M = molecular weight marker.

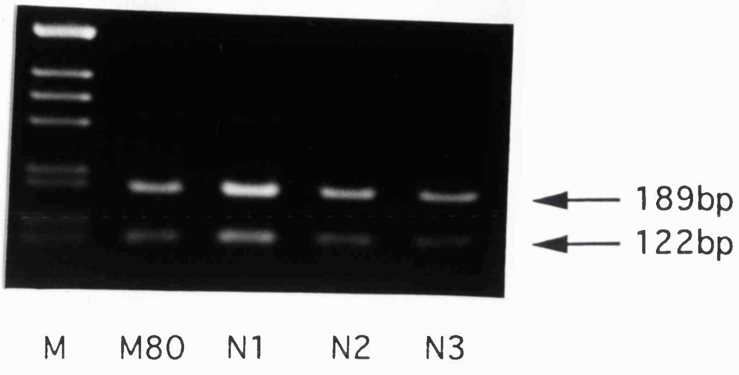


Figure 5.13 *Eco*RI digestion of *hyhbf* genomic DNA PCR products amplified with HYHBF.F and HYHBF.R. M = molecular weight marker.

rather than mRNA, in the JG RNA sample. Standard genomic PCR using the full panel of cDNA samples, and primers based upon intron sequences flanking exon 4 of the *PGM1* gene yielded a product of 295bp in all samples. This is the expected size of the PGM1 PCR product (Figure 5.14), indicating that genomic DNA was present in the RNA samples. The result also implies that the nested PCR products shown in figure 5.10 may have arisen from priming genomic DNA rather than cDNA.

The role of reverse transcriptase (RT) activity in the RT-PCRs was then investigated in an experiment using duplicate pairs of RNA samples amplified with or without RT. The results are summarized in figure 5.15. Nested RT-PCR of eight RNA samples initially amplified the expected 261bp band from two samples in which RT was included (JG and human kidney), but also from the TC B cell line, irrespective of the presence of RT. When the experiment was repeated, the 261bp product was amplified from K562, only in the presence of RT, but also from four other samples (including TC) irrespective of the presence of RT. Therefore, although there was inconsistency in the results between the two sets of experiments, successful PCR appeared to be independent of the presence of RT. This finding strengthens the evidence that the *hyhbf* band amplified from the RNA/cDNA samples is due, at least in part, to the presence of genomic DNA.

At first sight, the positive amplification of *hyhbf* from the K562 cDNA library provided evidence that the human sequence had been transcribed. The validity of this interpretation was investigated. Nested primer PCR was carried out using primers designed to the multiple cloning site of the cDNA library vector and the *hyhbf* reverse primers, such that amplification may occur whichever orientation the *hyhbf* sequence is cloned (Figure 5.16). A fresh aliquot of the K562 cDNA library was used as template and great care was taken to prevent contamination by extraneous DNA (from the environment). No amplification occurred as judged by ethidium bromide staining. Therefore, it would appear that the amplification of the *hyhbf* sequence from the K562 cDNA library may have been due to the presence of bacterial DNA, presumably *E. coli*, in the library plasmid DNA preparation.

5.4 CHROMOSOMAL LOCALIZATION OF *HYHBF*

A PCR based analysis using DNA from a panel of human-rodent somatic cell hybrids was carried out in an attempt to map the genomic DNA *hyhbf* sequence to a human chromosome. Touchdown PCR using the primers HYHBF.F and

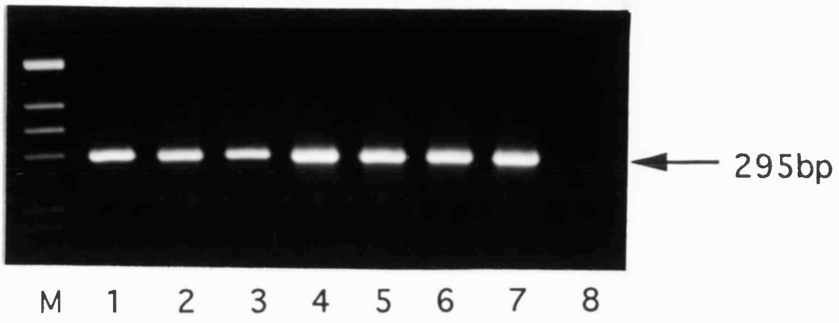


Figure 5.14 PCR of *PGM1* exon 4 from cDNA samples using intron based primers. cDNA samples: lane 1 JG cell line; lane 2 ED cell line; lane 3 K562 cell line; lane 4 6997 cell line; lane 5 Storey cell line; lane 6 human liver; lane 7 Molt4 cell line; lane 8 dH_2O control; M•= molecular weight size marker.

Figure 5.15 Results of RT-PCR experiments with and without reverse transcriptase followed by amplification using *hyhbf* nested primers.

Total RNA Sample	With (W) or Without (W/O) Reverse Transcriptase	Presence (+) or Absence (-) of 261bp Product	
		EXPT 1	EXPT 2
K562 cell line	W	-	+
	W/O	-	-
JG B cell line	W	+	+
	W/O	-	+
TC B cell line	W	+	+
	W/O	+	+
ED B cell line	W	-	+
	W/O	-	+
Storey B cell line	W	-	+
	W/O	-	+
Human Liver	W	-	-
	W/O	-	-
Molt 4 cell line	W	-	-
	W/O	-	-
Human Kidney	W	+	-
	W/O	-	-

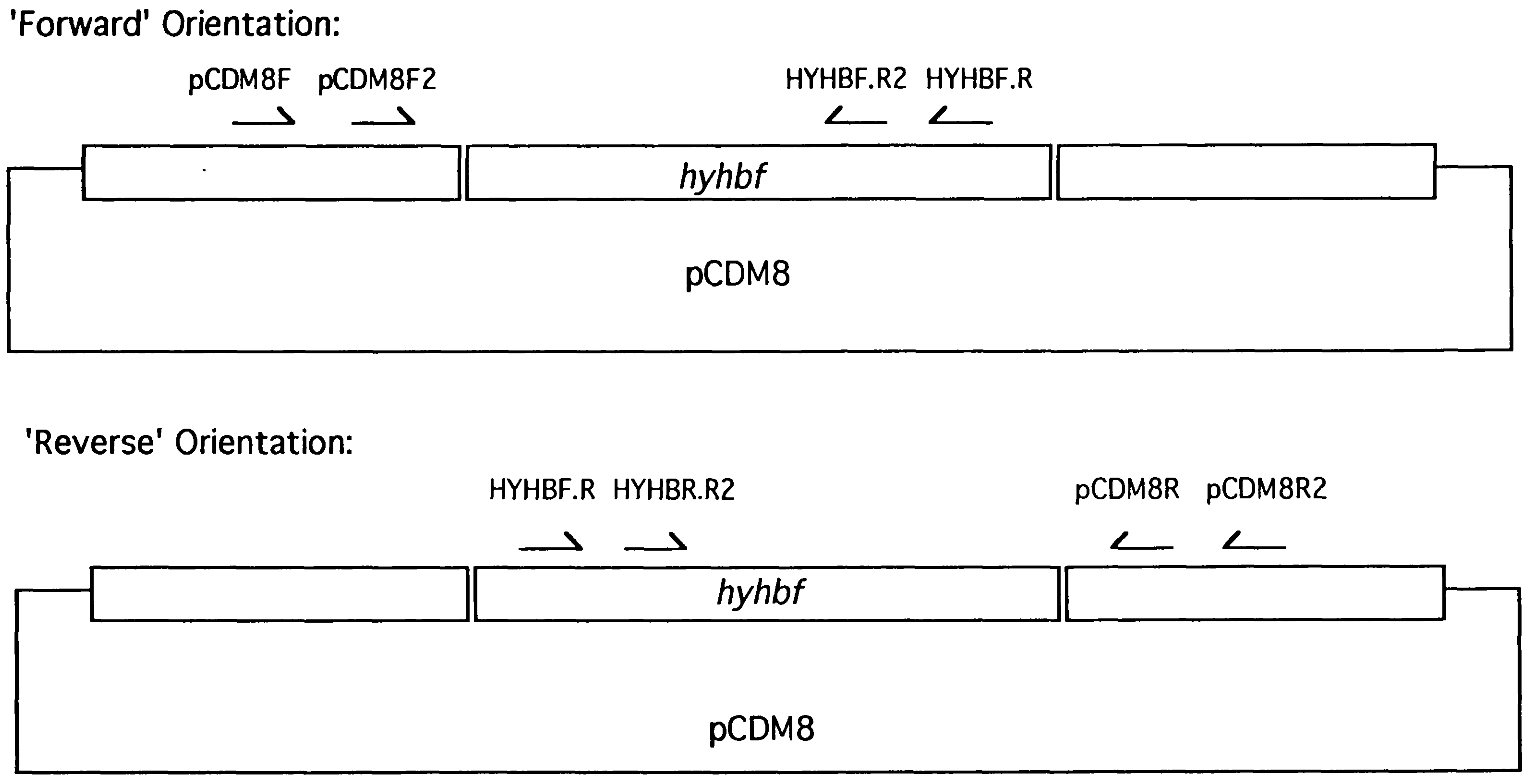


Figure 5.16 Nested primer PCR strategy to amplify *hyhbf* from the cDNA libraries. pCDM8 primers are sited in the multiple cloning site of the vector.

HYHBF.R was carried out. A 312bp PCR product was generated from all three rodent parents (rat, mouse, hamster) and a panel of hybrids (MCP6 [Fr. 6]; HHW416 [4]; SIF4A24E1 [4, 17, 21, X]; GM10478 [4, 6, 10, 20]; F4Sc13Cl12 [1p, 6, 7, 7, 13, 14]) (Figure 5.17). This indicated that the nucleotide sequence of *hyhbf* is highly conserved between rodents and humans.

Single strand conformational analysis (SSCA), which is capable of resolving PCR products of up to about 400bp that differ by only a single nucleotide, was employed in an attempt to distinguish between the human and the rodent parents. However, no differences in the electrophoretic patterns of the human and rodent ssDNAs were observed. Therefore, PCR products from a human control and the rodent parents were sequenced with a view to finding nucleotide differences which may lead to the identification of diagnostic restriction endonuclease recognition sites, or enable the development of an allele specific PCR method for mapping. Surprisingly, the nucleotide sequences obtained from human, rat, mouse and hamster *yhbf* PCR products were identical and raised the possibility that the rodent and hybrid DNA samples had become contaminated with either the *hyhbf* clone or *hyhbf* PCR products.

To investigate whether the plasmid HYHBF clone was responsible, PCR analysis was carried out using the M13 (-40) forward primer, from the pCRII vector, and HYHBF.F2: a PCR product of 494bp would be expected if the clone was present. DNA from the rodent parents and three leucocyte DNA samples were amplified. In addition, a further three leucocyte DNA samples which had not been opened previously were used as "clean" controls. The only band of approximately 494bp was seen in the rat parent, the other rodent samples and human controls were negative (Figure 5.18). Therefore it is concluded that the samples had not been exposed to the clone, with the possible exception of the rat sample. Non-specific PCR products of variable intensity and size were observed in all samples, these spurious bands may have resulted from the M13 or HYHBF.F2 sequences acting simultaneously as both forward and reverse primers, but this was not confirmed.

The possibility of *hyhbf* PCR products contaminating the DNA samples, either by transfer from the pipette or in the air was investigated. The pipette used for handling PCR products was used to add water to standard PCR reaction mixes in which no DNA was added and a second set of no DNA controls were opened for 5secs in the laboratory to check for aerial contamination. Following PCR, no products were observed by ethidium bromide staining. Therefore, *hyhbf* PCR products were demonstrated not to be contaminants.

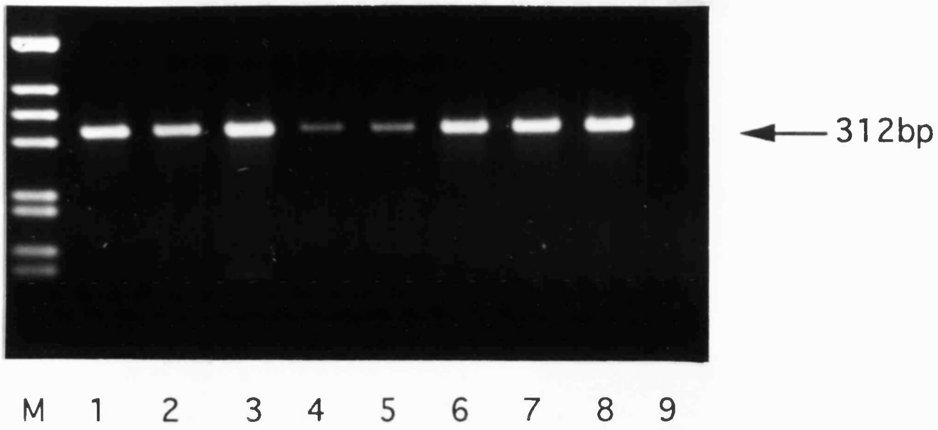


Figure 5.17 Touchdown PCR of human-rodent somatic cell hybrids and rodent parents DNA samples using HYHBF.F and HYHBR.R. DNA samples: lane 1 rat parent; lane 2 mouse parent; lane 3 hamster parent; lane 4 MCP6 ; lane 5 HHW416; lane 6 GM10478; lane 7 F4Sc13Cl12; lane 8 SIF4A24E1; lane 9 dH_2O control; M = molecular weight size marker.

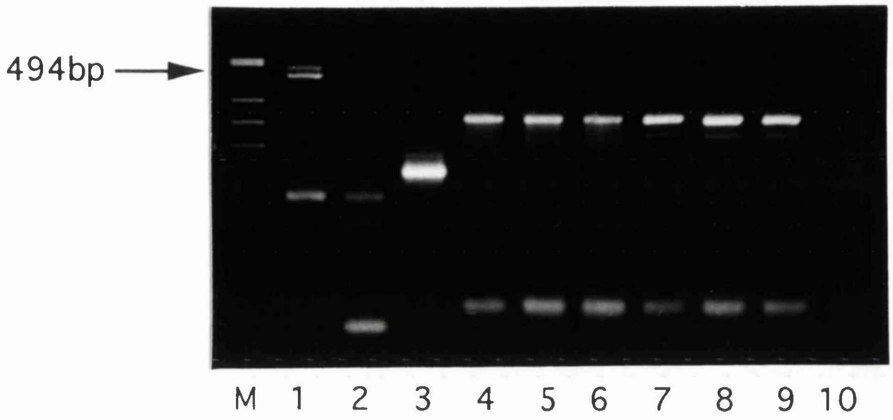


Figure 5.18 PCR products amplified from DNA samples using HYHBF.F2 and M13 (-40) forward primer to detect contamination with the *hyhbf* clone. Lane 1 faza - rat parent DNA; lane 2 rag - mouse parent DNA; lane 3 a23 - hamster parent DNA; lane 4 leucocyte DNA M80; lane 5 leucocyte DNA N1; lane 6 leucocyte DNA N2; lane 7 control leucocyte DNA N16; lane 8 leucocyte DNA N17; lane 9 leucocyte DNA N20; lane 10 dH_2O control; M = molecular weight size marker.

5.5 SOUTHERN BLOT ANALYSIS OF *HYHBF*

Southern blot analysis of genomic DNA was attempted in order to demonstrate the presence of the *hyhbf* sequence in the human genome. In addition to several leucocyte samples, genomic DNA from the K562, JG, ED and TC cell lines was digested with *Eco*RI and *Rsa*I. The HYHBF probe was prepared by excising the insert from the pCRII vector by digestion with *Bst*XI. Following a purification step, the probe was ^{32}P-labelled by the random priming method. At the first attempt, the efficiency of ^{32}P-dCTP incorporation was 45% and no hybridization signals were obtained from the Southern blot after prolonged exposure. Several adjustments to probe purification and labelling reactions were tried to improve the efficiency of labelling (Figure 5.19). The highest labelling efficiences of 50%, were obtained from using the Rediprime DNA labelling system. Although excellent results from Southern blots of PCR products had been obtained throughout this study, blots of genomic DNA digested with *Eco*RI, *Rsa*I, and rare cutter enzymes remained consistently negative.

The low incorporation of ^{32}P-dCTP in the probes, and the negative results of genomic Southern blots, may be due to the high G+C content of the hyhbf sequence which could favour the formation of highly stable single stranded DNA structures. (Computer analysis using the MFOLD program, in the GCG package, indicated that this may be the case for the *hyhbf* sequence) Such stable secondary structures may be very poor templates for labelling reactions and also cause problems during probe hybridisation.

5.6 SUMMARY

i) The complete *hyhbf* insert sequence was obtained. The cloned sequence is 448bp in length and encodes an ORF of 149 amino acids. The nucleotide sequence has a high G+C content of 0.61, compared to the average of 0.40 for coding DNA in man.

ii) The sequence shows unexpectedly high conservation with the *E.coli yhbf* gene at both the nucleotide (64.6% identity) and amino acid level (61.6% identity). Comparison of the hyhbf peptide sequence with human PGM1 shows 29.6% identity.

Figure 5.19 Results from labelling the HYHBF probe.

Probe[a]	Amount of DNA (ng)	Labelling kit[b]	Random Primed[c]	Temp (°C)	Time	Incorporation Efficiency (%)
A	25	M	Yes	RT	4hrs	9
A	25	M	Yes	RT	5hrs	9
A	25	R	Yes	37	45mins	44
A	25	R	Yes	37	20mins	17
B	30	R	Yes	37	2hrs	20
B	15	R	Yes	37	90mins	13
C	20	R	Yes	37	10mins	38
C	10	R	Yes	37	10mins	29
C	20	R	Yes	37	10mins	43
C	10	R	Yes	37	10mins	40
D	25	R	Yes	37	10mins	23
E	20	R	Yes	37	10mins	17
C	20	R	Yes	37	10mins	9
C	20	R	Yes	RT	2hrs	50
C	20	R	Yes	37	10mins	17
C	20	R	Yes	RT	5hrs	50
C	20	R	Yes	RT	2hrs	50
C	20	M	No	37	45mins	25
C	20	R	No	RT	2hrs	50
C	20	R	Yes	RT	4hrs	60
C	20	R	Yes	RT	4hrs	50
C	20	M	Yes	RT	4hrs	71
C	20	M	Yes	RT	4hrs	33

[a]The HYHBF probes A, B and C were prepared by spinning the DNA through glasswool, D by electroelution and E by Wizard PCR preps DNA purification system.

[b]Multiprime (M) or Rediprime (R) DNA Labelling System (Amersham) was used.

[c]Random primers were routinely used, although the PCR primers HYHBF.F2 and HYHBF.R were also tested.

iii) Nested RT-PCR was required to amplify the sequence from most RNA samples. Nested PCR also produced the expected size band from the K562 cDNA library. However, amplification of the *hyhbf* sequence using primers designed to the multiple cloning site in combination with the *hyhbf* reverse primers proved negative.

iv) Touchdown PCR amplified the same size band from genomic DNA as cDNA; evidence regarding the presence of DNA in the RNA preparations suggests that the bands amplified in the RT-PCR experiments originate from DNA rather than RNA.

v) The chromosomal localization of the *hyhbf* sequence by SSCA and restriction endonuclease analysis could not be achieved due to the human and rodent homologues being identical.

vi) Southern blot analyses of *hyhbf* were negative. This was probably due to the high G+C content of the probe, which led to low ^{32}P-dCTP incorporation efficiencies and problems during probe hybridization.

5.7 CONCLUSIONS

The *hyhbf* sequence represents a partial cDNA encoding an ORF of 149 amino acids. The hyhbf peptide shows, in addition to the active site and magnesium binding loop motifs, conservation of a number of amino acid residues with human PGM1. Multiple sequence alignment of a selection of eukaryotic and prokaryotic phosphohexomutases and hyhbf illustrates that these residues appear to be conserved throughout this gene family. Therefore, the conservation of these residues encoded by *hyhbf* suggests that this sequence is also a member of the gene family. Conservation of amino acids is found primarily among those residues located on the internal surfaces of the protein, with the non-polar nature of the side chains highly conserved (Creighton, 1993). Residues at reverse turns, primarily glycines and prolines, are also generally conserved in homologous proteins to maintain the tertiary structure. This is demonstrated in hyhbf, with a number of the conserved glycines and prolines between hyhbf and human PGM1 located at the beginning or end of an α-helix of PGM1. Although the multiple alignments identify three major regions of deletions in hyhbf and the other bacterial phosphohexomutases, these were found to correspond to a region of looping peptide and two ends of an a-helix in human PGM1, suggesting that these regions are not as constrained to evolutionary change.

The molecular characterization of this sequence was inconclusive. The expected size PCR product of *hyhbf* was amplified from the K562 cDNA suggesting it was a transcribed sequence. However, when vector arm primers were used in combination with the hyhbf primers, no PCR products were obtained. In addition, RT-PCR experiments were set up to determine the role of reverse transcriptase. Although the results were not reproducible, it was evident that the amplification of the expected 261bp band was independent of the presence of reverse transcriptase. Therefore, it is suggested that this sequence may not be transcribed, despite the initial template DNA for the PCR reaction being derived from a sample of total RNA.

Since the same size band is amplifiable both from RNA and DNA, and the RNA requires nested primers to be observed, it seemed probable that the sequence originated from genomic DNA. It has been shown through the amplification of exon 4 from *PGM1*, using intronic primers, that DNA is present in the RNA samples. Therefore, two possiblities exist; it may either be a pseudogene or intergenic DNA. If it were a pseudogene, the primers could either be amplifying a transcribed pseudogene or a large exon within a pseudogene. However, both of these possiblities and the intergenic DNA theory are not supported by the apparently identical nucleotide sequence observed in the rodent homologues. Confirmation of the sequence as human in origin would be easily demonstrated by Southern blot analysis. However, due to the high G+C content of the sequence, which was thought to form a highly stable single stranded structure, Southern blot data was not obtained.

CHAPTER SIX

DATABASE SEARCH FOR MEMBERS OF THE PGM GENE FAMILY

An alternative strategy for identifying other members of the PGM gene family utilized the expressed sequence tags (ESTs) databases available through the HGMP Resource Centre. ESTs represent the most rapidly expanding source of novel human sequences; by 1995, up to 25,000 human genes were represented by EST sequences, in addition to the 5,100 genes characterized and submitted to Genbank (Adams et al, 1995). ESTs are generated by single-pass, partial sequencing of cDNA clones from one or both ends, providing 300-400bp of nucleotide sequence from expressed genes. The nucleotide sequences are submitted to the Genbank and EMBL databases, and additionally to dbEST (Boguski et al, 1993). dbEST is a specialized database for ESTs, as in addition to the complete report which is submitted by the contributor, the database includes periodically updated information on homology searches between the EST and the nucleotide and protein databases.

6.1 IDENTIFICATION OF PGM-RELATED ESTs

The EST sequences in EMBL and Genbank were searched using the entire PGM1 amino acid sequence as well as the highly conserved protein motifs associated with PGM catalytic activity; the active site peptide GII**LTASHNP** and the magnesium binding loop GAAF**DGDGDR**. The tblastn option of the searching programme 'blast' compares the protein query sequence against the nucleotide sequence database dynamically translated in all six reading frames.

Numerous human clones were identified by both the full length probe and the active site motif. The majority of these were subsequently identified as human *PGM1*. However, three other human EST clones were identified: when the nucleotide sequences were translated, two showed complete conservation of the active site (human ESTI and human ESTII), and the third contained an Asn^{118} to Cys^{118} amino acid substitution. This EST has subsequently been identified as part of the cDNA for PGMRP (PGM-related protein), previously known as aciculin (Moiseeva et al, 1996).

Human ESTI (clone c-0qg02) was isolated from an infant brain cDNA library. The nucleotide sequence was translated and found to encode an ORF of 111 amino acids (Figure 6.1). Comparison of this sequence with human PGM1

Figure 6.1 The 5' nucleotide sequence of human ESTI (clone c-0qg02) encoding an ORF of 111 amino acids. The conserved active site peptide is shown in bold.

```
    AGCAGCAAAGGCATCGTGATCAGTTTTGACGCCCGAGCTCATCCATCCAGTGGGGGTAGC
  1 ---------+---------+---------+---------+---------+---------+ 60
    TCGTCGTTTCCGTAGCACTAGTCAAAACTGCGGGCTCGAGTAGGTAGGTCACCCCCATCG

    S  S  K  G  I  V  I  S  F  D  A  R  A  H  P  S  S  G  G  S

    AGCAGAAGGTTTGCCCGACTTGCTGCAACCACATTTATCAGTCAGGGGATTCCTGTGTAC
 61 ---------+---------+---------+---------+---------+---------+ 120
    TCGTCTTCCAAACGGGCTGAACGACGTTGGTGTAAATAGTCAGTCCCCTAAGGACACATG

    S  R  R  F  A  R  L  A  A  T  T  F  I  S  Q  G  I  P  V  Y

    CTCTTTTCTGATATAACGGCAACCCCCTTTGTGCCCTTCACAGTATCACATTTGAAACTT
121 ---------+---------+---------+---------+---------+---------+ 180
    GAGAAAAGACTATATTGCCGTTGGGGGAAACACGGGAAGTGTCATAGTGTAAACTTTGAA

    L  F  S  D  I  T  A  T  P  F  V  P  F  T  V  S  H  L  K  L

    TGTGCTGGAATCATGATAACTGCATCTCACAATCCAAAGCAGGATAATGGTTATAAGGTC
181 ---------+---------+---------+---------+---------+---------+ 240
    ACACGACCTTAGTACTATTGACGTAGAGTGTTAGGTTTCGTCCTATTACCAATATTCCAG

    C  A  G  I  M  I  T  A  S  H  N  P  K  Q  D  N  G  Y  K  V

    TATTGGGATAATGGAGCTCAGATCATTTCTCCTCACGATAAAGGGATTTCTCAAGCTATT
241 ---------+---------+---------+---------+---------+---------+ 300
    ATAACCCTATTACCTCGAGTCTAGTAAAGAGGAGTGCTATTTCCCTAAAGAGTTCGATAA

    Y  W  D  N  G  A  Q  I  I  S  P  H  D  K  G  I  S  Q  A  I

    GAAGAAAATCTAGAACCGTGGCCTCAAGCTTGGG
301 ---------+---------+---------+---- 334
    CTTCTTTTAGATCTTGGCACCGGAGTTCGAACCC

    E  E  N  L  E  P  W  P  Q  A  W
```

revealed that they were 29.7% identical and 47.7% similar at the amino acid level (Figure 6.2). Since an identity between two protein sequences of more than 20% probably indicates homology (Creighton, 1993), this suggests that the human ESTI may represent a member of the PGM gene family. The dbEST entry for the human ESTI contained sequences for both ends of the clone. The 5' nucleotide sequence was the sequence which encoded the TASHNP peptide and the 3' nucleotide sequence appeared to be the 3'UTR: nineteen thymine bases had been removed prior to submission of the 3' sequence to the database, suggesting the presence of a poly A tail, and there was no ORF.

In addition to the human EST clones, a pig EST clone encoding the peptide TASHNP was identified from a small intestine cDNA library. This clone was found to be orthologous to human ESTI. The sequences were 87.8% identical at the nucleotide level and 92.7% identical at the amino acid level (Figure 6.3). Identification of orthologous sequences from two species, in two tissue types, each independently isolated, was thought to provide good evidence that human ESTI was a transcribed gene, rather than a cloning artefact.

Human ESTII (clone 55g09) was identified from a T-lymphoblastoid cell line and encodes an ORF of 85 amino acids (Figure 6.4). The dbEST entry only contains sequence for the 5' end of the cloned cDNA.

The magnesium binding loop peptide, GAAFDGDGDR, identified a single human EST, which was found to be human PGM1.

6.2 CHARACTERIZATION OF HUMAN ESTI

The research reported in this chapter was carried out during the final stages of my PhD and forms part the initial stages of a project which is being continued by other members of the PGM research group. The strategies employed here to investigate the human ESTI sequence are expected to be applied to human ESTII and any other candidate EST clones which are identified in the future, in the continuing search for members of the PGM gene family.

6.2.1 RT-PCR ANALYSIS

The expression of the gene including the human ESTI sequence was investigated using an RT-PCR strategy. EST.F and EST.R primers were designed from the 5' EST sequence, (Figure 6.5) and used to amplify cDNA derived from human liver and the cell lines Storey, Molt4 and K562. In each of

Figure 6.2 Amino acid sequence comparison of human PGM1 and the translation of the 5' sequence of human ESTI. The active site peptide is shown in bold.

```
                   .         .         .         .         .
PGM1   42 SIISTVEPAQRQEATLVVGGDGRFYMKEAIQLIARIAAANGIGRLVIGQN 91
          |  :.| . : .. .  .|:. || . .|. :|     ..||. .:::
ESTI    1 SSKGIVISFDARAHPSSGGSSRRFARLAATTFI.....SQGIPVYLFS.. 43

                   .         .         .         .         .
PGM1   92 GILSTPAVSCIIRKIKAIGGIILTASHNPGGPNGDFGIKFNISNGGPAPE 141
          :| .|| |. .:..:| .:||::|||||| ..||     :.:  :..| :
ESTI   44 DITATPFVPFTVSHLKLCAGIMITASHNPKQDNG.....YKVYWDNGA.Q 87

                   .         .         .         .
PGM1  142 AITDKIFQISKTIEEYAVCPDLKVDLGVLGKQQFDLENKFKPF 184
           |...  .||..|||                   :|| . .::
ESTI   88 IISPHDKGISQAIEE...................NLEPWPQAW 111
```

Similarity = 47.7%
Identity = 29.7%

Figure 6.3 Amino acid sequence comparison of translations of the 5' sequences of human ESTI and pig EST. The active site peptide is shown in bold.

```
                   .         .         .         .         .
Human  16 SSGGSSRRFARLAATTFISQGIPVYLFSDITATPFVPFTVSHLKLCAGIM 65
          .||||||||||||||.|||||||||||  ||:|||||:||||||||||||
Pig     1 ESGGSSRRFARLAATPFISQGIPVYLFXXITPTPFVPYTVSHLKLCAGIM 50

                   .         .         .         .
Human  66 ITASHNPKQDNGYKVYWDNGAQIISPHDKGISQAIEENLEPWPQAW 111
          |||| ||||||||||||||||||||||||||.||||:|||||| ||
Pig    51 ITASXNPKQDNGYKVYWDNGAQIISPHDKGIAQAIEGNLEPWPXAW 96
```

Similarity = 92.7%
Identity = 89.6%

Figure 6.4 The 5' nucleotide sequence of human ESTII (clone 55g09) encoding an ORF of 85 amino acids. The conserved active site peptide is shown in bold.

```
    GGCCGGAACTGTCTTTTGCTGTGCGAGAATTGGGGACATTTGCTGGTATCATGATTACGG
  1 ---------+---------+---------+---------+---------+---------+ 60
    CCGGCCTTGACAGAAAACGACACGCTCTTAACCCCTGTAAACGACCATAGTACTAATGCC

     P  E  L  S  F  A  V  R  E  L  G  T  F  A  G  I  M  I  T  A

    CATCACACAATCCCAAGGNATACAATGGCTATAAGGTTTATGGTGAAGATGGTGGCCAAA
 61 ---------+---------+---------+---------+---------+---------+ 120
    GTAGTGTGTTAGGGTTCCNTATGTTACCGATATTCCAAATACCACTTCTACCACCGGTTT

     S  H  N  P  K  ?  Y  N  G  Y  K  V  Y  G  E  D  G  G  Q  M

    TGGTACCGGAAGCCGTTGATGCGGTTGTTAACGAATTAGCGGGCATTTCTGATATCTTTA
121 ---------+---------+---------+---------+---------+---------+ 180
    ACCATGGCCTTCGGCAACTACGCCAACAATTGCTTAATCGCCCGTAAAGACTATAGAAAT

     V  P  E  A  V  D  A  V  V  N  E  L  A  G  I  S  D  I  F  N

    ATATTGCCCTTGATGAAGACCAAACTTACGTTGAAGTGATTGATCANGCCATTGACGAGC
181 ---------+---------+---------+---------+---------+---------+ 240
    TATAACGGGAACTACTTCTGGTTTGAATGCAACTTCACTAACTAGTNCGGTAACTGCTCG

     I  A  L  D  E  D  Q  T  Y  V  E  V  I  D  ?  A  I  D  E  Q

    AATATTTGGCAGCTATG
241 ---------+------- 257
    TTATAAACCGTCGATAC

     Y  L  A  A  M
```

Figure 6.5 The 5' nucleotide sequence of human ESTI illustrating the location of primers.

```
    AGCAGCAAAGGCATCGTGATCAGTTTTGACGCCCGAGCTCATCCATCCAGTGGGGGTAGC
  1 ---------+---------+---------+---------+---------+---------+ 60
    TCGTCGTTTCCGTAGCACTAGTCAAAACTGCGGGCTCGAGTAGGTAGGTCACCCCCATCG

                           EST.F2
    AGCAGAAGGTTTGCCCGACTTGCTGCAACCACATTTATCAGTCAGGGGATTCCTGTGTAC
 61 ---------+---------+---------+---------+---------+---------+ 120
    TCGTCTTCCAAACGGGCTGAACGACGTTGGTGTAAATAGTCAGTCCCCTAAGGACACATG

           EST.F
    CTCTTTTCTGATATAACGGCAACCCCCTTTGTGCCCTTCACAGTATCACATTTGAAACTT
121 ---------+---------+---------+---------+---------+---------+ 180
    GAGAAAAGACTATATTGCCGTTGGGGGAAACACGGGAAGTGTCATAGTGTAAACTTTGAA

    TGTGCTGGAATCATGATAACTGCATCTCACAATCCAAAGCAGGATAATGGTTATAAGGTC
181 ---------+---------+---------+---------+---------+---------+ 240
    ACACGACCTTAGTACTATTGACGTAGAGTGTTAGGTTTCGTCCTATTACCAATATTCCAG
             EST.R2                                        EST.R

    TATTGGGATAATGGAGCTCAGATCATTTCTCCTCACGATAAAGGGATTTCTCAAGCTATT
241 ---------+---------+---------+---------+---------+---------+ 300
    ATAACCCTATTACCTCGAGTCTAGTAAAGAGGAGTGCTATTTCCCTAAAGAGTTCGATAA

    GAAGAAAATCTAGAACCGTGGCCTCAAGCTTGGG
301 ---------+---------+---------+---- 334
    CTTCTTTTAGATCTTGGCACCGGAGTTCGAACCC
```

the samples, an intensely staining PCR product of the expected size, 128bp, was obtained (Figure 6.6). This indicates that the gene may be constitutively expressed as the mRNA is present in liver, a fibroblast cell line (Storey), a lymphoid cell line (Molt4) and the erythroleukaemic cell line (K562).

Amplification of genomic DNA from the same sources under identical conditions did not produce a 128bp product, (section 6.2.4), and therefore, amplification of the EST sequence is derived from RNA rather than genomic DNA.

6.2.2 SOUTHERN BLOT ANALYSIS

Southern blot analysis was carried out to investigate the human ESTI gene in the human genome; for example, whether the sequence was present in single or multiple copies. The RT-PCR product was reamplified and the DNA purified to provide a 128bp probe. The incorporation efficiency of ^{32}P-dCTP for this probe was 75%. Genomic DNA from K562 and leucocytes was digested with *Eco*RI, *Taq*I and *Msp*I, none of which have recognition sites present in the human ESTI sequence. Following hybridization with the human ESTI probe, the filters were washed to high stringency (0.1 x SSC, 0.1% SDS). After 2 weeks autoradiography, distinct hybridization signals were evident (Figure 6.7).

DNAs digested with *Msp*I showed a single band of 2.5kb. However, DNAs digested with *Taq*I and *Eco*RI showed four and three bands respectively. In the *Taq*I digests, of leucocycte DNA, there were two stronger hybridization signals of 2.2 and 1.7kb and two weaker signals of 6.9 and 0.9kb. The two stronger hybridization signals suggest that there is an intron present in the sequence covered by the probe. The two weaker hybridization signals suggest the presence of a closely related gene. This idea is supported by the result of the *Eco*RI digestion, in which the leucocyte DNA samples show strong hybridization signals of 7.5 and 6.6kb band and a less intense 7.9kb band.

The disparity observed in the hybridization signals from K562 DNA compared to the leucocyte DNAs may reflect the copy number of the sequence/gene(s). K562, although generally triploid, shows many chromosomal rearrangements (Fox et al, 1996), and therefore not all genes may be represented equally. This hypothesis can be investigated once suitable genomic probes have been obtained, as these can be used for fluorescence in-situ hybridization of K562 metaphase spreads.

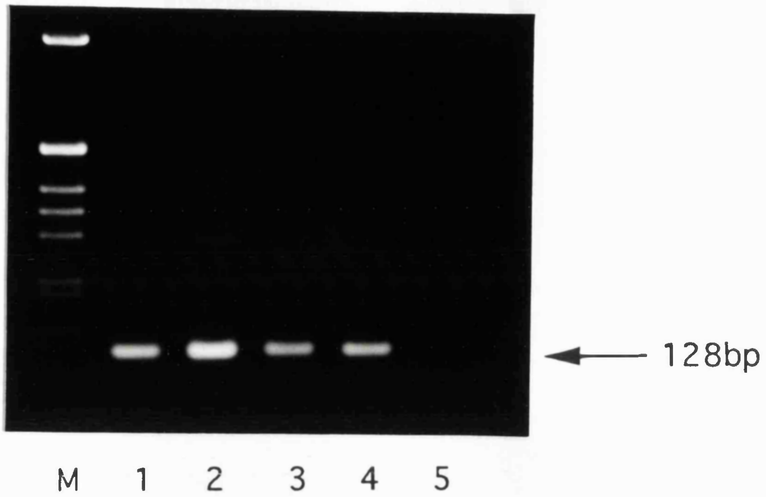


Figure 6.6 RT-PCR products amplified from total RNA using EST.F and EST.R. Lane 1 Storey cell line; lane 2 human liver; lane 3 K562 cell line; lane 4 Molt4 cell line; lane 5 dH_20 control; M = molecular weight size marker.

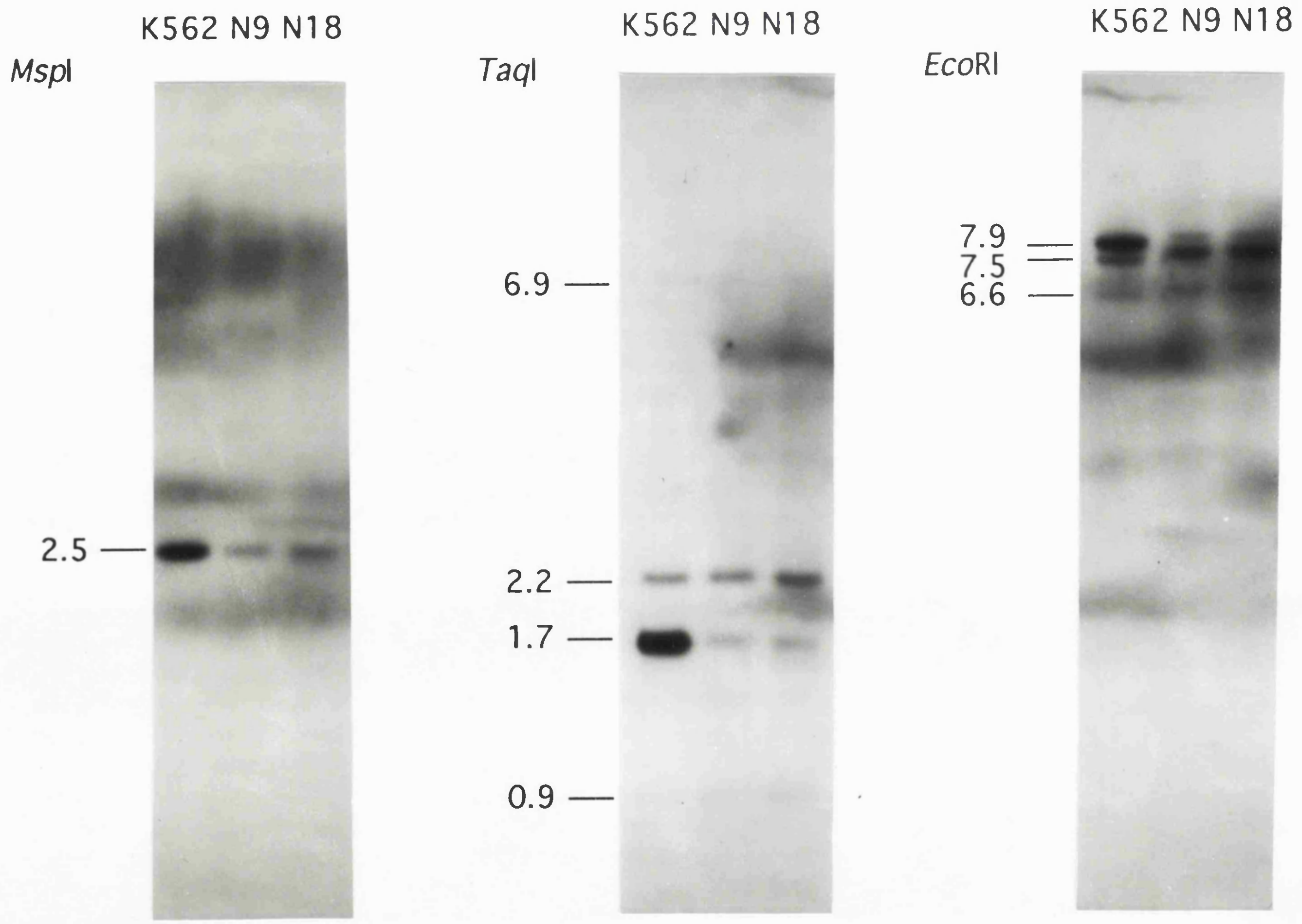


Figure 6.7 Southern blot analysis using the human ESTI RT-PCR product as probe. K562 and leucocyte DNAs were digested with *Msp*I, *Taq*I and *Eco*RI.

6.2.3 NORTHERN BLOT ANALYSIS

Northern blot analysis was carried out to investigate if human ESTI was widely expressed, as suggested by the RT-PCR results, and to determine the size of the transcript. Northern blots (Clontech) were hybridized with the human ESTI probe by J. Lovegrove. After high stringency washing (0.1 x SSC, 0.1% SDS) and 5 days autoradiography, four hybridization signals were observed (Figure 6.8). The transcripts are approx. 4.5kb, 2.4kb 1.6kb and 1.35kb. All four bands are present in heart, brain, liver, skeletal muscle, kidney and pancreas. Only the 1.6kb and 1.35kb bands appears to be present in placenta, whilst the 1.35kb and the 4.5kb transcript are present in lung and liver. These hybridization signals may represent either alternative transcripts from a single gene or transcripts from related genes, possibly those observed by Southern blot analysis.

6.2.4 GENOMIC DNA PCR ANALYSIS

Standard genomic DNA PCR was carried out on leucocyte DNA using the EST.F and EST.R primers in an attempt to amplify related genomic sequences. Amplification of a 128bp product would indicate the primers are sited within an exon. However, if they are separated by a small intron, a larger PCR product may be obtained which could be used as a probe for fluorescence in-situ hybridization, to allow chromosomal localization of human ESTI.

Following amplification, numerous PCR products were obtained (Figure 6.9a). Optimization of the PCR was attempted by using the touchdown procedure, but there was no improvement in the specificity. Southern blot analysis of the PCR products revealed no highly specific band of hybridization, under high stringency washing conditions. The low level of hybridization signals observed following 3 days autoradiography (Figure 6.9b) were thought to be due to the probe hybridizing to the primer sequences.

If *PGM1* and the human ESTI genes are members of a gene family, it is possible that the intron/exon structure may be conserved between the two genes. The position of the introns with respect to the site of the primers may indicate the size of band which could be expected, or, if the primer is sited over an intron/exon boundary, explain why no specific PCR product is obtained. As can be seen from figure 6.10, if the structure was conserved, the expected band size from genomic DNA would be 128bp, the two primers being sited within the corresponding exon 2 of *PGM1.* However, no band of this size was

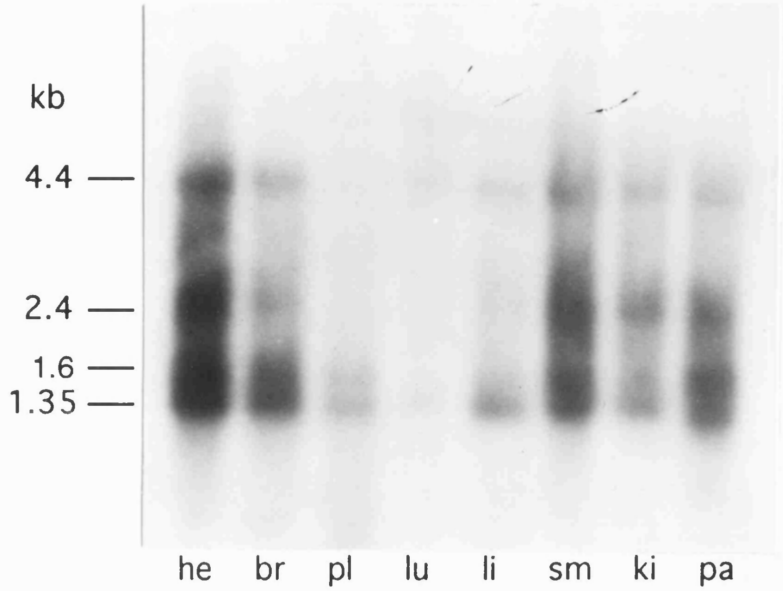


Figure 6.8 Northern blot analysis of human tissue samples using the human ESTI RT-PCR product as probe. Tissues samples are: he = heart; br = brain; pl = placenta; lu = lung; li = liver; sm = skeletal muscle; ki = kidney; pa = pancreas.

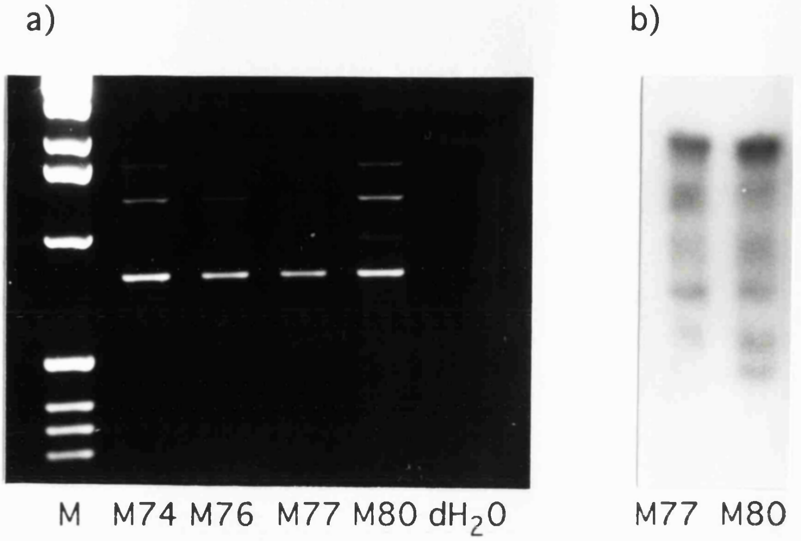


Figure 6.9 Genomic DNA PCR of leucocyte DNA with EST.F and EST.R. a) Ethidium bromide stained gel to detect PCR products. b) Southern blot analysis after 3 days autoradiography following hybridization with the human ESTI RT-PCR product. M = molecular weight size marker.

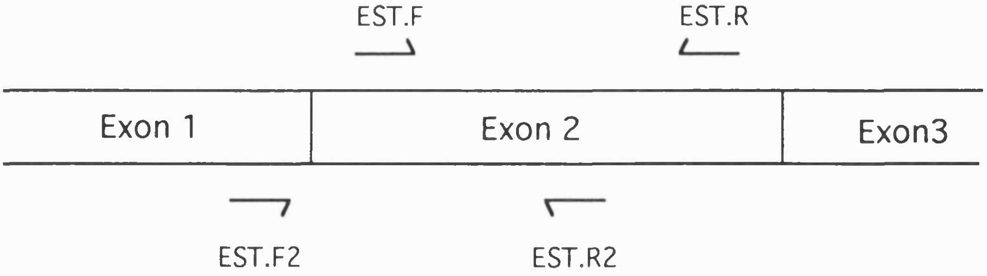


Figure 6.10 Position of putative intron/exon boundaries based on the genomic structure of human *PGM1*. The location of the EST primers are also shown.

amplified from genomic DNA. Thus, the intron/exon structure is not conserved. Therefore, it is not possible to determine if the EST primers are sited over an intron/exon boundary in the human ESTI sequence and whether this is affecting the specificity of the PCR. Alternatively, the primers may be sited either side of a large intron and the PCR conditions are not suitable for amplification of the DNA.

A new set of primers, EST.F2 and EST.R2, were designed (Figure 6.5 & 6.10). These were sited 20-30bp upstream of each of the original primers; if the original primers had covered an intron/exon boundary, these should not. Amplification of leucocyte genomic DNA with EST.F2 and EST.R2 also produced numerous non-specific bands (Figure 6.11a; lanes 1-4). PCR was then carried out using the old EST.F was used in combination with the new EST.R2 (Figure 6.11a; lanes 5-8) and the new EST.F2 with the old EST.R (Figure 6.11a; lanes 9-12). Again, numerous bands were produced. However, Southern blot analysis of these products identified a highly specific doublet following 4hrs autoradiography produced by the EST.F/ESTR2 primers (Figure 6.11b). These products are estimated to be approximately 2.3kb.

6.3 IDENTIFICATION OF YEAST PMM-RELATED EST

In addition to the identification of PGM-related ESTs, a human EST was identified which showed homology to the yeast phosphomannomutase (PMM) (Bernstein et al, 1985; Smith et al, 1992). The yeast *PMM* genes in *Saccharomyces cerevisiae* (*sec53*) and *Candida albicans* (*pmm*) encode a protein of 29,000mw. They do not show any of the characteristic protein motifs encoded by the other cloned *PMM* and *PGM* genes. However, the *sec53* gene product has been shown to function additionally as a phosphoglucomutase. The PMM EST (clone b4HB3MA-COT8-HAP-Ft261) was identified by searching the sequence databases using a keyword, in this case 'phosphomannomutase' (stringsearch, GCG), rather than a peptide sequence. Isolated from human neonate brain, both the 5' and 3' ends of the clone had been sequenced (Figure 6.12).

An alternative strategy for the preliminary RT-PCR experiments was carried out, with the forward primer, PMM.F, designed from the 5' sequence, and the reverse primer, PMM.R, from the 3' sequence (Figure 6.12). If the homologous gene was conserved in man, the expected PCR product would be approximately 450bp. RT-PCR was carried on total RNA from the erythroleukaemic cell line K562, and the lymphoblastoid cell lines 6997 and

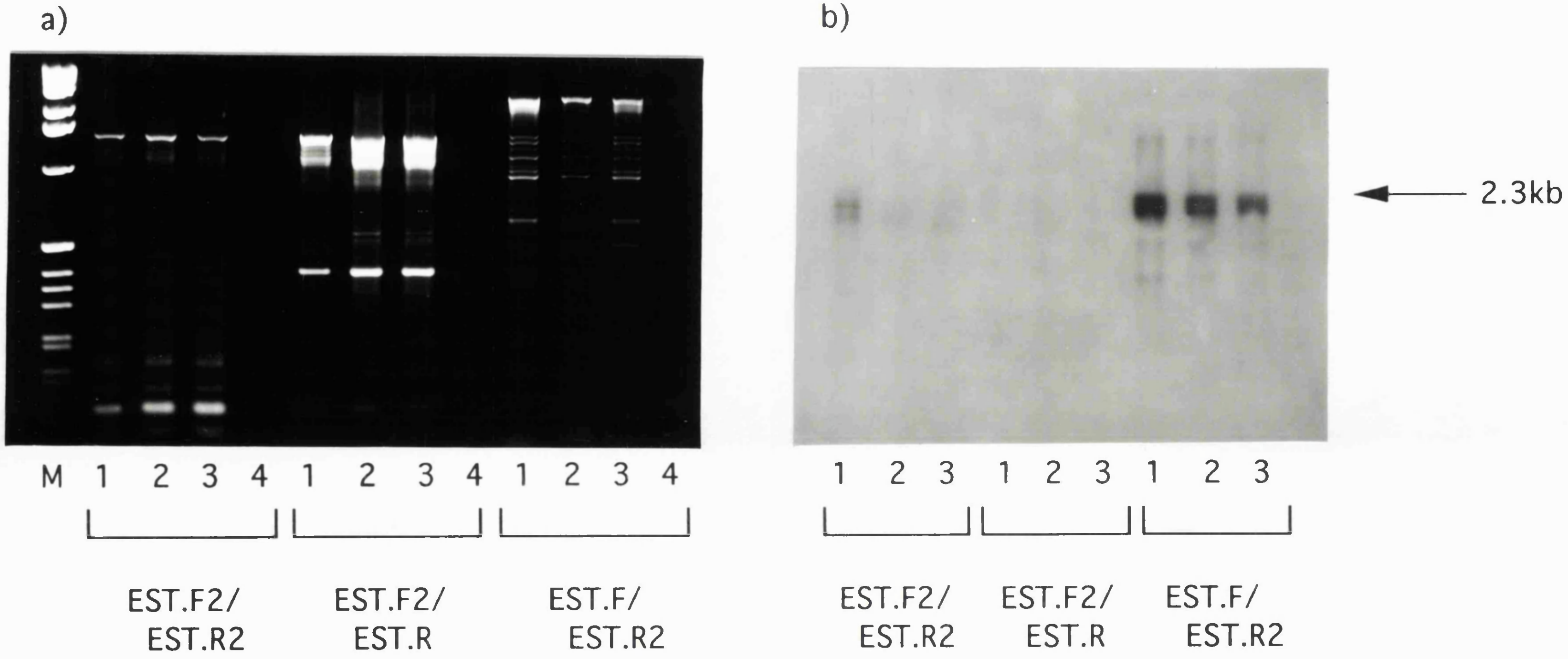


Figure 6.11 Genomic PCR of leucocyte DNA with EST.F2 and EST.R2 and in combination with the primers EST.F and EST.R. a) Ethidium bromide stained gel to detect PCR products. b) Southern blot analysis after 4hrs autoradiography following hybridization with the human ESTI RT-PCR product. Lane 1 M77 DNA; lane 2 M79 DNA; lane 3 M80 DNA; lane 4 dH_2O control; M = molecular weight size marker.

Figure 6.12 Nucleotide sequence of EST homologous to yeast PMM. Location of the RT-PCR primers PMM.F and PMM.R are shown in bold.

```
5' nucleotide sequence

     AAGCTTGGCACGAGGCTCGCAAAGTGTTGGGATTGCAGACCTGAGCCACAGTGTCCAACC
   1 ---------+---------+---------+---------+---------+---------+ 60
     TTCGAACCGTGCTCCGAGCGTTTCACAACCCTAACGTCTGGACTCGGTGTCACAGGTTGG

     TGTCTAATTTTTAGTGTCTAAGCTTTGTACTGCTTCAGATCCAGGTAGAATGTGGCTTC
  61 ---------+---------+---------+---------+---------+---------+ 120
     ACAGATTAAAAATCACAGATTCGAAACATGACGAAGTCTAGGTCCATCTTACACCCGAAG

     CTGGGTTCTCAGCACTAAGTGAGGGCTAAGTGGAGGTCCCAGACATGTTGAAAGCCAGAA
 121 ---------+---------+---------+---------+---------+---------+ 180
     GACCCAAGAGTCGTGATTCACTCCCGATTCACCTCCAGGGTCTGTACAACTTTCGGTCTT

     TGCTATGCTTCCCCTCTCCCCCCATAGAAAATTGACCCTGAGGTGGCCGCCTTCCTGCAG
 181 ---------+---------+---------+---------+---------+---------+ 240
     ACGATACGAAGGGGAGAGGGGGGTATCTTTTAACTGGGACTCCACCGGCGGAAGGACGTC
                                   PMM.F
     AAGCTACGAAGTAGAGTGCAGATCGGTGTGGTGGGCGGCTCTGACTACTGTAAGATCGCT
 241 ---------+---------+---------+---------+---------+---------+ 300
     TTCGATGCTTCATCTCACGTCTAGCCACACCACCCGCCGAGACTGATGACATTCTAGCGA

     GAGCAGCTGGGTGACGGGGATGAAGTCATTGAGAAGTTTGATTATGTGTTTGGCGAGAAC
 301 ---------+---------+---------+---------+---------+---------+ 360
     CTCGTCGACCCACTGCCCCTACTTCAGTAACTCTTCAAACTAATACACAAACCGCTCTTG

     GGGACGGTGCAGTATAAGCACGGACGACTGCTCTCCAAG
 361 ---------+---------+---------+--------......................
     CCCTGCCACGTCATATTCGTGCCTGCTGACGAGAGGTTC

               TTTGTTCTGGGAACTTTAATACTGTGACAAAGTTCTCTAAAATAGGCACC
   1 ..........---------+---------+---------+---------+---------+ 50
               AAACAAGACCCTTGAAATTATGACACTGTTTCAAGAGATTTTATCCGTGG

     TTCCCCACCGTACCTCATCGCCCAGGGCAGGCAGGCAGGGCAGGCTAGATCTCGTACCGA
  51 ---------+---------+---------+---------+---------+---------+ 110
     AAGGGGTGGCATGGAGTAGCGGGTCCCGTCCGTCCGTCCCGTCCGATCTAGAGCATGGCT

     TACTTGAGCACGCCTCCTCCTGGTGCAGAAAGAAACCTCTTCTGTACCGAAATACAAGCA
 111 ---------+---------+---------+---------+---------+---------+ 170
     ATGAACTCGTGCGGAGGAGGACCACGTCTTTCTTTGGAGAAGACATGGCTTTATGTTCGT

     GCAGCTGTGGCCTGGGCCACCAGGTGGAGCATGGGGAACACTCTGGGCCCTGGGAGGACG
 171 ---------+---------+---------+---------+---------+---------+ 230
     CGTCGACACCGGACCCGGTGGTCCACCTCGTACCCCTTGTGAGACCCGGGACCCTCCTGC

     AAGCCAGTGCCACTAGGAGCAGACTGGCTGGGGACGGTTGTCCACACAGACTCTGGCCCC
 231 ---------+---------+---------+---------+---------+---------+ 290
     TTCGGTCACGGTGATCCTCGTCTGACCGACCCCTGCCAACAGGTGTGTCTGAGACCGGGG

     ATCTGGGTGGGCTTGCAGCAGGCGTCCTGGGCCAGAGGAGGGGGCCTGGCATCTATCCA
 291 ---------+---------+---------+---------+---------+---------- 349
     TAGACCCACCCGAACGTCGTCCGCAGGACCCGGTCTCCTCCCCCGGACCGTAGATAGGT
          PMM.R                              3' nucleotide sequence
```

7014. No products were amplified as detected by ethidium bromide staining. Further analysis of the translation of the 5' sequence revealed a stop codon upstream of the region of homology. In fact, in all six frames, at least two stop codons were present. It was estimated that the automated single-pass sequencing results in a 3% error or base ambiguity rate (Boguski et al, 1993). For this EST clone, however, 3% may be a conservative estimate.

6.4 SUMMARY

i) The strategy of searching for EST sequences which encode conserved amino acid motifs for the identification of PGM-related genes has been successful. Both the full PGM1 amino acid sequence and the active site peptide probe identified three novel human sequences: human ESTI, human ESTII and an EST which has subsequently been identified as the *PGMRP* gene. A further EST clone originating from pig was also identified and found to be orthologous to human ESTI. The magnesium binding loop peptide probe did not identify any novel PGM-related sequences, only a clone which was identified as *PGM1*.

ii) Preliminary characterization of the human ESTI clone has been carried out. The 5' sequence of the human ESTI clone encodes an ORF of 111 amino acids. Sequence comparison at the protein level with human PGM1 revealed an identity of 29.7% between the two sequences. This suggests a common ancestry; the human ESTI sequence may therefore represent a member of the PGM gene family.

iii) Molecular analysis of human ESTI was carried out at both the RNA and DNA level. RT-PCR of three cell lines and human liver RNA using EST.F and EST.R produced the expected 128bp band. A product of this size was not amplified from genomic DNA extracted from the same samples, indicating that the sequence is derived from RNA, rather than DNA. Northern blot analysis detected up to four distinct transcripts in a variety of tissue types. This may be explained by the occurrence of alternate transcripts and/or related genes. Southern blot analysis indicated that there was a related sequence present. Genomic DNA PCR with EST.F and EST.R2 primers amplified a highly specific 2.3kb product, as observed by hybridization with the human ESTI probe, in addition to a number of non-specific products.

iv) Searching the databases using a keyword, rather than protein sequence, identified an EST clone orthologous to *sec53* and *PMM* of *S.cerevisiae* and

C.albicans respectively. However, following RT-PCR, no products were detected from the cell lines K562, 6997 and 7014.

6.5 CONCLUSIONS

The human ESTI sequence is a candidate member of the PGM gene family; the 5' nucleotide sequence encodes a peptide, including the active site motif, which shows a significant level of identity with human PGM1. The sequence has been shown to be derived from RNA and it is widely expressed, being amplified from a variety of cell lines and detected on Northern blots of numerous tissues, such as kidney, brain and skeletal muscle. Northern blot analysis also suggests the presence of alternative transcripts; four transcripts were observed. The absence of some of these transcripts in placenta, lung and liver may represent some form of regulated expression.

Southern analysis indicates the presence of a closely related sequence to human ESTI in the genome. The identification of this second sequence by hybridization of the 128bp RT-PCR product suggests that the two sequences are greater than 67.6% identical at the nucleotide level over this region. *PGMRP* is 67.6% identical to human *PGM1*, but is not detected by Southern blot analysis using the human *PGM1* cDNA as probe. This second sequence may represent a paralogous gene or a pseudogene. If it is expressed, screening a cDNA library with the 128bp probe should identify cDNA clones representing both human ESTI and the related sequence.

Preliminary mapping data has been obtained by members of the PGM research group for the human ESTI clone. Using primers sited in the 3' UTR of the sequence, a panel of human-rodent somatic cell hybrids were amplified to determine the chromosomal localization of the gene. All the human-rodent hybrids containing human chromosome 4, and the chromosome 4 only hybrid HHW416, consistently produced an intensely staining PCR product. However, amplification of hybrids containing chromosome 7, including the chromosome 7 only hybrid clone 21, produced a low intensity PCR product. Increasing the annealing temperature to improve the specificity of the reaction did not abolish amplification. Thus, this data may suggest that the ESTI related sequence, identified by Southern blot analysis, is located on chromosome 7.

Further evidence to support the localization of human ESTI to chromosome 4 is provided by another human EST clone, 130882, which has been mapped to chromosome 4. The 3' nucleotide sequence of this clone is almost identical to human ESTI, whilst the 5' nucleotide sequence encodes a putative magnesium binding site motif DPDADR. Thus, it is suggested that the gene represented by human ESTI is a candidate for PGM2. The regional localization of the PGM2 locus is 4p14-q12. Further localization of the human ESTI clone will be obtained using human-rodent somatic cell hybrids containing chromosomes with known breakpoints. In addition, clone 130882 is available from the HGMP Resource Centre. It may be possible to use the partial cDNA sequence as a probe for fluorescence in-situ hybridization to determine the exact map postition.

Since the related sequence maps to chromosome 7, it does not represent the third PGM isozyme, PGM3, which maps to chromosome 6. Thus, if this sequence is expressed, it may represent a further phosphohexomutase locus, or possess an alternative, possibly structural, function.

The negative results from the RT-PCR of the yeast PMM-related EST may possibly be due to sequencing error(s) incorporated into the PCR primer(s). Errors at the 3' end of the primer are likely to inhibit amplification. The presence of sequencing errors is supported by the absence of an ORF in the 5' EST sequence. Alternatively, with no ORF present, the status of the sequence as expressed is questionable.

Searching the EST databases with conserved protein motifs appears to be a very powerful and resourceful strategy to identify novel related genes. Since partial cDNA sequence was available, the use of PCR allowed a rapid molecular characterization of the EST clone to be carried out. And as the number of clones submitted to the EST databases continues to increase, this resource should be searched periodically for further novel PGM-related sequences.

CHAPTER SEVEN:

EVOLUTION OF THE *PGM1* GENE IN PRIMATES

PGM1 in man is a highly polymorphic marker at the protein level, the four commonest alleles arising from two mutations and intragenic recombination. A phylogeny for these alleles, put forward by Carter et al, (1979) and Takahashi et al, (1981), suggested the *PGM1**1+ allele is ancestral (section 1.2.2.1). Isozyme studies of Hominoidea great apes (orangutan, gorilla and chimp) showed that, following both starch gel electrophoresis and IEF, the primary isozyme of PGM appears identical in electrophoretic mobility and isoelectric point (pI) to the PGM1*1+ isozyme of man (Carter et al, 1979). An isozyme identical to PGM1*1+ was also found in some Old World and New World Monkeys (langurs, guenons, macaque, marmoset), but not in others (baboon, squirrel monkey and owl monkey). Therefore, the protein data indicates that the emergence of a PGM1*1+ like isozyme predated the division between Old World simians and New World simians (Figure 7.1).

The primary aim of this investigation was to determine if the PGM1*1+ like protein found in the great apes has the same characteristic amino acid substitutions as the human PGM1*1+ isozyme i.e. Arg^{220} and Tyr^{419} (section 1.3.1.3). Nucleotide sequencing of exons 4 and 8, which contain the polymorphic substitutions in human *PGM1*, was carried out on samples from all the great apes, and the amino acid sequence deduced.

In addition, the sites corresponding to the 3' untranslated region (3' UTR) polymorphism in man were investigated in the great apes. In man, this polymorphism exhibits substitutions at three sites in exon 11: nt 1773, nt 1788 and nt 1844 (section 1.3.1.4). On the basis of its high frequency in the British population, it has been proposed that the +++ haplotype (allele 1) is ancestral. In order to assess this view, the presence of the +++ haplotype in the great ape species was investigated.

In the analysis of *PGM1* exons 4, 8 and 11, multiple samples of presumed unrelated individuals were investigated and this allowed a search for common sequence polymorphisms in these exons to be carried out.

The third aim recorded in this chapter was to investigate the levels of nucleotide and amino acid conservation in primate PGM1. Sequences of exons 1A and 5 in chimpanzee, gorilla and orangutan (suborder anthropoids) were compared

Figure 7.1 A guide to primate classification.

SUBORDER	INFRAORDER	SUPERFAMILY	FAMILY	SUBFAMILY
PROSIMII (prosimians)	LEMURIFORMES (lemuriforms)	LEMURIDEA (lemurs)	CHEIROGALEIDAE (mouse and dwarf lemurs)	
			LEMURIDAE	**LEMURINAE (true lemurs)**
				LEPILEMURINAE (sportive lemurs)
			INDRIDAE (indri group) DAUBENTONIIDEA (aye-aye)	
	LORISFORMES (lorisforms)	LORISOIDES (loris group)		LORISINAE (lorises)
				GALAGINAE (bushbabies)
	TARSIIFORMES (tarsiers)	TARSIOIDEA	TARSIIDAE (tarsiers)	
ANTHROPOIDEA (simians or anthropoids)	PLATYRRHINI (New World simians)	CEBOIDEA (New World Monkeys)	CEBIDAE (true monkeys)	CEBINAE (capuchins, etc.)
				AOTINAE (owl monkeys, etc)
				ATELINAE (spider monkeys, etc)
				SAIMIRIINAE (squirrel monkeys)
			CALLITRICHIDAE* (marmosets and tamarins)	

Figure 7.1 cont.

SUBORDER	INFRAORDER	SUPERFAMILY	FAMILY	SUBFAMILY
ANTHROPOIDEA (simians or anthropoids)	CATARRHINI (Old World simians)	CERCOPITHECOIDEA (Old World monkeys)	CERCOPITHECIDEA	CERCOPITHECINA* (macaques, baboons, mandrills, etc.)
				COLOBINAE (leaf monkeys)
		HOMINOIDEA (apes and humans)	HYLOBATIDAE	HYLOBATINAE (gibbons)
			PONGIDAE	**PONGINAE* (orangutans)**
			HOMINIDAE	**GORILLINAE* (gorilla and chimps)**
				HOMINIMAE* (humans)

Subfamilies highlighted indicate primates included in this study.
*Samples which have shown the PGM1*1+ allele on IEF (Carter et al, 1979).

with lemur (suborder prosimians). Exon 1A is the site of the third mutation in *PGM1* which gives rise to the 3/7 alleles found at polymorphic frequencies in some Asian-Pacific populations.

In addition to sequencing data obtained from the primate samples, nucleotide and amino acid sequences of exons 4 , 8, 11, 1A and 5 from rabbit (Whitehouse et al, 1992) and rat (Auger et al, 1993) and of exons 4, 8 and 5 from mouse (Friedman, personal communication) have also been included for comparison.

7.1 PRIMATE SAMPLES

A number of primate samples were available for this analysis, including gorilla (2), chimpanzee (5), orangutan (5) and lemur (1). These samples varied from whole blood and white blood cells to live cells frozen in liquid nitrogen and DNA (Figure 7.2). Blood was heated at 95^{o}C for 10mins, centrifuged at 12,000g and the supernatant collected. Following a washing step in distilled water, the live cells were prepared by the same method. In both cases, 1µl of supernatant was added to the PCR reaction. For primate white blood cell samples, 1µl was used directly in the PCR.

7.2 ISOZYME ANALYSIS OF THE PRIMATE SAMPLES

Isoelectric focusing of either whole blood or white blood cells from the great apes was carried out to determine if they possessed an isozyme equivalent to the PGM1*1+ in man. In all of the samples investigated (Figure 7.2), a two banded pattern, characteristic of the PGM1*1+ primary and secondary isozymes was observed. In addition, a third more cathodal band was observed in one gorilla (Daniel) and in one chimpanzee sample (Halfpenny). These bands may represent polymorphic protein alleles.

7.3 DNA SEQUENCE ANALYSIS

7.3.1 ANALYSIS OF EXONS 4 AND 8

To determine if the molecular basis of the *PGM1**1+ allele is conserved in the three great apes, the nucleotide sequences of exons 4 and 8 was obtained. Exon 4 was amplified with primers E4F and E4R (March et al, 1993a). The complete coding region of exon 4, 47bp of IVS3 and 28bp of IVS4, was subsequently sequenced from man (1), gorilla (2), chimpanzee (5) and

Figure 7.2 Primate samples.

Primate	Name	Source of DNA[a]	Putative PGM1*1+ isozyme	Exons Sequenced
Gorilla	Daniel	WBC	Yes	1A, 4, 5, 8, 11
	Sampson	DNA	-	4, 8, 11
Chimpanzee	Halfpenny	Blood & WBC	Yes	4, 8, 11
	Farthing	Blood & WBC	Yes	4, 8, 11
	Katja	WBC	Yes	1A, 4, 8, 11
	Jane	WBC	Yes	4, 5, 8, 11
	Masikini	DNA	-	4, 8, 11
Orangutans: Bornean	Kate	Blood & WBC	Yes	4, 8, 11
	Blossom	WBC	Yes	1A, 4, 8, 11
	Kibriah	WBC	Yes	4, 5, 8, 11
Sumatran	Annie	WBC	Yes	1A, 4, 5, 8, 11
	Henry	DNA	-	4, 8, 11
Lemur	-	Blood	-	-
	Columbo	Live Cells	-	1A, 5
Human	N14	DNA	-	1A, 4, 5, 8, 11

[a]WBC = white blood cells

orangutan (5). The exon 8 sequence was amplified with E8F2 and E8R2. The complete coding region of exon 8, 23bp of IVS7 and 22bp of IVS8, was subsequently sequenced from man (1), gorilla (2), chimpanzee (5) and orangutan (5).

The nucleotides which form the structural basis of the PGM1 protein type in man were identified. All twelve great ape samples carried a C base at nucleotide 723 in exon 4, (Figure 7.3), conserving the codon **C**GT, which encodes the amino acid Arg^{220}, associated with 1 allele. In exon 8, at nucleotide 1320, the samples carried a T base (Figure 7.5), conserving the codon **T**AT, which encodes the amino acid Tyr^{419}, associated with the + allele. This suggests that the great apes' PGM1*1+ protein type may be identical to that of man and that the ancestral allele is *PGM1**1+.

DNA sequencing of exon 4 and exon 8 detected no nucleotide substitutions which would lead to changes in the amino acid sequence. However, the orangutans did show two nucleotide substitutions in the coding sequence of exon 4, at nt 647 and 707. The second substitution was identified as polymorphic; the Bornean orangutans were all heterozygotes, carrying both G and A bases (Figure 7.4). The Sumatran orangutans were both homozygotes; one homozygous for the G and the other homozygous for the A. The presence of these two alleles in both sub-species of the orangutans suggests the polymorphism was established prior to the events that led to the geographical isolation which exists today.

The other nucleotide substitutions identified in the great apes occurred in intron sequence. In IVS3, in orangutan, three substitutions were seen at nt -35, nt -20 and nt -16. The same substitution at nt-20 is seen in gorilla, in addition to a deletion at nt -14. Both of these changes were also seen in chimpanzee, along with a substitution at nt -30 (Figure 7.6). In IVS7, two nucleotide substitutions were seen at nt -6 and nt -13 in orangutan. A single substitution was seen in chimpanzee at nt -10, whilst the gorilla was identical to man at this position (Figure 7.7). Therefore, in contrast to the conservation of the coding sequences, the introns of each of the great ape species were found to be unique.

Intron data for exons 4 and 8 was not available for rabbit, rat or mouse. As would be expected, these species show more extensive nucleotide substitutions in the coding DNA, primarily at the third base of the codon, although two changes are present, which lead to amino acid substitutions at

Figure 7.3 Autoradiograph of exon 4 nucleotide sequences from gorilla, chimpanzee, orangutan and human (716-744bp). Nucleotide 723, which is the site of the 2/1 polymorphism, is shown in bold. All the primates carry a C residue, associated with the *PGM1**1 allele.

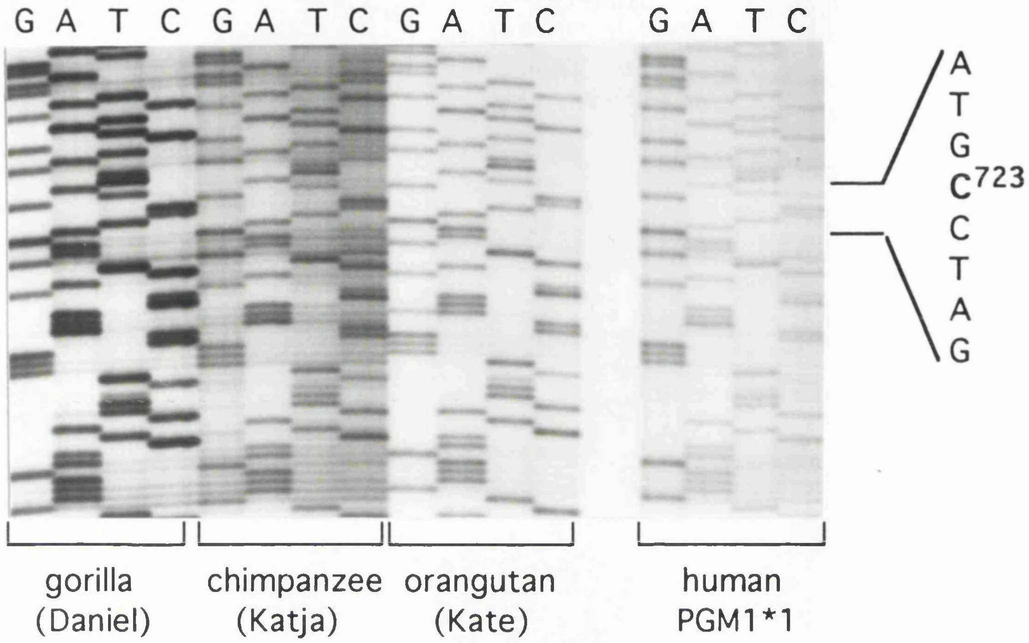


Figure 7.4 Autoradiograph of exon 4 nucleotide sequences from Sumatran and Bornean orangutans (693-723bp), demonstrating the nucleotide polymorphism at nt 707.

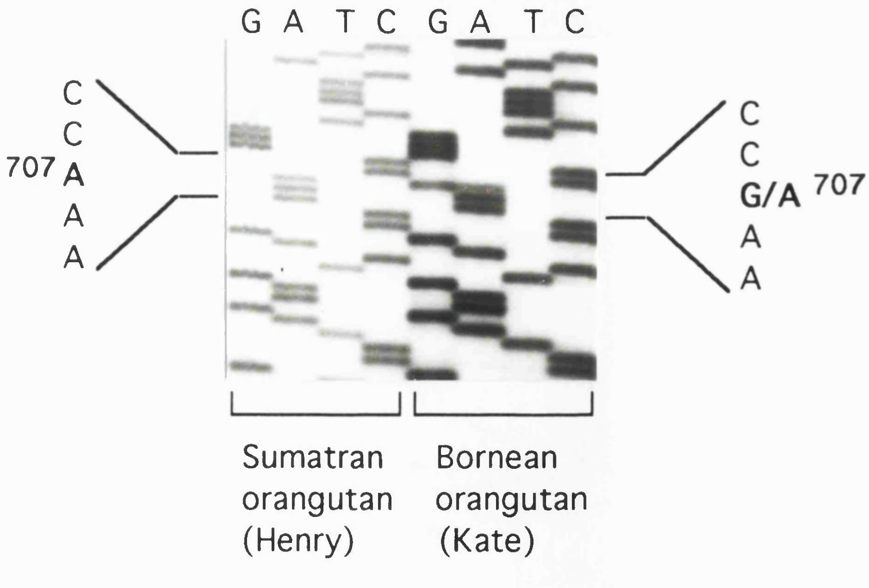


Figure 7.5 Autoradiograph of exon 8 nucleotide sequences from gorilla, chimpanzee, orangutan and human (1300-1333bp). Nucleotide 1320, which is the site of the +/- polymorphism, is shown in bold. All the primates carry a T residue, associated with the *PGM1**+ allele.

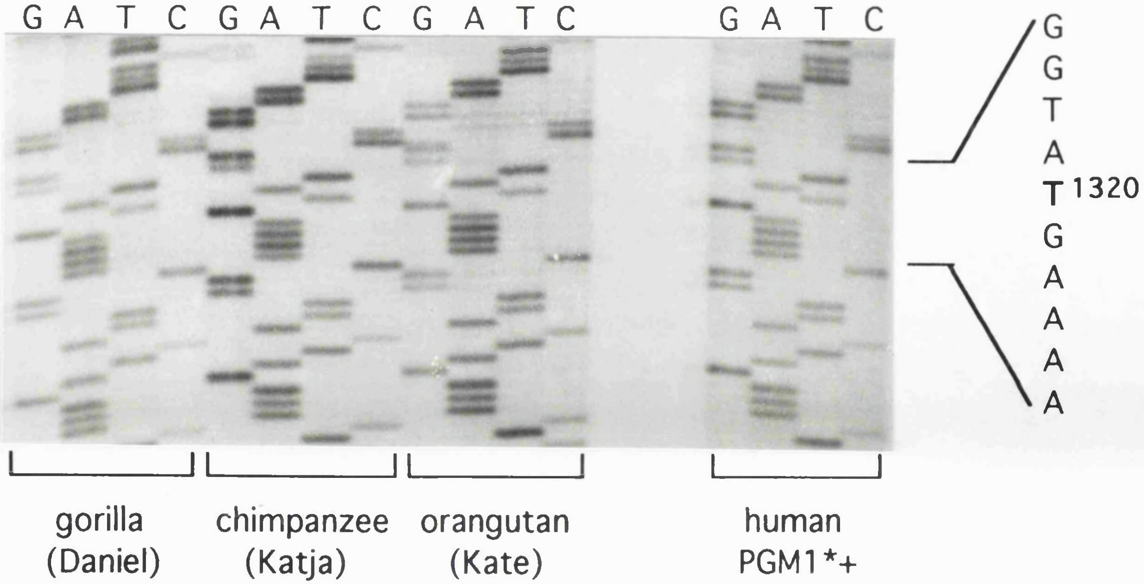


Figure 7.6 Multiple sequence alignments of primate, rabbit and rodent exon 4 nucleotide sequences.
Consensus sequence is from the human *PGM1**1 allele.
Coding sequence is in upper case and introns in lower case.
The bar (-) indicates an identical nucleotide, the asterisk (*) indicates deletions and the dot (.) indicates data not available.

```
          -49                                                619
consensus tctaaatgtg tttaatcctt ccatcttttg atgttgcttg ttctcacagT
  human1  ---------- ---------- ---------- ---------- ----------
  human2  ---------- ---------- ---------- ---------- ----------
   chimp  ---------- ---------a ---------a -----*---- ----------
 gorilla  ---------- ---------- ---------a -----*---- ----------
 orangB1  ---------- ----g----- ---------a ---c------ ----------
 orangB2  ---------- ----g----- ---------a ---c------ ----------
 orangS1  ---------- ----g----- ---------a ---c------ ----------
 orangS2  ---------- ----g----- ---------a ---c------ ----------
  rabbit  .......... .......... .......... .......... .......---
     rat  .......... .......... .......... .......... .......---
   mouse  .......... .......... .......... .......... ..........

          620                                                669
consensus GGAAATTGTG GATTCGGTAG AAGCTTATGC TACAATGCTG AGAAGCATCT
  human1  ---------- ---------- ---------- ---------- ----------
  human2  ---------- ---------- ---------- ---------- ----------
   chimp  ---------- ---------- ---------- ---------- ----------
 gorilla  ---------- ---------- ---------- ---------- ----------
 orangB1  ---------- ---------- -------C-- ---------- ----------
 orangB2  ---------- ---------- -------C-- ---------- ----------
 orangS1  ---------- ---------- -------C-- ---------- ----------
 orangS2  ---------- ---------- -------C-- ---------- ----------
  rabbit  ---------- -----A--G- ---------- ---G------ ----A-----
     rat  ---G--C--- --C--A--C- -G--C----- C--------- ----A-----
   mouse  .--G------ --C--A--G- -G--C----- C--------- ----A-----

          670                                                719
Consensus TTGATTTCAG TGCACTGAAA GAACTACTTT CTGGGCCAAA CCGACTGAAG
  human1  ---------- ---------- ---------- ---------- ----------
  human2  ---------- ---------- ---------- ---------- ----------
   chimp  ---------- ---------- ---------- ---------- ----------
 gorilla  ---------- ---------- ---------- ---------- ----------
 orangB1  ---------- ---------- ---------- ---------- ----------
 orangB2  ---------- ---------- ---------- -------G-- ----------
 orangS1  ---------- ---------- ---------- ---------- ----------
 orangS2  ---------- ---------- ---------- -------G-- ----------
  rabbit  ---------A ----T----- -----G--C- ---------- ------A---
     rat  ---------A C--------G --G-----C- ----C----- -A--------
   mouse  -C-------A C--------G --G-----C- ----T----- -A--------

          720                     744                           +27
consensus ATCCGTATTG ATGCTATGCA TGGAGgtata caatcatttc ttttcaattc cc
  human1  ---------- ---------- ---------- ---------- ---------- --
  human2  ---T------ ---------- ---------- ---------- ---------- --
   chimp  ---------- ---------- ---------- ---------- ---------- --
 gorilla  ---------- ---------- ---------- ---------- ---------- --
 orangB1  ---------- ---------- ---------- ---------- ---------- --
 orangB2  ---------- ---------- ---------- ---------- ---------- --
 orangS1  ---------- ---------- ---------- ---------- ---------- --
 orangS2  ---------- ---------- ---------- ---------- ---------- --
  rabbit  --------A- ----C----- -----..... .......... .......... ..
     rat  -----C--C- -C--C----- C----..... .......... .......... ..
   mouse  -----C--A- -C--C----- C----..... .......... .......... ..
```

Figure 7.7 Multiple sequence alignments of primate, rabbit and rodent exon 8 nucleotide sequences.
Consensus sequence is from the human *PGM1**+ allele.
Coding sequence is in upper case and introns in lower case.
The bar (-) indicates an identical nucleotide and the dot (.) indicates data not available.

```
          -23                      1207                         1233
consensus gcagcttgct gtcccccctc cagGTTCTGA CCACATCCGT GAGAAAGATG
 human+   ---------- ---------- ---------- ---------- ----------
 human-   ---------- ---------- ---------- ---------- ----------
  chimp   ---------- ---g------ ---------- ---------- ----------
gorilla   ---------- ---------- ---------- ---------- ----------
 orangB   ---------- c------t-- ---------- ---------- ----------
 orangS   ---------- c------t-- ---------- ---------- ----------
 rabbit   .......... .......... ...----C-- ------T--- ----------
    rat   .......... .......... ...----A-- ---------A ----------
  mouse   .......... .......... ...----G-- ---T-----A ----------

          1234                                              1283
consensus GACTGTGGGC TGTCCTTGCC TGGCTCTCCA TCCTAGCCAC CCGCAAGCAG
 human+   ---------- ---------- ---------- ---------- ----------
 human-   ---------- ---------- ---------- ---------- ----------
  chimp   ---------- ---------- ---------- ---------- ----------
gorilla   ---------- ---------- ---------- ---------- ----------
 orangB   ---------- ---------- ---------- ---------- ----------
 orangS   ---------- ---------- ---------- ---------- ----------
 rabbit   -G-------- ---G------ ---------- -T--G----- ------A---
    rat   ---------- ------G--- ---------- -T--G----- ------A---
  mouse   ---------- C-----G--- ---------- -T--G----- ------A---

          1284                                              1333
consensus AGTGTGGAGG ACATTCTCAA AGATCATTGG CAAAAGTATG GCCGGAATTT
 human+   ---------- ---------- ---------- ---------- ----------
 human-   ---------- ---------- ---------- ------C--- ----------
  chimp   ---------- ---------- ---------- ---------- ----------
gorilla   ---------- ---------- ---------- ---------- ----------
 orangB   ---------- ---------- ---------- ---------- ----------
 orangS   ---------- ---------- ---------- ---------- ----------
 rabbit   ---------- ----C----- ---C--C--- --C----TC- -------C--
    rat   --G------- ---------- ---C--C--- --G----T-- -T-----C--
  mouse   --C------- ----C----- ---C--C--- --G----T-- -T-----C--

          1334 1342                  +22
consensus CTTCACCAGg tgagccacag cccagctggg g
 human+   ---------- ---------- ---------- -
 human-   ---------- ---------- ---------- -
  chimp   ---------- ---------- ---------- -
gorilla   ---------- ---------- ---------- -
 orangB   ---------- ---------- ---------- -
 orangS   ---------- ---------- ---------- -
 rabbit   ---------. .......... .......... .
    rat   ---------. .......... .......... .
  mouse   ---T-----. .......... .......... .
```

residues 200 and 205 (Figure 7.8a). All three retain the C base at nt 723 which is associated with the 1 allele. However, in exon 8, the + allele is not conserved. Although they carry the T residue at position 1320, a nucleotide substitution at the following base changes the codon from **TAT** to **TTT**, such that a Phe^{419} is encoded in all three species (Figure 7.8b).

7.3.2 ANALYSIS OF EXON 11

To determine whether the +++ haplotype (allele 1) of the 3' UTR polymorphism in man could be identified in any of the great apes, the samples were amplified with exon 11 primers E11F and E11R (March et al, 1993b). Nucleotides 1712 to 1899 of exon 11 were subsequently sequenced from man (1), gorilla (2), chimpanzee (5) and orangutan (5).

The PCR products included the last 36 nucleotides of coding sequence: no nucleotide substitutions were found in the great apes. The 3' UTR in man contains three polymorphic sites at nt 1773, 1788 and 1844. No polymorphisms were demonstrated at these sites in the apes. All showed the +++ haplotype, apart from the two Sumatran orangutans which both carried a G base at nt 1788. Since this is one of the polymorphic nucleotides observed in man at this site (haplotype +-+), it is possible that the orangutans may also be polymorphic.

Five other base changes were found in the 3' UTR of the great apes, one in chimpanzee at nt 1757 and four in orangutan at nt 1847, nt 1866, nt 1881 and nt 1895. The change at nt 1847 is also seen in gorilla. The three unique substitutions in orangutan are illustrated in figure 7.9. The observed levels of nucleic acid sequence conservation supports the current view of primate evolution, based upon both molecular and morphometric data, that chimpanzees and gorillas are more closely related to man than orangutans. A large number of base changes were seen in rabbits and rats, which reflects the great level of divergence between lagomorphs, rodents and primates (Figure 7.10).

7.3.3 ANALYSIS OF EXONS 1A AND 5

The lemur DNA sample was not amplified by the exon 4, 8 and 11 primers. This was thought to reflect the level of nucleic acid divergence which has occurred between the intron sequences of these primates. Therefore, in order to investigate the level of nucleotide and amino acid conservation of PGM1 in

Figure 7.8 Multiple sequence alignments of primate, rabbit and rodent amino acid sequences from a) exon 4 and b) exon 8. Consensus sequence is from the human *PGM1**1+ allele. The bar (-) indicates an identical amino acid.

a)

```
          186                                      226
consensus EIVDSVEAYA TMLRSIFDFS ALKELLSGPN RLKIRIDAMH G
 human1   ---------- ---------- ---------- ---------- -
 human2   ---------- ---------- ---------- ----C----- -
  chimp   ---------- ---------- ---------- ---------- -
gorilla   ---------- ---------- ---------- ---------- -
 orangB   ---------- ---------- ---------- ---------- -
 orangS   ---------- ---------- ---------- ---------- -
 rabbit   ---------- ----N----N ---------- ---------- -
    rat   ---------- ----N----N ---------- ---------- -
  mouse   ---------- ----N----N ---------- ---------- -
```

Human polymorphism: Arg^{220} to Cys^{220}

b)

```
          382                                         425
consensus SDHIREKDGL WAVLAWLSIL ATRKQSVEDI LKDHWQKYGR NFFT
 human+   ---------- ---------- ---------- ---------- ----
 human-   ---------- ---------- ---------- -------H-- ----
  chimp   ---------- ---------- ---------- ---------- ----
gorilla   ---------- ---------- ---------- ---------- ----
 orangB   ---------- ---------- ---------- ---------- ----
 orangS   ---------- ---------- ---------- ---------- ----
 rabbit   ---------- ---------- ---------- -----H-F-- ----
    rat   ---------- ---------- -----R---- -------F-- ----
  mouse   ---------- ---------- ---------- -------F-- ----
```

Human polymorphism: Tyr^{419} to His^{419}

Figure 7.9 Autoradiograph of exon 11 nucleotide sequences from gorilla, chimpanzee and orangutan (1862-1898bp). Nucleotide substitutions in the 3'UTR sequence are shown in bold.

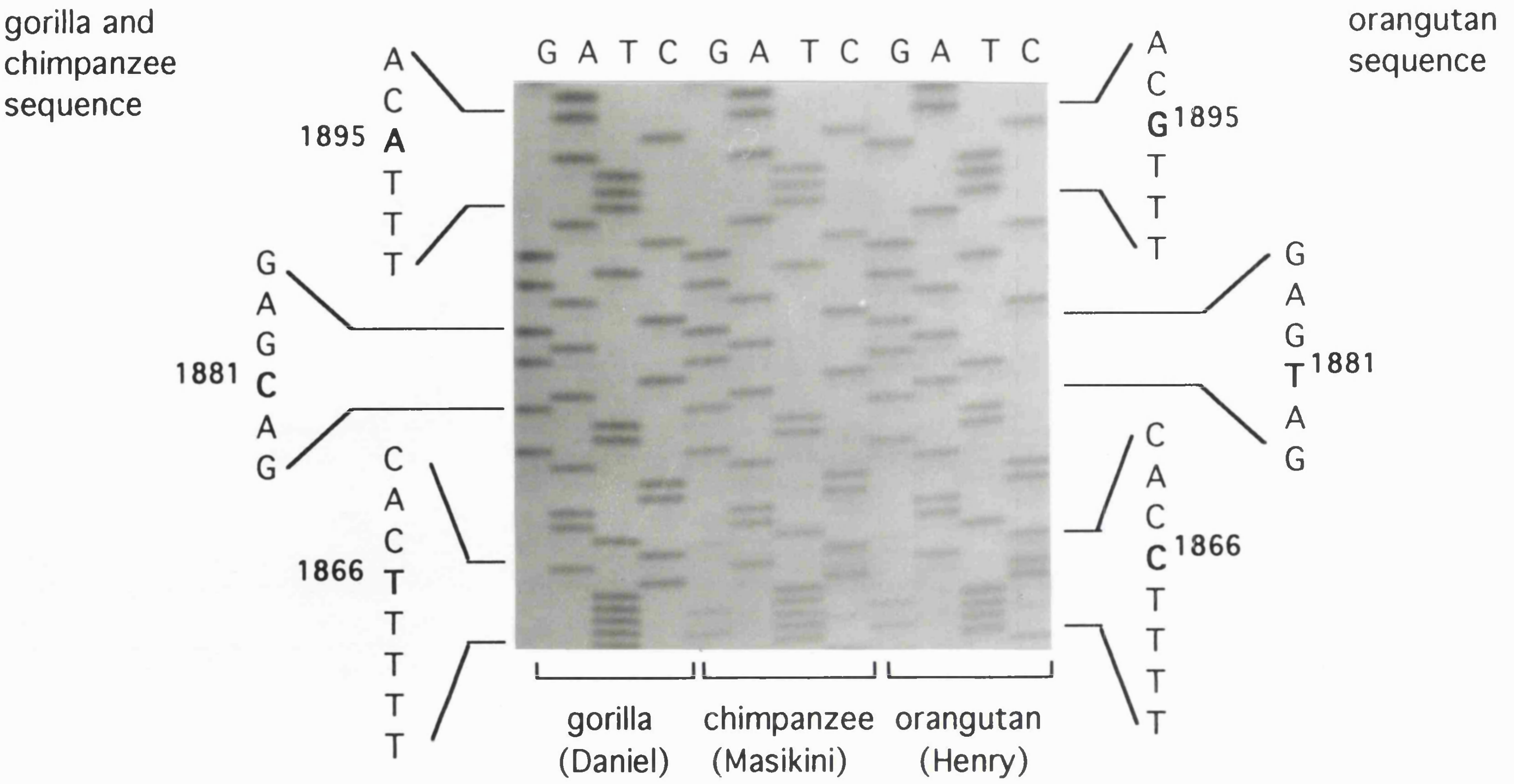


Figure 7.10 Multiple sequence alignments of primate, rabbit and rat exon 11 nucleotide sequences.
Coding sequence is in upper case, 3' UTR is in lower case.
Consensus sequence is from the human *PGM1**3'UTR 1 allele.
The bar (-) indicates an identical nucleotide and the asterisk (*) indicates deletions.

```
          1712                                                1761
consensus GGAGAGGACG GGACGCACTG CACCCACTGT CATCACCtaa gaagacaggc
 human1   ---------- ---------- ---------- ---------- ----------
 human2   ---------- ---------- ---------- ---------- ----------
 human3   ---------- ---------- ---------- ---------- ----------
 human4   ---------- ---------- ---------- ---------- ----------
  chimp   ---------- ---------- ---------- ---------- -----g----
gorilla   ---------- ---------- ---------- ---------- ----------
 orangB   ---------- ---------- ---------- ---------- ----------
 orangS   ---------- ---------- ---------- ---------- ----------
 rabbit   A--A-----A ---------- -------C-- ---------- ---c----a-
    rat   ---------A --C------- -C--A----- ---------g ag---*t---

          1762                                                1811
consensus ctgatgtggt acgtccctcc acccccggac ccatccaagt catctgattg
 human1   ---------- ---------- ---------- ---------- ----------
 human2   ---------- -t-------- ---------- ---------- ----------
 human3   ---------- -t-------- ------a--- ---------- ----------
 human4   ---------- ---------- ------a--- ---------- ----------
  chimp   ---------- ---------- ---------- ---------- ----------
gorilla   ---------- ---------- ---------- ---------- ----------
 orangB   ---------- ---------- ------a--- ---------- ----------
 orangS   ---------- ---------- ---------- ---------- ----------
 rabbit   -a----at-- ---------- g--******* ****------ ----------
    rat   tg-------c -****-a--- t----a---a -tg----cac tgc-------

          1812                                                1860
consensus aagagcat*g acagaaacaa aatgtattca ccaagcattt taggatttga
 human1   ---------- ---------- ---------- ---------- ----------
 human2   ---------- ---------- ---------- ---c------ ----------
 human3   ---------- ---------- ---------- ---c------ ----------
 human4   ---------- ---------- ---------- ---c------ ----------
  chimp   ---------- ---------- ---------- ---------- ----------
gorilla   ---------- ---------- ---------- ------t--- ----------
 orangB   ---------- ---------- ---------- ------t--- ----------
 orangS   ---------- ---------- ---------- ------t--- ----------
 rabbit   --------g- -g-------- -g--g--agg --cc-ac--- ct--ga--tg
    rat   -------ca- -----***-c -g-------g a-ct-gcc-- ---ac-catc

          1861                                   1898
consensus ctttttcact aaccagttga cgagcagtgc atttacaa
 human1   ---------- ---------- ---------- --------
 human2   ---------- ---------- ---------- --------
 human3   ---------- ---------- ---------- --------
 human4   ---------- ---------- ---------- --------
  chimp   ---------- ---------- ---------- --------
gorilla   ---------- ---------- ---------- --------
 orangB   -----c---- ---------- t--------- ----g---
 orangS   -----c---- ---------- t--------- ----g---
 rabbit   a-c---acac t-a-t----c -a-a---ca- g--gcgt-
    rat   t-a---t-
```

lemur, primers designed to the coding sequences of exons 1A and 5 were used to amplify both the lemur and the great ape samples.

The samples were amplified by exon 1A primers E1F and E1R. Nucleotides 136 to 276 were sequenced from lemur, gorilla, chimpanzee, orangutan (both a Sumatran and Bornean) and man. Primers E5F and E5R were used to amplify exon 5. Sequence was obtained from nt 802 to 925 of the six primates.

The PGM1 3/7 alleles, found at polymorphic frequencies in some Asian-Pacific populations, are characterized by a mutation in exon 1A. An A to T transition at nt 265 leads to a substitution of Lys^{67}, encoded by AAG, for Met^{67}, encoded by ATG, to give rise to the 3/7 alleles. In all the primate samples tested, an A base was found at nt 265, conserving the codon AAG. Thus, Lys^{67}, which is associated with the 2/1 alleles of human PGM1, is found in primates. This codon is also conserved in rabbits. Rats, in contrast, show two base changes, such that the codon ACC encodes the amino acid Thr^{67}.

The exon 1A sequences of man, chimpanzee and orangutan were identical. An A to G transition was seen in gorilla at nt 269, although it was a synonymous mutation. The lemur showed greater variation in the coding sequence, with four synonymous mutations at nt 176, nt 200, nt 230 and nt 252 (Figures 7.11 and 7.12). A missense mutation was also identified, at nt 214, an A to T transversion giving rise to an amino acid substitution of Gln^{50} to Leu^{50}. This missense mutation is also seen in rat.

In exon 5, man, chimpanzee and gorilla nucleotide sequences were identical. A synonymous mutation at nt 815 was identified in orangutan. The lemur exhibited four base changes (Figure 7.13), of which two at nt 866 and nt 872 were synonymous. The first missense mutation was at nt 900, a G to C transversion leading to the substitution of Glu^{279} to Gln^{279}, and the second at nt 911, a T to G transversion leading to the substitution of Phe^{282} to Leu^{282}. Unexpectedly, in rabbit, rat and mouse PGM1, despite a number of nucleotide substitutions in exon 5, the amino acid sequences are identical to man (Figure 7.14).

Figure 7.11 Autoradiograph of exon 1A nucleotide sequences from lemur, gorilla, chimpanzee, orangutan and human (171-208bp). Nucleotide substitutions between the lemur and other primates are shown in bold.

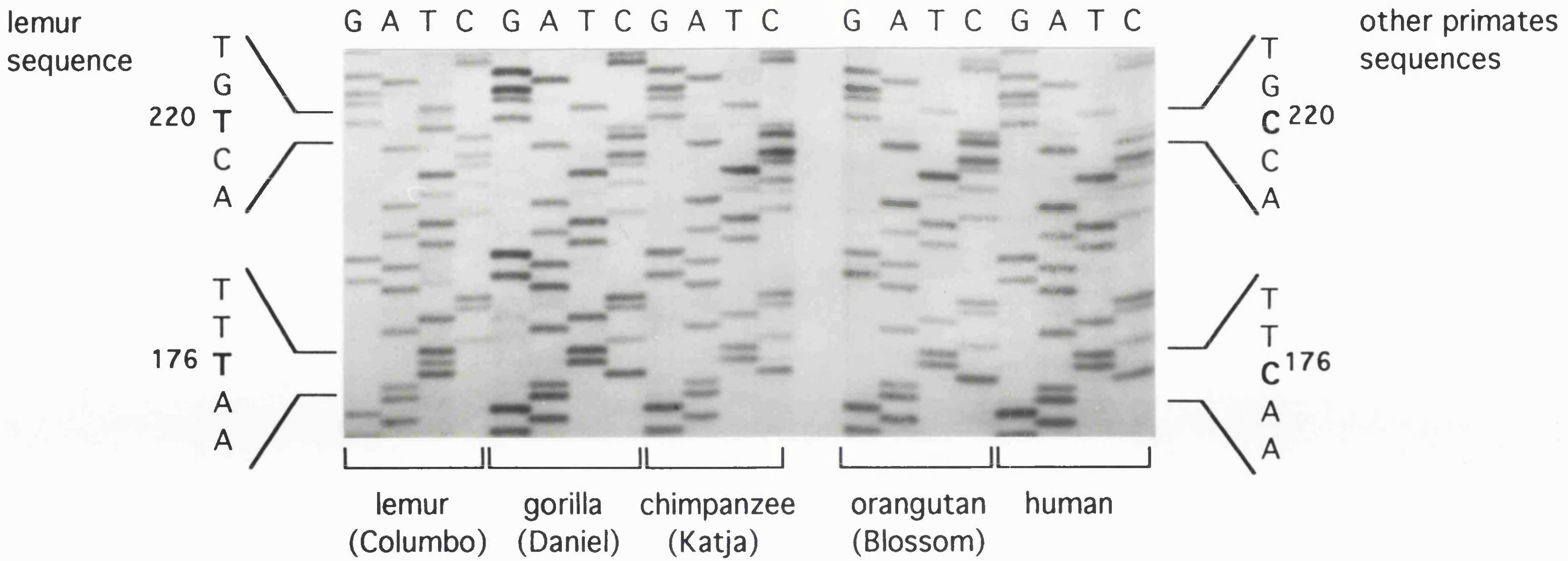


Figure 7.12 Multiple sequence alignments of primate, rabbit and rat exon 1A sequences at the a)DNA level and b) amino acid level. Consensus sequence is from the human PGM1*1+ allele. The bar (-) indicates an identical nucleotide or amino acid.

a)

```
          136                                                185
consensus GGGTGAAGGT GTTCCAGAGC AGCGCCAACT ACGCGGAGAA CTTCATCCAG
 humanA   ---------- ---------- ---------- ---------- ----------
 humanB   ---------- ---------- ---------- ---------- ----------
  chimp   ---------- ---------- ---------- ---------- ----------
gorilla   ---------- ---------- ---------- ---------- ----------
 orangS   ---------- ---------- ---------- ---------- ----------
 orangB   ---------- ---------- ---------- ---------- ----------
  lemur   ---------- ---------- ---------- ---------- T---------
 rabbit   ---------- ---------- ---A------ -T-------- ----------
    rat   -A-------- -------G-- -A---T---- -T-------- T--------A

          186                                                235
consensus AGTATCATCT CCACCGTGGA GCCGGCGCAG CGGCAGGAGG CCACGCTGGT
 humanA   ---------- ---------- ---------- ---------- ----------
 humanB   ---------- ---------- ---------- ---------- ----------
  chimp   ---------- ---------- ---------- ---------- ----------
gorilla   ---------- ---------- ---------- ---------- ----------
 orangS   ---------- ---------- ---------- ---------- ----------
 orangB   ---------- ---------- ---------- ---------- ----------
  lemur   ---------- ----T----- --------T- ---------- ----C-----
 rabbit   ---------- ---------- ---------- ---------- ----C-----
    rat   --C---G--- ---------- -------TT- A--------- -T--C-----

          236                                     276
consensus GGTGGGCGGG GACGGCCGGT TCTACATGAA GGAGGCCATC C
 humanA   ---------- ---------- ---------- ---------- -
 humanB   ---------- ---------- ---------T ---------- -
  chimp   ---------- ---------- ---------- ---------- -
gorilla   ---------- ---------- ---------- ---A------ -
 orangS   ---------- ---------- ---------- ---------- -
 orangB   ---------- ---------- ---------- ---------- -
  lemur   ---------- ------A--- ---------- ---------- -
 rabbit   ---------- ------A--- ---------- ---------- -
    rat   T--------- -----T--C- ---------C C--------- -
```

b)

```
          25                                            70
consensus VKVFQSSANY AENFIQSIIS TVEPAQRQEA TLVVGGDGRF YMKEAI
 humanA   ---------- ---------- ---------- ---------- ------
 humanB   ---------- ---------- ---------- ---------- --M---
  chimp   ---------- ---------- ---------- ---------- ------
gorilla   ---------- ---------- ---------- ---------- ------
 orangB   ---------- ---------- ---------- ---------- ------
 orangS   ---------- ---------- ---------- ---------- ------
  lemur   ---------- ---------- -----L---- ---------- ------
 rabbit   -------T-- ---------- ---------- ---------- ------
    rat   -----GN--- --------V- -----L---- ---------- --T---
```

Human polymorphism: Lys^{67} to Met^{67}

(The authors have numbered the amino acids to include Met^{1})

Figure 7.13 Autoradiograph of exon 5 nucleotide sequences from lemur, gorilla, chimpanzee, orangutan and human (855-912bp). Nucleotide substitutions between the lemur and other primates are shown in bold.

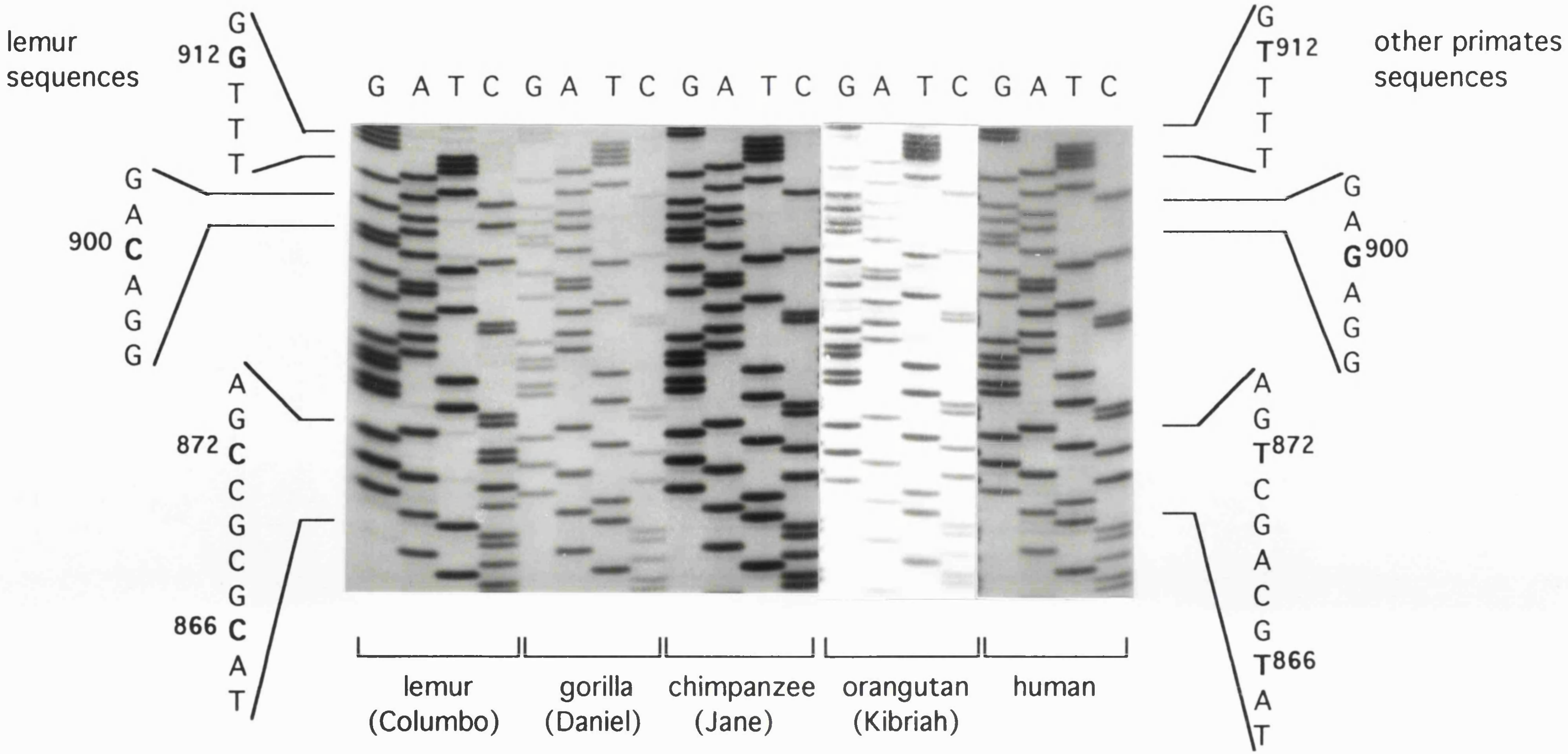


Figure 7.14 Multiple sequence alignments of primate, rabbit and rat exon 5 sequences at the a)DNA level and b) amino acid level. Consensus sequence is from the human *PGM1**1+ allele.
The bar (-) indicates an identical nucleotide or amino acid.

a)

```
          802                                                  851
consensus CGGCAGTTAA CTGCGTTCCT CTGGAGGACT TTGGAGGCCA CCACCCTGAC
    human ---------- ---------- ---------- ---------- ----------
    chimp ---------- ---------- ---------- ---------- ----------
  gorilla ---------- ---------- ---------- ---------- ----------
   orangB ---------- ---T------ ---------- ---------- ----------
   orangS ---------- ---T------ ---------- ---------- ----------
    lemur ---------- ---------- ---------- ---------- ----------
   rabbit -C--T--C-- ---T------ ---------- -C--G----- ----------
      rat -A--T--G-- ---T-----C --------T- ---------- ----------
    mouse -A--T--G-- ---T--C--C --------T- ---------- ---T--C---

          852                                                  901
consensus CCCAACCTCA CCTATGCAGC TGACCTGGTG GAGACCATGA AGTCAGGAGA
    human ---------- ---------- ---------- ---------- ----------
    chimp ---------- ---------- ---------- ---------- ----------
  gorilla ---------- ---------- ---------- ---------- ----------
   orangB ---------- ---------- ---------- ---------- ----------
   orangS ---------- ---------- ---------- ---------- ----------
    lemur ---------- ----C----- C--------- ---------- --------C-
   rabbit ---------- ----C----- ---------- -----G---- ----C-----
      rat ---------- ----C--T-- ---------- --A------- ----------
    mouse -----T---- -------T-- ------A--- ---------- ----------

          902                    925
consensus GCATGATTTT GGGGCTGCCT TTGA
    human ---------- ---------- ----
    chimp ---------- ---------- ----
  gorilla ---------- ---------- ----
   orangB ---------- ---------- ----
   orangS ---------- ---------- ----
    lemur ---------G ---------- ----
   rabbit ------C--C ---------- ----
      rat ---------C ---------- ----
    mouse ---------C ---------- ----
```

b)

```
          247                                       286
consensus AVNCVPLEDF GGHHPDPNLT YAADLVETMK SGEHDFGAAF
    human ---------- ---------- ---------- ----------
    chimp ---------- ---------- ---------- ----------
  gorilla ---------- ---------- ---------- ----------
   orangB ---------- ---------- ---------- ----------
   orangS ---------- ---------- ---------- ----------
    lemur ---------- ---------- ---------- --Q--L----
   rabbit ---------- ---------- ---------- ----------
      rat ---------- ---------- ---------- ----------
    mouse ---------- ---------- ---------- ----------
```

7.4 SUMMARY

i) The gorilla, chimpanzee and orangutan samples all carry the C at nt 723 and T at nt 1320 characteristic of the PGM1*1+ protein phenotype seen in man. Therefore, the primate isozyme which appears to be PGM1*1+ on IEF also exhibits the *PGM1**1+ characteristics at the DNA level.

ii) The gorilla and chimpanzee samples all carry the nucleotides associated with the +++ haplotype (allele 1) of the 3' UTR polymorphism observed in man. In orangutans, at nt 1788, the two polymorphic nucleotides, characteristic of the +-+ and +++ haplotypes seen in man, were observed, but they were not demonstrated to be polymorphic; the G nucleotide was confined to the Sumatran and the A nucleotide to the Bornean orangutans.

iii) In the three exons 4, 8 and 11 where multiple, presumedly unrelated samples of each species were investigated, no polymorphic nucleotide substitutions leading to changes in the amino acid sequence were demonstrated in the coding sequence of *PGM1*. The only nucleotide polymorphism identified was in exon 4 at nt 707 in the orangutans. The three Bornean orangutans were all found to be heterozygous, whilst one of the Sumatran orangutans, Henry, was homozygous for the A nucleotide, and the other, Annie, was homozygous for the G.

iv) In exons 1A and 5, a greater level of nucleotide and amino acid divergence was evident in the lemur. The amino acids were completely conserved in the gorilla, chimp and orangutans, whilst the lemur contained one amino acid substitution in exon 1A and two in exon 5. Interestingly, in exon 5, although the rabbit and rodent show a greater level of nucleotide diversity than the lemur, the amino acid sequence is identical to the great apes and man.

7.5 CONCLUSIONS

The molecular basis of the *PGM1**1+ allele in man is conserved in the great apes which suggests that the *PGM1**1 and the *PGM1**+ alleles are conserved among primates of the hominoidae superfamily. This conclusion also provides support that the *PGM1**1+ is the ancestral allele in man. Conservation of the *PGM1**1 allele is retained in rabbits and rodents, although the *PGM1**+ allele is not.

Nucleotide sequence data from exon 11 of *PGM1* in the great apes supports the proposal of the +++ haplotype of the 3'UTR polymorphism being ancestral. Although no polymorphisms were detected at the three sites in the primates, the two polymorphic bases at nt 1788 in man, G and A, are observed in the two populations of orangutans. Whether these two populations are indeed polymorphic at this locus, or whether these mutations are fixed cannot be determined with the limited number of samples available. However, two suggestions emerge from this data. First, the nucleotide substitutions in man and orangutans may have occurred independently. Alternatively, if the orangutans are polymorphic at this locus, it may be inferred that of the three polymorphic sites in man, this was the initial polymorphism, and it occurred prior to the divergence of orangutans and man.

The IEF data provides evidence of intraspecific variation with an additional cathodal band present in the chimpanzee Halfpenny and the gorilla Daniel. DNA sequence analysis of exons 4, 8 and 11 of *PGM1* in Halfpenny identified no heterozygous nucleotides, indicating the mutation which underlies this polymorphism is not located in these exons. In Daniel, exons 1A, 4, 5, 8 and 11 of *PGM1* were sequenced. Again, no heterozygous nucleotides were identified. Therefore, this data suggests that the molecular basis of intraspecific variation in the primates occurs at a site distinct from those in man.

Exons 1A and 5 from *PGM1* of lemur show a greater number of nucleotide changes compared to human *PGM1*, than the great apes, reflecting the evolutionary distance between the species. The lemur belongs to the suborder prosimii, whereas man, gorilla, chimpanzee and orangutan belong to the suborder anthropodiea. The divergence of these two suborders is estimated to have occurred between 65 and 56 million years ago. Since a number of nucleotide changes were demonstrated in the coding sequence of lemur *PGM1*, the intron sequences would be expected to show even greater nucleotide divergence from human *PGM1*. Therefore, the failure of the *PGM1* intron sited primers to amplify exons 4, 8 and 11 is most probably due to mismatches between the primers and the template DNA.

CHAPTER EIGHT:

EVOLUTION OF THE PHOSPHOHEXOMUTASES

Phosphoglucomutases (PGM) and phosphomannomutases (PMM) have been cloned from a wide variety of organisms, including prokaryotes and eukaryotes, protozoans and metazoans. Comparison of orthologous sequences (divergence following speciation) from these species would allow a phylogenetic tree to be constructed, from which the evolution of the species can be inferred. However, in this chapter, knowledge of the evolution of the species is used to investigate the molecular evolution of the phosphohexomutases. Therefore both orthologous and paralogous sequences (divergence following duplication) have been included in the analysis.

Immunological studies using anti-rabbit PGM polyclonal antibodies (Chapter Three) and the low stringency and degenerate primer PCR approaches (Chapter Four) suggest that the genes encoding the PGM2 and PGM3 isozymes are not as closely related to PGM1 as was first thought from simple comparison of isozyme patterns. Therefore, the primary aim of this investigation was to place PGM, PMM and related sequences within an evolutionary framework and to see if there are divergent clusters of sequences that may suggest alternative pathways of evolution for PGM2 and PGM3. Identification of these pathways may provide additional information, such as conserved protein motifs, which may lead to the identification and characterization of these loci.

In addition, the phylogenetic analysis should identify duplications of the common ancestral gene and allow the evolutionary relationship between the many prokaryotic sequences to be investigated. Finally, the possibility of *Agrobacterium tumefaciens PGM* having arisen by divergence following a trans-kingdom horizontal gene transfer (an example of a xenologous sequence) will be examined through its phylogenetic relationship to eukaryotic and prokaryotic sequences.

8.1 CONSTRUCTION OF PHYLOGENETIC TREES

Amino acid sequences were used to construct the phylogeny, since they allow more distantly related sequences to be identified. Proteins evolve more slowly than nucleotide sequences, in part due to constraints which conserve the structure and function of the protein. The nucleotide sequences from the diverse range of species in this study would not detect distant relationships due to the excessive number of random mutation events which have occurred throughout evolution.

8.1.1 COMPILATION OF SEQUENCES

PGM, PMM and related sequences were obtained from the Genbank and EMBL nucleotide databases, using two searching strategies. The first approach used 'stringsearch' (GCG) to identify any sequences in which the keyword phosphoglucomutase or phosphomannomutase appeared in the definition. Ten PGM and thirteen PMM sequences were identified (Figure 8.1). The second approach used the peptide sequences of the active site and magnesium binding loop motifs, TASHNP and FDGDGDR respectively, as probes to identify PGM-related sequences. The nucleotide databases were searched using the tfasta option of the 'fasta' database searching programme, which enables comparison of the protein query sequence against nucleotide sequences. Five sequences which showed high conservation of the characteristic peptide motifs, but were not characterized as PGM or PMM, were identified, along with the majority (17) of the PGM and PMM sequences identified previously (Figure 8.1).

Amino acid sequences were derived from the cDNA using 'translate' (GCG). Any unusual codon usage was edited using the sequence editor 'seqed' (GCG); for example, in *Mycoplasma pirum*, UGA codons code for tryptophan and in *Paramecium tetraurelia*, UAA codons code for glutamine. All 28 peptide sequences showed an overall identity to human PGM1 of more than 20% using bestfit (GCG), and an identity of greater than 20% is thought to indicate a common ancestry (Creighton, 1993). A multiple sequence file of all the sequences was compiled using the text editor 'emacs'.

8.1.2 MULTIPLE SEQUENCE ALIGNMENTS

The 28 peptide sequences were aligned using the multiple sequence alignment programme 'pileup' (GCG) (Appendix A). The gap weight used was 3.0 and gap length weight was 0.1. These values produced the most optimal results, aligning the active site and magnesium binding loop motifs whilst minimizing the number of gaps.

Figure 8.1 List of protein sequences for the phylogenetic analysis

Species	PGM, PMM or related protein (Gene)	Genbank Acc. No	Amino Acid Identity to Human PGM1	Reference
Homo sapiens	PGM (*PGM1*)	M83033	-	Whitehouse et al, 1992
Homo sapiens	related (*PGMRP*)	L40933	67.6 %	Moiseeva et al, 1996
Spinacia oleracea	PGM (*pgm*)	X75898	20.1 %	Penger et al, 1994
Parafusin tetraurelia	related (*PFUS*)	L12471	54.6 %	Subramanian et al, 1994
Saccharomyces cerevisiae	PGM (*PGM1*)	X72016	51.5 %	Boles et al, 1994
Saccharomyces cerevisiae	PGM (*PGM2*)	X74823	52.4 %	Boles et al, 1994
Saccharomyces cerevisiae	related (*AGM1*)	X75816	20.8 %	Boles et al, 1994
Azospirillum brasilense	PMM (*exoC*)	U20583	23.4 %	Peterson & Vanderleyden, unpub.
Agrobacterium tumefaciens	PGM (*PGM*)	L24117	55.8 %	Uttaro et al, 1994 Uttaro et al, 1995
Neiserria meningitidis	PGM (*pgm*)	U02490	25.6 %	Zhou et al, 1994
Neisseria gonorrhoeae	PGM (*pgm*)	U02489	26.7 %	Zhou et al, 1994
Salmonella enterica B/LT2	PMM (*rfbK*)	X56793	23.3 %	Jiang et al, 1991
Salmonella enterica C1/M40	PMM (*rfbK*)	M84642	25.5 %	Lee et al, 1992b
Salmonella enterica B/LT2	PMM (*cpsG*)	X59886	25.1 %	Stevenson et al, 1991
Escherichia coli K12	PGM (*pgm*)	M77127	24.7 %	Tal et al, unpub.
Escherichia coli	PGM (*pgm*)	U08369	25.7 %	Lu & Kleckner, 1994

Figure 8.1 cont.

Escherichia coli 07/K1	PMM (*rfbK*)	L04596	24.5 %	Marolda & Valvano, 1993
Escherichia coli 09/E69	PMM (*rfbK1*)	L27646	22.2 %	Jayaratne et al, 1994
Escherichia coli 09/E69	PMM (*rfbK2*)	L27632	22.2 %	Jayaratne et al, 1994
Escherichia coli 09/F719	PMM (*rfbK*)	D13231	21.6 %	Sugiyama et al, 1994
Escherichia coli K12	PMM (*cpsG*)	L11721	24.7 %	Aoyama et al, 1994
Escherichia coli K12	related (*yhbf*)	L12968	25.8 %	Dallas et al, 1993
Coxiella burnetti	PMM (*pmm*)	X79075	27.0 %	Thiele et al, unpub.
Xanthamonas camprestris	PMM (*xanA*)	M83231	23.5 %	Koplin et al, 1992
Pseudomonas aeroginosa	PMM (*algC*)	M60873	23.3 %	Zielenski et al, 1991
Acetobacter xylinum	PGM (*celB*)	L24077	26.3 %	Brautaset et al, 1994
Heliobacter pylori	related (*ureC*)	X57132	24.8 %	Labigne et al, 1991
Mycoplasma pirum	PMM (*pmm*)	L13289	20.5 %	Tham et al, 1993

8.1.3 PHYLOGENETIC ANALYSIS

Phylogenetic trees were produced using both maximum parsimony and neighbour-joining distance methods. Parsimony trees were constructed using the phylogenetic package PAUP (Phylogenetic Analysis Using Parsimony; Swofford, 1990); the resulting phylogeny is the tree which shows the minimum total tree length, that is, the minimum number of evolutionary steps required to obtain the data. Two trees were constructed. In the first, amino acid changes were unweighted, such that changes from one amino acid to another were equally probable, whilst in the second, the amino acid changes were weighted according to the minimum number of nucleotide substitutions encoding that change (Fitch & Margoliash, 1967).

In contrast to parsimony methods, the neighbour-joining distance method constructs the tree by first linking the least distant pair of sequences, and then adds the next most closely related sequence (Saitou & Nei, 1987). Thus the tree is constructed according to the calculated genetic distances between the sequences. Pairwise genetic distances were calculated for all of the sequences (Appendix B), using the point accepted mutation (PAM) matrix (Dayhoff, 1978). This weights amino acid substitutions according to their chemical properties and frequency of occurrence in proteins.

8.1.4 THE BOOTSTRAP RESAMPLING METHOD

The support for the major nodes within both the parsimony and distance trees were evaluated by the bootstrap. This procedure samples amino acid positions randomly from the data matrix, in this case the multiple sequence alignment file, to build a new data set the same size as the original. The new data set is then used to construct a new tree. A consenus tree of the specified number of bootstrap replicates provides a measure of support for the nodes within the tree. Generally, those nodes found in 95% of bootstrap replicates are thought to be strongly supported. For the unweighted parsimony tree and distance trees, 100 bootstrap replicates of the whole data set were examined. However, the size of the data set made computation of the bootstrap values for the weighted parsimony trees impossible. Therefore, to estimate the bootstrap support for these trees, one of a pair of sequences which were shown to be 90% or more identical were eliminated.

8.2 THE CONSENSUS PHYLOGENETIC TREE

The phylogenies obtained from both the parsimony and distance methods gave the same basic topology (Figure 8.2). The major feature of the tree was the presence of three well supported groups, the sequences of which were characterized by the conserved putative glucose binding loop or equivalent motif. The first group contains the majority of eukaryotic PGM and PGM-related sequences and the prokaryotic *Agrobacterium tumefaciens* PGM, all of which show the GEESFG putative glucose binding loop motif. The other two major groups contain prokaryotic sequences; one group consists of enterobacteria, which are characterized by the GEMSAH motif and the other group contains the non-enterobacteria proteobacteria which show the GEMSGH motif. In addition, there are several other sequences which did not fall into these three groups. In most, but not all, a distinct motif corresponding to the sugar binding loop was identified.

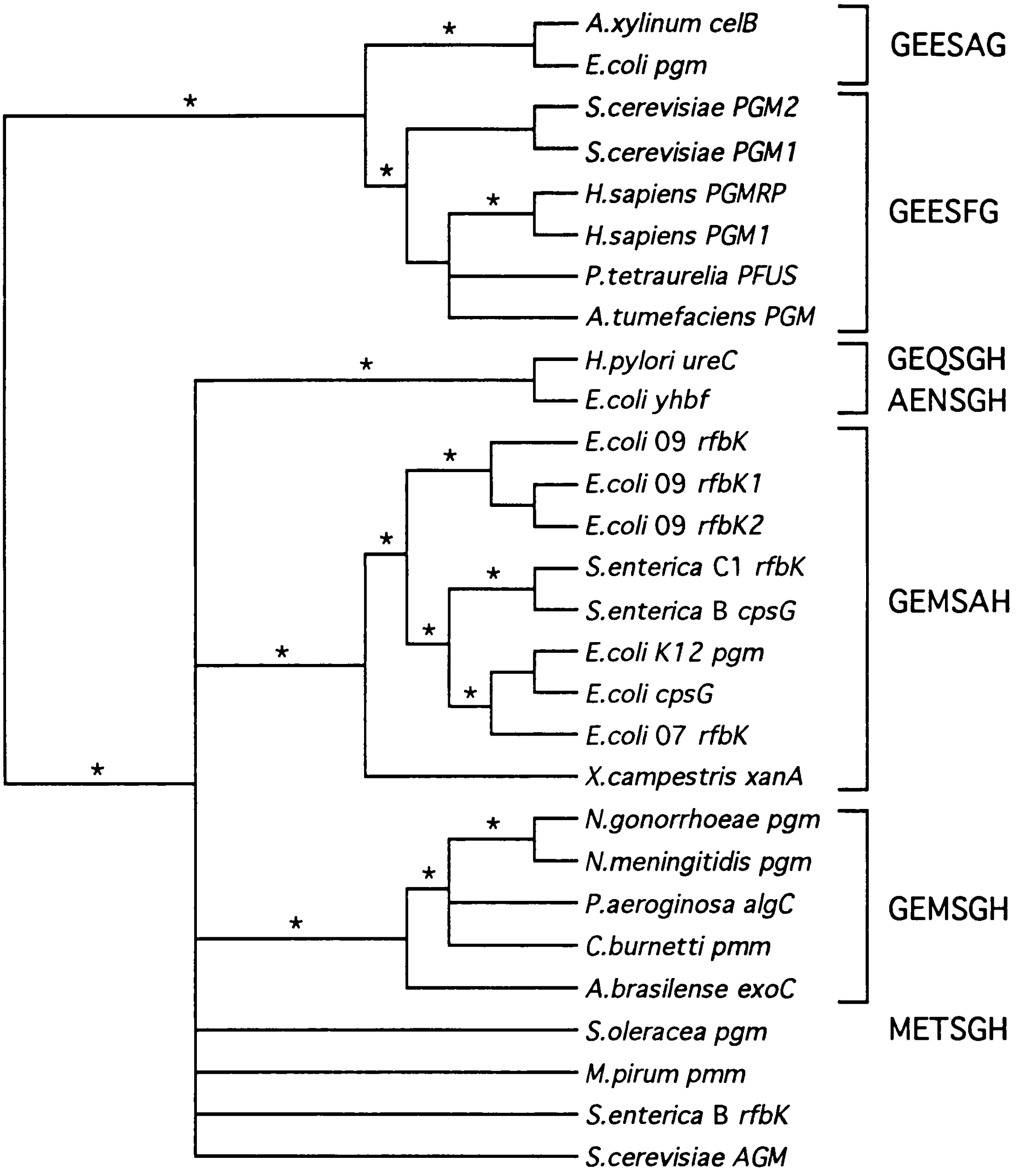
8.2.1 PROKARYOTIC PGM AND RELATED SEQUENCES

The prokaryotic sequences show extensive diversity within this putative gene family. Although there are the two major clusters of sequences, the phylogenetic analysis identifies six distinct evolutionary pathways of prokaryotic PGM and related sequences. The first pathway contains the true bacterial PGM sequences, identified in *E.coli* and *A.xylinum*. The second and third are represented by the two major clusters of enterobacteria and non-enterobacterial proteobacteria sequences. The fourth pathway is represented by a single sequence from *S.enterica*, which is quite distinct from the other enterobacterial sequences. The fifth pathway contains the *ureC* and *yhbf* gene products which may represent a change in function of the ancestral gene; the *ureC* gene product is required for urease activity. The sixth pathway is represented by the PMM of the gram positive bacteria *Mycoplasma pirum*. In addition, the *A.tumefaciens* PGM may represent a further pathway although this gene is later suggested to be the result of a trans-kingdom horizontal gene transfer event (section 8.2.2.1)

8.2.1.1 *Esherichia coli* and *Acetobacter xylinum* PGM

The phylogenetic analysis indicates that *E.coli* and *A.xylinum* PGM are more closely related to eukaryotic PGM than to the other prokaryotic sequences (with the exception of *A.tumefaciens* PGM). These two proteins possess a conserved sugar binding loop motif, GEESAG, which differs from the eukaryotic

Figure 8.2 Phylogeny of the proteins encoded by the PGM, PMM and PGM-related genes, obtained by both maximum parsimony and neighbour-joining distance methods. The tree is unrooted. The three major groups of sequences are characterized by the conserved putative sugar binding loop. Nodes indicated with an asterisk (*) indicate bootstrap support of above 95%.



GEESFG motif by a single amino acid substituition. They are also of a similar size to the eukaryotic proteins, containing approximately 550 amino acids, whereas the other prokaryotic sequences are smaller, generally containing about 460 amino acids. Therefore, the GEESAG containing sequences are probably the true bacterial homologues of *PGM1*. In addition to their role in glycolysis, the PGM proteins are involved in the biosynthetic pathways of the outer membrane in *E.coli* and cellulose in *A.xylinum* (Lu & Kleckner, 1994; Brautaset et al, 1994)

8.2.1.2 Proteobacterial phosphohexomutases

The phylogeny of the phosphohexomutase sequences reflects the evolution of the proteobacteria as determined by their 16sRNA sequences and DNA-rRNA hybridization studies (DeLey et al, 1990) (refer to figure 8.3). All the sequences show complete conservation of the sugar binding motif GEMSGH. Both the *N.gonorrhoeae* PGM and *P.aeroginosa algC* gene product have been shown to possess phosphoglucomutase and phosphomannomutase activity (Sandlin & Stein, 1994; Coyne et al, 1994). They also appear to be the only PGM/PMM protein in these bacteria, since no activity is detected when the gene is deleted or disrupted by site directed mutagenesis.

8.2.1.3 Enterobacterial phosphomannomutases

The third and fourth distinct evolutionary pathways of the prokaryotes consist primarily of enterobacterial phosphomannomutases. The third group comprises of PMMs from several serotypes of the enterobacteria *Salmonella enterica* and *E.coli* and, unexpectedly, the plant pathogen *Xanthomanas campestris*. These proteins, transcribed from the gene loci *rfbK, cpsG, pgm* and *xanA*, all show complete conservation of the GEMSAH sugar binding loop motif. The fourth pathway is represented by a PMM transcribed from *rfbK* in *S.enterica* group B. This gene is distinct from the other enterobacterial proteins, with no sugar binding loop motif easily identifiable. The promiscuity of the ancestral gene during evolution, in its capability to transfer between bacterial species and exchange between gene clusters, may provide an explanation for both the presence of the *xanA* gene product from *Xanthomanas campestris* and the divergent *rfbK* gene product.

The phylogenetic analysis indicates that the majority of *rfbK* and *cpsG* gene products are closely related. These PMMs are transcribed from two gene clusters *rfb* and *cps*. The *rfb* cluster contains genes invovled in the

Figure 8.3 Classification of bacteria included in the phylogenetic analysis. Only two of the eleven major phyla are represented.

PHYLUM	16s RNA SUBCLASS	RNA SUPERFAMILY	GENUS
Proteobacteria	α subclass	superfamily IV	Azospririllum
			Agrobacterium
	β subclass	superfamily III	Neiserria
	γ subclass	superfamily I	Enterobacteria: Salmonella Escherichia
			Coxiella
		superfamily II	Xanthamonas
			Pseudomonas
			Acetobacter
	δ subclass		
		superfamily VI	Heliobacter
Gram positive bacteria			Mycoplasma
Cyanobacteria			
Spriochaetes			
Gram negative anaerobic rods, cytophaga & flavobacteria		superfamily V	
Green sulphur bacteria			
Chlamydiae			
Planctomyces & relatives			
Deinococcus & relatives			
Green non-sulphur bacteria & relatives			
Thermotoga & relatives			

biosynthesis of the polysaccharide component of the O-antigen and the *cps* cluster contains genes responsible for the synthesis of the polysaccharide colanic acid or M-antigen (Reeves, 1993). The two clusters have been mapped to approximately the same region on the chromosomes of both *S.enterica* and *E.coli*, and in *S.enterica* they are separated by approximately 10kb (Stevenson et al, 1991).

Sequence analysis of prokaryotic genes has shown that the genomes of bacterial species have characteristic G+C contents. For example, genes from mycoplasmas have G+C contents of 0.23-0.41, genes from the Enterobacteriaceae *S.enterica* and *E.coli* have an average G+C content of 0.5 and genes from Pseudomonadaceaes have a G+C contents of 0.58-0.71 (Logan, 1994). The G+C content of the genes *rfbK*$_{C1}$ and *cpsG*$_{B}$, encoding the *S.enterica* PMM proteins is approximately 0.61. In *E.coli* genes *rfbK*$_{O7}$, *cpsG* and *pgm*$_{K12}$ the G+C content is 0.55. In both species, the higher than expected G+C content has been attributed to the horizontal transfer of the gene from a species with a characteristically high G+C content (Stevenson et al, 1991; Aoyma et al, 1994). However, due to the difference between the G+C contents in these two species it has been proposed that the transfer of the ancestral gene to *E.coli* occurred prior to its transfer into *S.enterica* (Aoyma et al, 1994). This proposed ability to transfer between bacterial genomes may account for the inclusion of the PMM gene *xanA* from *X.campestris* in this group.

The phylogenetically distinct *rfbK* gene from *S.enterica* group B represents the fourth pathway for the evolution of the prokaryotic phosphohexomutases. Although this protein is the only representative in the phylogenetic analysis, homologous genes to *rfbK*$_{B}$ have been identified in groups A, D and E1 by Southern blot analysis, (Verma et al. 1988; Wang et al, 1992) and in group C2, with the cloning of the gene (Brown et al, 1992). In contrast to the genes described in the previous paragraph, *rfbK*$_{B}$ has a low G+C content, of 0.40. This finding has also been suggested to be due to the acquistion of the gene by horizontal gene transfer, but from a species with a characteristically low G+C content (Stevenson et al, 1991).

The analysis of the G+C contents of these genes has allowed an insight into the evolution of the S.enterica *rfbK*$_{C1}$ gene (Lee et al, 1992b). Hybridization studies show that the *rfbK*$_{C1}$ and the *cpsG*$_{C1}$ genes are highly conserved. However, the 3' end of *rfbK*$_{C1}$ has a much lower G+C content than the rest of the sequence. Therefore, it is proposed by Lee et al that duplication of the

$cpsG_{C1}$ gene followed by recombination with the *rfb* gene cluster has given rise to the *$rfbK_{C1}$* gene, with the subsequent loss of the *S.enterica* archetypal *rfbK* gene. *cpsG* and *rfbK* genes have not, however, been isolated from a single *E.coli* group and therefore, whether both loci are present and whether similar recombination events may have occurred to generate two highly conserved loci remains to be determined.

8.2.1.4 *E.coli yhbf* and *Heliobacter pylori ureC*

The phylogenetic analysis identified *E.coli yhbf* and *H.pylori ureC* gene products as related sequences. Both proteins show conservation of the active site and magnesium binding loop motifs. However, the sugar binding loop motif is not conserved, even between these two sequences. The function of the *yhbf* and *ureC* gene products is unknown, but *ureC* is required for urease activity (Labigne et al, 1991). Therefore, these sequences may represent a change in function of the ancestral PGM gene.

8.2.1.5 *Mycoplasma pirum* PMM

The *M.pirum* PMM represents a further distinct pathway in the evolution of the prokaryotic phosphohexomutases. A sugar binding loop motif similar to those present in the other prokaryotic and eukaryotic phosphohexomutase proteins could not be identified. The gene was located in a cluster of genes involved in the salvage pathway of nucleotides, which led Tham et al (1993) to suggest that the protein may actually be a phosphopentomutase rather than a phosphohexomutase. However, no enzyme activity analysis was carried out.

8.2.2 EUKARYOTIC PGM AND PGM-RELATED SEQUENCES

Three branches of evolution are evident from analysis of the eukaryotic sequences. The first is a major group, consisting of six sequences (although one is bacterial) which are characterized by the presence of the GEESFG protein motif. The second is represented by the *S.cerevisiae* N-acetylglucosamine phosphomutase and the third by chloroplast PGM from *Spinacia oleracea.*

8.2.2.1 Human PGM1 homologues, paralogues and xenologues

The majority of eukaryotic PGM and PGM-related protein sequences cluster together in the third major evolutionary pathway identified by the phylogenetic analysis. The phylogeny identifies "recent" duplication events which have occurred in both yeast and humans. The yeast, *S.cerevisiae*, possesses two homologues of human PGM1. Following duplication and translocation, these paralogous genes have undergone a change in regulation (Boles et al, 1994). The major PGM isozyme, encoded by *PGM2*, is induced by galactose, whilst the *PGM1* gene is constitutively expressed at low levels (Oh & Hopper, 1990). In humans, the PGM1 paralogue PGMRP has undergone a change in function. PGMRP is located in the adherens-type cellular junctions and interacts with the cytoskeletal proteins dystrophin and utrophin (Belkin & Burridge, 1995). Although it shows a high level of identity with human PGM1 at the protein level (67.6%) (Moiseeva et al, 1996), and shows cross reactivity with the anti-PGM1 antibodies, it is unable to function as a phosphoglucomutase (Belkin et al, 1994).

The *Paramecium tetraurelia* PGM-like protein, parafusin, also appears to represent a change in function of the ancestral gene. It is a glucosylphosphotransferase acceptor protein, which undergoes rapid dephosphorylation upon stimulation of secretion (Satir et al, 1990). No PGM activity was found in parafusin enriched fractions, suggesting the presence of an additional gene, homologous to *PGM1* (Andersen et al, 1994).

The most unexpected result of the phylogenetic analysis was the clustering of the *A.tumefaciens* PGM with human PGM1 and parafusin. This sequence is more closely related to human PGM1 than the yeast PGM sequences, and it is quite distinct from the true bacterial PGM sequences. The node is well supported in the unweighted tree (90%) and receives stronger support from the weighted parsimony tree (99%). This unexpected topology may be explained by trans-kingdom horizontal gene transfer of a PGM gene. The criteria which should support the suggestion of horizontal transfer include: i) the sequences under consideration are from a wide range of species, ii) the rest of the tree should correspond to a conventional phylogeny and iii) more than one type of tree building programme should be used (Smith et al, 1992). All these criteria are fulfilled in the case of PGM: the analysis contains sequences from mammals, yeast, protozoa and bacteria, and *A.tumefaciens* is the only sequence which does not conform with the expected phylogeny, given by both the parsimony and distance methods.

When suggesting a horizontal gene transfer, the potential gene donor and gene acceptor should have contact so that the transfer is feasible. For this example, a mode of transfer for a gene from a eukaryote, perhaps a plant, to transfer to

A.tumefaciens is put forward, which although speculative, is plausible. *A.tumefaciens* is a gram negative soil bacteria which is able to induce crown galls at wound sites of a range of dicotyledonous plants. These crown galls or tumours are caused by the T-DNA of the tumour inducing (Ti) plasmid integrating into the nuclear genome of the plant. Transcription of the *onc* genes of the T-DNA causes cell proliferation. This is the only example of trans-kingdom DNA transfer occurring as part of the natural life cycle of an organism. However, it may also allow a method of transfer in the reverse direction from the plant to the bacteria. Of the two PGM isozymes expressed in plants, the gene encoding chloroplastic PGM of spinach has been cloned (section 8.2.2.2) (Penger et al, 1994). Since this shows a distinct evolutionary pathway it is suggested here that the cytosolic PGM protein will be of a more eukaryotic nature, homologous to human PGM1, and this may be the source of the sequence which has transferred into the *A.tumefaciens* genome.

8.2.2.2 Chloroplastic PGM

The chloroplastic PGM represents an alternative pathway in the evolution of the ancestral *PGM* gene in eukaryotes. The motif thought to correspond to the sugar binding loop is quite distinct and the protein is highly diverged from the other eukaryotic sequences. This may reflect its evolution as a chloroplastic PGM, which includes a 55 amino acid transit peptide at the amino terminus to allow transport into the chloroplast.

8.2.2.3 *S.cerevisiae* N-acetylglucosamine phosphomutase (AGM)

The AGM protein represents a further alternative pathway in the evolution of the ancestral gene. This sequence is the most diverged from the other sequences, but the active site (TASHNP) and magnesium binding loop (DGDADR) motifs are highly conserved. The absence of an identifiable sugar binding loop most probably reflects the change in sugar specificity to N-acetylglucosamine.

8.3 SUMMARY

i) The phylogenetic trees constructed using maximum parsimony and neighbour-joining distance methods gave the same basic topology. Three major groups of sequences, one consisting of primarily eukaryotic proteins, and the other two of prokaryotic proteins, were identified. These major nodes received strong support from resampling of the data using the bootstrap. There were also a number of sequences which did not fall into one of the three catergories. These six additional distinct phylogenetic groups may represent alternative evolutionary pathways of the ancestral gene. The phylogeny appears to be based primarily upon the number of amino acids and the distinctive sugar binding loop motif identified in the majority of the sequences.

ii) Phylogenetic analysis of prokaryotic sequences suggest a total of six possible evolutionary pathways to account for the variety of PGM and related sequences: i) the true bacterial PGMs, the most likely homologues of human PGM1, ii) the proteobacterial phosphohexomutases, which form one of the major groups, characterized by the presence of the GEMSGH motif, iii) the enterobacterial phosphohexomutases, another major group, characterized by the GEMSAH motif, iv) the single sequence encoded by the *rfbK* gene in *S.enterica*, group B, v) the cluster of two proteins, encoded by the *yhbf* and *ureC* gene loci, which may represent a change in function and vi) the mycoplasma PMM, which may actually be a phosphopentomutase.

iii) The eukaryotic sequences suggest three pathways of phosphohexomutase evolution: i) the major group of eukaryotic sequences, which also includes possible trans-kingdom horizontal gene transfer event involving *A.tumefaciens*, ii) the yeast AGM protein and ii) the chloroplastic PGM.

8.4 CONCLUSIONS

The phylogenetic analysis of the phosphohexomutase sequences indicates that a number of gene duplications, possibly as many as eight, may have occurred during the evolution of the ancestral gene to account for the diversity observed. The more diverse sequences may represent examples of convergent evolution. However, although examples of structural and mechanistic convergence have been reported, there are no reports of sequence convergence (Doolittle, 1994). Thus, these sequences are most likely to have evolved from a single ancestral gene. Evidence of more recent gene duplications, in both eukaryotes and prokaryotes, is also seen. In yeast, the duplication of the *PGM* gene is associated with a change in regulation, whereas in humans, duplication is associated with a change in function; PGMRP does not show PGM activity. In prokaryotes, recent duplications are seen among the enterobacteria at the *rfbK* and *cpsG* loci. In *S.enterica*, group C1, the *cpsG* gene appears to have duplicated and recombined in to the *rfb* gene cluster, resulting in the loss of the archetypal *rfbK* gene. In *E.coli* group 09, the *rfbK* gene has tandemly duplicated, to give the*rfbK1* and *rfbK2* loci.

The evolution of the prokaryotic loci considered in this study is quite complex, involving duplications, recombination and horizontal transfer of genes between bacteria. The total number of loci encoding PGM and PGM-related proteins is difficult to determine. In *N.gonorrhoeae* and *P.aeroginosa* only a single isozyme encoding both PGM and PMM was detected. However, *E.coli* appears to possess a number of loci: *cpsG, rfbK, yhbf* and *pgm*. Although it has not been demonstrated that *E.coli* possess both the *rfbK* and *cpsG* loci, the *yhbf* and *pgm* loci map to separate regions of the chromosome, suggesting at least three loci are present. Therefore, the greater distribution of loci in the enterobacteria may be partly due to the acquisition of genes from other bacteria, as suggested by the G+C content of the genes, rather than by a more conventional duplication of the ancestral gene.

The phylogenetic relationship of *A.tumefaciens* PGM to both the eukaryotic and prokaryotic sequences was investigated. It was found to cluster with the two human sequences, PGM1 and PGMRP, and parafusin from *P.tetraurelia*. It was shown to be more closely related to these sequences than either of the yeast PGM sequences were, suggesting an unconventional descent. This received further support from the identification of two true bacterial PGM homologues. If *A.tumefaciens* PGM had evolved from the prokaryotic PGM1 homologue, it would have been expected to group with these two sequences. Therefore, it is proposed that the *A.tumefaciens PGM* gene was acquired by a trans-kingdom horizontal transfer event. The criteria for proposing the existence of such an event are met, and since this bacteria is a plant pathogen, the source of the eukaryote-like gene may have been a dicotyledon plant.

The main aim of the phylogenetic analysis was to investigate the evolution of the phophohexomutases and to identify alternative pathways which may represent the evolution of PGM2 and PGM3. A total of eight phylogenetically distinct pathways have been identified, and two of these may represent the evolution of PGM2 and PGM3. First, the yeast AGM sequence provides evidence of an alternative pathway in the evolution of eukaryotic sequences. If one of the other human PGM isozymes is a candidate for the human homologue, it is more likely to be PGM3, since i) they are both monomers, ii) the molecular weights of the two proteins are comparable; AGM 62,000mw, PGM3 estimated 65,000mw and iii) both AGM and PGM3 are poor phosphoglucomutases.

The second pathway, which may represent the evolution of PGM2, is shown by the mycoplasma PMM sequence. This is thought to represent a phosphopentomutase, since it is located on a DNA fragment encoding genes involved in the salvage pathway for nucleosides. In man, PGM2 is thought to be the true phosphopentomutase, in addition to possessing phosphoglucomutase and phosphomannomutase activities. Although the sizes of the two proteins are not comparable, *M.pirum* PMM is 61,400mw (Tham et al, 1993) and PGM2 is 71,000mw, the *M.pirum* PMM is larger than the majority of the PGM and PMM proteins in the gram negative bacteria. The difference in size may reflect the evolutionary divergence between these two species, with mycoplasmas under greater selection pressure to maintain a small genome size.

To determine if these interpretations are correct, further homologous sequences from a variety of species are required. Although the active site and magnesium binding loop motifs are conserved in these proteins, conserved motifs characteristic of the AGM and putative phosphoribomutase proteins may be identified. These could then be the basis for gene identification strategies, such as those presented in this thesis. From these investigations, the method of choice would be to search the EST databases with the entire protein sequence or conserved motifs. Identification of any human related sequences may relatively quickly lead to the molecular characterization of the sequence.

CHAPTER NINE:

DISCUSSION

The research described in this thesis focuses on the molecular and evolutionary investigations of the phosphoglucomutase gene family. The gene for *PGM1* was cloned and characterized four years ago (Whitehouse et al, 1992; Putt et al, 1993) and the main focus for my research was the investigation of approaches for cloning other members of the gene family. The strategies included the use of antibodies raised against PGM1, low stringency PCR, degenerate primer PCR and identification of ESTs with homology to PGM1. The advantages and disadvantages of each of these approaches and the resources used, will be discussed, with respect to the specific investigation of PGM and gene identification in general.

A number of novel PGM-related sequences have been identified either by degenerate primer PCR or by searching EST databases. Two of these sequences have been investigated and partially characterized. Although the sequence identified by degenerate primer PCR, *hyhbf*, is a member of the PGM gene family, its origin is not certain. Preliminary molecular characterization of the human ESTI sequences indicate there may be more than one homologous locus and there is also evidence of alternative transcripts. Recent mapping data suggests that two of the loci may be localized on chromosomes 4 and 7. Therefore, one of the genes homologous to human ESTI may be *PGM2*; circumstantial evidence for this is discussed.

The evolution of the mammalian *PGM1* gene has been investigated at the nucleotide level. Comparative studies of exons 1, 4, 8 and 11, (which contain genetic polymorphisms in the human population), and exon 5 were carried out on man, great apes (gorilla, chimpanzee and orangutan), rabbit, rat and mouse. The data confirm the previous hypothesis that the PGM1*1+ protein allele is ancestral and also indicate an extremely high level of nucleotide sequence conservation among the hominoids. Conservation at the protein level is also evident among more distantly related species, for example in frog, as determined by immunological criteria.

Phylogenetic analysis of PGM and PGM-related sequences was carried out to identify possible alternative evolutionary pathways of the ancestral PGM gene along which *PGM2* and *PGM3* may have evolved. During this analysis an example of a trans-kingdom horizontal gene transfer event may have been identified. The *A.tumefaciens PGM* is far more eukaryotic in nature than prokaryotic, and shows a greater identity with human *PGM1* than either of the paralogous genes in yeast. This example will be compared with other claims of horizontal gene transfer. In addition, two PGM-related genes, parafusin (*PFUS*) and N-acetylglucosamine phosphomutase (*AGM*), reported from paramecium and yeast respectively, are considered as candidate paralogues for human *PGM3* (and *PGM2*). Finally, the possibility of convergent evolution, rather than divergent evolution, giving rise to the three PGM isozymes is discussed.

9.1 RESOURCES USED FOR GENE IDENTIFICATION

The primary aim was to identify and clone partial cDNAs for the constitutively expressed genes *PGM2* and *PGM3*. The principal sources of cDNA were the erythroleukaemic cell line K562 and the K562 and placental cDNA libraries obtained from the HGMP Resource Centre. These were thought to be ideal resources: K562 showed a marked reduction in PGM1 transcripts and an increase in activity of the PGM2 and PGM 3 isozymes (Chapter Three), and placental extracts showed a relatively high level of PGM3 expression in comparison to most other tissues.

9.1.1 THE HGMP cDNA LIBRARIES

The two cDNA libraries were used as template DNA for degenerate primer PCR. However, the only PGM-related sequences obtained from these libraries were of bacterial origin suggesting that the degenerate primers annealed preferentially to bacterial DNA in the plasmid preparations. The inability to select recombinant plasmids containing human *PGM1* or other PGM-related sequences from these libraries using degenerate primer PCR, may reflect the inherent low efficiency of this strategy; of 291 plasmids analyzed, only two contained novel PGM-related sequences, and these were identical. It was also shown that the average insert size of both libraries was approximately 500bp. Since the libraries were constructed using oligo dT primers, the majority of cDNA inserts would be from the 3' end of the mRNA transcripts. The forward degenerate primers, based on the active site motif, anneal near to the 5' end of the *PGM1* gene; therefore inserts of approximately 1900bp would be required for successful amplification of *PGM1*. Thus, the apparent low abundance of full-length inserts may also, in part, explain why *PGM1* and PGM-related sequences were not identified. This idea was supported by PCR results from the *PGM1* cDNA primer pairs 2, 3, 4, 5 and 6 (Figure 3.16). Amplification of *PGM1* from the cDNA libraries was only achieved with forward primers sited downstream of the active site region (pairs 3, 4, 5, and 6). However, no products were obtained using primer pair 2, in which the forward primer is sited just upstream of the active site. This suggests the largest PGM1 transcripts are of approximately 1600bp.

9.1.2 THE K562 ERYTHROLEUKAEMIC CELL LINE

The absence of PGM1 activity in the K562 cell line is a unique and specific characteristic. The molecular basis of this deficiency was shown to be associated with a marked reduction in the level of PGM1 mRNA transcript. Thus, this cell line was thought to be an ideal resource for the cloning of PGM-related genes. Deficiency of PGM1 is also associated with an increase in the activities of PGM2 and PGM3. This is in contrast to null PGM1 phenotypes observed in man, where there is no associated increase (Ward et al, 1985). Thus, this may also be a specific characteristic of K562.

In the K562 RT-PCR experiments, following cDNA synthesis primed using random hexamers, the reduction in PGM1 mRNA transcripts was expected to lead to a reduction in the ratio of *PGM1* to PGM-related sequences amplified, thereby improving the chances of amplifying *PGM2* and *PGM3*. However, no PGM-related sequences were identified. This may have been due to the inefficiency of the degenerate primer PCR strategy, with PGM-related sequences amplified but not cloned or cloned but not selected (section 9.2.3). Alternatively, it may reflect a greater divergence of *PGM2* and *PGM3* than expected.

SSCP and restriction enzyme analysis of K562 genomic DNA determined the putative *PGM1* genotype to be 2+1+. Cytogenetic analysis, however, identified three chromosome 1s, each carrying the *PGM1* gene. Since restriction enzyme analysis of the PGM1 transcripts showed unequal expression of the alleles, with the *PGM1**2 allele expressed at greater levels than *PGM1**1, the proposed genotype is 2+2+1+. The expression from the *PGM1**2 allele compared to the *PGM1**1 is, however, greater than the 2:1 ratio expected. It is suggested that this may be due to a trans-acting element affecting all three genes, but not equivalently. This possiblity may be supported by a report of a K562 subclone K562[S]P in which *PGM1* has been reactiviated (Ravazzolo et al, 1985). Following cellulose acetate electrophoresis, activity staining of the gel revealed a single PGM1 band in K562[S]P, which was not present in the standard K562 cell line. The authors did not comment on the phenotype of PGM1, but our interpretation is that the phenotype of the K562[S]P subclone is *PGM1**2. This may be explained by either the trans-acting element allowing transcription from the *PGM1**2 allele(s) but not from the *PGM1**1 allele or by a proportional increase in the activity of all three genes, such that only the PGM1*2 isozyme could be detected.

The cytogenetic analysis provides evidence of an evolving karyotype in K562. The cell line was originally characterized by possession of a Philadelphia chromosome. However, more recent published karyotypes do not appear to show this marker (Ajmar et al, 1983; Selden et al, 1983). Cytogenetic analysis of our cell line, using chromosome specific paints and fluorescence in-situ hybridization, has identified the Philadelphia chromosome as a duplicated acrocentric marker (Fox et al, 1996). Thus there is evidence of genetic instability in this cell line. Therefore, although K562 has been shown to express PGM2 and PGM3, the gross structural rearrangements which are found are likely to affect gene transcription of numerous genes, including, possibly other PGM-related genes.

Genetic instability of lymphoblastoid and lymphoma lines has previously been investigated by isozyme analysis (Povey et al, 1980). It was found that lines derived from patients with Burkitts lymphoma (BL), which possess chromosomal rearrangements, often lose gene function. This is exemplified by the BL line JIJOYE, which was originally heterozygous for PEP A, APRT, and ACP_1, and positive for PEP D. Subsequent analysis has shown cultures homozygous for PEP A, APRT X, ACP_1B, and/or negative for PEP D. Therefore, Povey and colleagues suggested that other lymphoma and leukaemia lines may also show genetic instability such that they are unrepresentative of the tumour cells from which they are derived. This appears to be the case with K562.

9.1.3 OTHER RESOURCES

RNA was available from a number of lymphoblastoid cell lines in which *PGM2* and *PGM3*, are both constitutively expressed. During the project it became evident that there were PGM-related genes, such as *PGMRP*, which are tissue specific, and others, such as the gene represented by human ESTI, which show alternative transcripts, in specific tissues. Therefore, the availability of RNA from a wider range of tissues could have increased the chance of identifying PGM-related genes.

9.2 GENE IDENTIFICATION APPROACHES

All the laboratory approaches were based on the assumption that *PGM1, PGM2* and *PGM3* are members of a diverged gene family whose protein products would show conservation of epitopes, or peptide motifs such as those located in the active site cleft of the protein. These include the active site loop (TASHNP), the magnesium binding loop (DGDGDR) and the putative glucose binding loop (GEESFG).

9.2.1 IMMUNOLOGICAL APPROACHES

Anti-rabbit PGM polyclonal antibodies have been shown to immunoprecipitate PGM1, but not PGM2 or PGM 3 (Drago et al, 1992). Two anti-human PGM1 specific polyclonal antibodies, anti-6' PGM and anti-10' PGM, were investigated to determine their immunoreactivity with the PGM2 and PGM3 isozymes. If the PGM gene family in man represents a relatively recent divergence of the loci, these antibodies may recognize human-specific epitopes shared between PGM1, PGM2 and PGM3. The anti-6' PGM antibodies were raised against a fusion protein containing most of domain 4 of PGM1, whilst anti-10' PGM antibodies were raised against a fusion peptide containing containing domains 2, 3 and 4 (Figure 2.1); domain 2 includes the magnesium binding loop motif, and domain 3, the glucose binding loop. The antibodies recognized both human and rabbit PGM1. However, no immunoreactivity was observed between the antibodies and the PGM2 and PGM3 isozymes.

The anti-PGM1 antibodies are, however, capable of identifying other members of the human PGM gene family. For instance, the anti-rabbit PGM polyclonal antibodies recognize PGMRP, the PGM1 related protein previously known as aciculin (Critchley & Whitehouse, personal communication). The *PGMRP* gene is a paralogue of *PGM1* and the protein, which is catalytically inactive, has evolved a new function as a structural element in the adherens-type cellular junctions (Moiseeva et al, 1996). Since strong immunoreactivity is observed between anti-PGM1 antibodies and the PGMRP protein, and PGM1 shows 68% amino acid identity with PGMRP, PGM2 and PGM3 are likely to be far less than 68% identical to PGM1. A *PGMRP* homologue has been identified in mice, suggesting the evolution of *PGMRP* predates mammalian radiation. This would imply that the lineages for PGM1, PGM2 and PGM3 were established at a very much earlier point in evolutionary history.

9.2.2 LOW STRINGENCY PCR

The low stringency PCR experiments were based upon the use of primers to two conserved regions of the PGM1 protein; the active site and the magnesium binding loop. Since these motifs are highly conserved at the nucleotide level among the eukaryotic PGM-like sequences, it was thought that they may also be conserved in PGM2 and PGM3. The K562 cell line has been shown to possess very low levels of PGM1 mRNA transcript, and thus, in RT-PCR experiments, the ratio of *PGM1* to PGM-related sequences in the cDNA pool would be expected to be less than in controls. Therefore, low stringency PCR was carried out on K562 total RNA and control cell lines 6997 and 7014. Any products amplified from any of the three samples of a different size to *PGM1*, or of a similar size from K562, would be worthy of further investigation.

The primary screening procedure for these products used the HPGM1 probe at low stringency on Southern blots to search for closely related sequences. However, only *PGM1* was identified from the two control cell lines, with no hybridization signals evident from K562. At first sight the negative finding in K562 contradicts the previous observation of low levels of RT-PCR products reported in Chapter Three. This apparent discrepancy can be explained by the quantity of cDNA used for each of the experiments: for low stringency PCR (and also for degenerate primer PCR) only a fifth of the cDNA reaction was added to the PCR, whilst for the characterization of PGM1 transcripts in K562, the entire cDNA reaction mix was used.

The low stringency PCR results suggest that sufficient divergence has occurred to prevent amplification of *PGM2* and *PGM3.* Although there may be amino acid conservation between PGM1 and PGM2 and PGM3, nucleotide sequences appear to be less conserved. Mismatches between the 3' end of the primer and the target sequence will in general lead to reduced amplification. The forward primer (Ser116) is derived from the amino acid sequence ILTASHNP. The proline residue is completely conserved in all species and the 3' end of the primer (-CC 3') allows any of the four proline codons to anneal. In contrast, the reverse primer (MgR) is derived from the amino acid sequence AAFDGDGDR. The two alanine residues are a eukaryotic feature, being found in mammals, yeast and parafusin. At the amino acid level it is less likely that these residues will be conserved in PGM2 and PGM3. Thus, the 3' end of the reverse primer (-GGCAGC 3') is likely to possess a number of mismatches with the templates. With hindsight, it might have been better to design the reverse primer to encode one of the highly conserved residues, such as Asp^{287} at the 3' end (-ATC 3').

Low stringency PCR would, however, be expected to amplify highly conserved PGM-related sequences such as the *PGMRP* gene from the appropriate RNA source. Comparison of the primer sequences with the *PGMRP* cDNA identifies four nucleotide substitutions in each of the primers, (although not at the 3' end) and this would not be expected to prevent amplification of *PGMRP*. For instance, Scharf et al, (1986) reported the amplification of allelic variants in the HLA DQα gene in which there were eight mismatches in the 26-mer forward primer and seven mismatches in the 28-mer reverse primer. However, the RNA sources for the low stringency PCR were all cell lines: K562 is erythroleukaemic, 6997 and 7014 are lymphoblastoid, and expression studies of PGMRP show that the protein is generally found in visceral and vascular smooth muscle (Moiseeva et al, 1996). Therefore, expression from *PGMRP* might not have been expected in any of the cell lines used, and this would explain why this gene was not amplified. In summary, the low stringency approach was found to be unsuitable for the cloning of *PGM2* and *PGM3* athough closely related sequences, such as *PGMRP*, should be identifiable using the appropriate cDNA sources.

9.2.3 DEGENERATE PRIMER PCR

The principle of degenerate primer PCR is to use primers that allow for nucleotide divergence whilst retaining the amino acid sequence of the highly conserved PGM protein motifs of the active site and the magnesium binding loop. With the publication of further PGM and PMM cDNA sequences, the primers were modified to incorporate limited amino acid changes, so that redundancy was allowed for both at the nucleotide and amino acid level.

This technique was successful in amplifying *PGM1* in man, its homologues in rat and *E.coli*, and an apparently novel PGM-related sequence, *hyhbf*, from human RNA. (This sequence will be discussed in section 9.3.1.) However, the two most significant problems encountered using this strategy were the inefficiency of the technique and the identification of suitable motifs on which to base the degenerate primers.

There are a number of modifications which could increase the efficiency of the technique. First, size selection of the PCR products could be employed; thus instead of using a sample of the PCR product directly, the products could be separated by electrophoresis and DNA of the 'expected' size, in this case between 250bp and 750bp, could then be extracted and cloned. Alternatively, size selection may be applied by estimating the size of the insert in a recombinant plasmid by PCR, prior to the preparation of plasmid DNA.

Another way of improving efficiency might have been to screen the recombinants with an oligonucleotide probe which could hybridize to related sequences. A probe based upon sequence between the active site and the magnesium binding loop would have been ideal. However, this region is not highly conserved between eukaryote and prokaryote PGMs and PMMs, and there were no distinctive motifs present. Therefore, this approach was not thought to be suitable for the identification of PGM-related sequences.

The complete *PGM1* cDNA, (HPGM1), was considered as a screening tool. However, Southern blot analysis with low stringency hybridization failed to identify any bands in addition to *PGM1.* Together with the low stringency PCR results and the immunological studies, this suggests that the level of divergence between *PGM1* and the other isozymes is far too great to allow identification by these homology based procedures. Further, during an attempt to locate *PGM3* by low stringency hybridization of the HPGM1 probe to a chromosome 6 flow-assorted genomic library, a single clone was identified which was shown to be a contaminant chromosome 9 (Ives, 1995). The sequence corresponded to exon 5 of the *PGMRP* gene (Moiseeva et al, 1996). Thus the HPGM1 probe detected no related sequences derived from chromosome 6. In summary, the HPGM1 probe is most unlikely to be suitable for the identification of *PGM2* and *PGM3.*

A modification which may have increased the probability of obtaining PGM-related sequences would be repeated transformations from the same PCR product. By doing this, less abundant cDNAs may have been cloned. The efficiency of degenerate primer PCR strategy may also have been increased if the appropriate source of RNA was used, such that PGM-related sequences are expressed in the cells from which the RNA is extracted.

The other main problem with the technique concerns the identification of protein motifs on which to base the degenerate primers, as illustrated by the N-acetylglucosamine phosphomutase (AGM) degenerate primers. In both normal and nested degenerate primer PCR, primers based on the *Saccharomyces cerevisiae* AGM protein sequence, identified no human AGM-related sequences. However, this may not indicate the absence of an AGM homologue, since a distinct AGM protein has been identified and partially characterized from mammalian tissue (Fernandez-Sorensen & Carlson, 1971).

Recently, a partial peptide sequence of a putative AGM protein was submitted to Swissprot from the yeast *Schizosaccharomyces pombe* (Acc. No. Q09687). Amino acid sequence analysis with AGM from *S.cerevisiae* shows 46% identity between the two sequences (Figure 9.1). Divergence between the two pathways which gave rise to these two yeasts is estimated to have occurred around the same time as the pathway which gave rise to mammals (Sprague, 1991). Therefore, this would suggest that the human homologue also shares approximately 50% identity with *S.cerevisiae* AGM. In addition to the active site and the magnesium binding loop motifs which are completely conserved, there are also additional segments of amino acids showing conservation. The location of the AGM degenerate primers was compared with these segments, to determine if the primers encode conserved amino acids.

Degenerate primer PCR with the first set of AGM specific primers used the forward primer DegAGMF1 and the reverse primer DegAGMR1. DegAGMF1 was sited over the magnesium binding loop, and therefore the amino acid sequence encoded by the primer was conserved, with the exception of the most 5' encoded residue (Figure 9.2). In contrast, the peptide encoded by the DegAGMR1 primer shows two amino acid changes. Thus, this primer is not sited over a conserved segment of the protein. Comparison of the peptides encoded by the nested degenerate primers also shows that DegAGMF2 does not lie in a conserved region, with three amino acid changes observed between the primer peptide and *S.pombe*. Thus, these AGM-specific primers were not ideally sited for degenerate primer PCR. Therefore, redesigned primers, based on the segments of conserved amino acids, may amplify an AGM homologue in man.

Figure 9.2 Diagram comparing the peptides encoded by the AGM-specific degenerate primers with the the corresponding peptides in the AGM protein from *S.pombe*.

Primer	Peptide encoded by primer sequence	Peptide sequence in *S.pombe*
DegAGMF1	**F**DGDADR	**I**DGDADR
DegAGMR1	D**M**LA**V**	D**L**LA**T**
DegAGMF2	G**IL**A**V**	G**VA**A**A**
DegMgR2	GD(G/A/F)DR	GDADR
DegSer116F	(G/A)SHNP	ASHNP
DegAGMR2	GAD**Y**V	GAD**F**V

Figure 9.1 Amino acid sequence comparison of AGM proteins from *S.pombe* and *S.cerevisiae*. The active site and the magnesium binding loop motifs are shown in bold.

```
                     .         .         .         .         .
S.pombe   1 MTKNKKYSYGTAGFRTKASDLEAAVYSSGVAAALRSMELKGKTIGVMITA 50
            .||| .:||||||||| |.:|:..::|.|: |.|||:.|.|. :||||||
S.cere   17 RTKNVQFSYGTAGFRTLAKNLDTVMFSTGILAVLRSLKLQGQYVGVMITA 66

                     .         .         .         .         .
S.pombe  51 SHNPVEDNGVKIIDADGGMLAMEWEDKCTQLANAPS........KAEFDF 92
            |||| :||||||:::||:||  .||. . |||||:|        :.|:.
S.cere   67 SHNPYQDNGVKIVEPDGSMLLATWEPYAMQLANAASFATNFEEFRVELAK 116

                     .         .         .         .         .
S.pombe  93 LIKQ..FLTPTTCQPKVIIGYDTRPSSPRLAELLKVCLDEM.SASYIDYG 139
            ||.:  :   ||. |.:::| |.|.||| |   |. ::..:  |  :|.|
S.cere  117 LIEHEKIDLNTTVVPHIVVGRDSRESSPYLLRCLTSSMASVFHAQVLDLG 166

                     .         .         .         .         .
S.pombe 140 YITTPQLHWLVRLINKSTAASFLEEGPPITEYYDTLTSAFSKIDPS..MQ 187
            ::||||||::. | |:.. .:  ...:.  :||. :.:||..: :.  ::
S.cere  167 CVTTPQLHYITDLSNRRKLEGDTAPVATERDYYSFFIGAFNELFATYQLE 216

                     .         .         .         .         .
S.pombe 188 DSPTVSRVVVDCANGVGSQPLKTV...AGLVKDSLSIELVNTDVRASELL 234
            .. .|.::.:|.|||:|:..||.:   .::  ..  :|::|.   ..|||
S.cere  217 KRLSVPKLFIDTANGIGGPQLKKLLASEDWDVPAEQVEVINDRSDVPELL 266

                     .         .         .         .         .
S.pombe 235 NNGCGADFVKTKQSPPLALEGKIKPNQLYASIDGDADRLIFYYINQNRKF 284
            | :||||:|||.|. | :|... . : ||.|:||||||::|||:: ..||
S.cere  267 NFECGADYVKTNQRLPKGLSPS.SFDSLYCSFDGDADRVVFYYVDSGSKF 315

                     .         .         .         .         .
S.pombe 285 HLLDGDKISTALVGYLNILVKKSGMPFSL..GVVQTAYANGASTEYLQD. 331
            |||||||||| :. :|.  :. . :. ||  ||||||||||.||.|:.:
S.cere  316 HLLDGDKISTLFAKFLSKQLELAHLEHSLKIGVVQTAYANGSSTAYIKNT 365

                     .         .         .         .         .
S.pombe 332 LGITTVFTPTGVKHL.HKAAKEFDIGVYFEANGHGTVLFSDKALANLAHP 380
            | ...  |.|||||| |.||.::|||:|||||||||::||:| : .  ..
S.cere  366 LHCPVSCTKTGVKHLHHEAATQYDIGIYFEANGHGTIIFSGK.FHRTIKS 414

                     .         .         .         .         .
S.pombe 381 FFTPSPVQAA..AIEQLQSYSVLINQAIGDAISDLLATISVLNALHWDAS 428
             :..|.:.:.  |:  |.::| ||||.:||||||:||.:..|. |. .:
S.cere  415 ELSKSKLNGDTLALRTLKCFSELINQTVGDAISDMLAVLATLAILKMSPM 464

                     .         .         .         .         .
S.pombe 429 AWSNTYKDLPNKLAKVKVSDRTIYKSTDAERRLVSPDGLQEKIDALVAKY 478
            .|.:.|.||||||.|. |.||.|:..||.||:|:.| |||:||| :||||
S.cere  465 DWDEEYTDLPNKLVKCIVPDRSIFQTTDQERKLLNPVGLQDKIDLVVAKY 514

                     .         .         .
S.pombe 479 EKGRSFVRASGTEDVVRVYAEASTKQAADELCEKVCQLV 517
            ..||||||||||||.||||||....   :::|:.|.: |
S.cere  515 PMGRSFVRASGTEDAVRVYAECKDSSKLGQFCDEVVEHV 553
```

In summary, degenerate primer PCR was successful in identifying *PGM1* homologues in rat and *E.coli*, as well as a novel PGM-related sequence, *hyhbf*. For optimal results, a number of conserved protein motifs are required for its success, since nested degenerate primers can increase the specificity of the PCR and therefore increase the efficiency of the technique.

9.2.4 IDENTIFICATION OF ESTs

Expressed PGM-related sequences were identified by searching the EST databases with the human PGM1 amino acid sequence and this strategy has so far proven to be the most efficient for identifying PGM-related sequences. Three PGM-related sequences were identified, one of which, human ESTI, has been characterized further. The partial cDNA sequence available from the 5' EST enabled the rapid characterization of the sequence by RT-PCR, genomic PCR, Southern blot and Northern blot analysis

An EST becomes a tool with which to carry out further searches of the EST databases. This may result in the identification and assembly of further cDNA sequences from the same candidate gene. For example, in the case of human ESTI, both 5' and 3' nucleotide sequences were available: the 5' nucleotide sequence encodes the protein in the region of the active site, and the 3' nucleotide sequence the 3' untranslated region (3' UTR). Both ends of the clone were used to search the databases for other ESTs (Figure 9.3), and the 3' sequence identified an additional EST (130882) in which the 3' nucleotide sequence was almost identical. The 5' sequence of clone 130882 was translated and found to encode a peptide which included a putative magnesium binding loop motif DPDADR. Thus, a partial cDNA sequence has been assembled for human ESTI (Figure 9.4).

It appears EST analysis is powerful and efficient, but there are limitations. There may not be any novel EST clones in the database which show significant homology to the query sequence. Alternatively, a protein motif encoded by the EST clone may have diverged significantly from the query peptide and thus go undetected, even though it remains functionally conserved. This is exemplified by the EST clone encoding the magnesium binding loop DPDADR, which was not detected by PGM1. Therefore, in hindsight, it would be wise to extend the search by including all possible combinations of amino acid changes conserved within a motif. Furthermore, the single-pass automated nucleotide sequencing results in an approximately 3% error or base ambiguity rate (Boguski et al, 1993). The presence of undetermined nucleotides restricts the choice of PCR

Figure 9.3 Strategy for EST database searches.

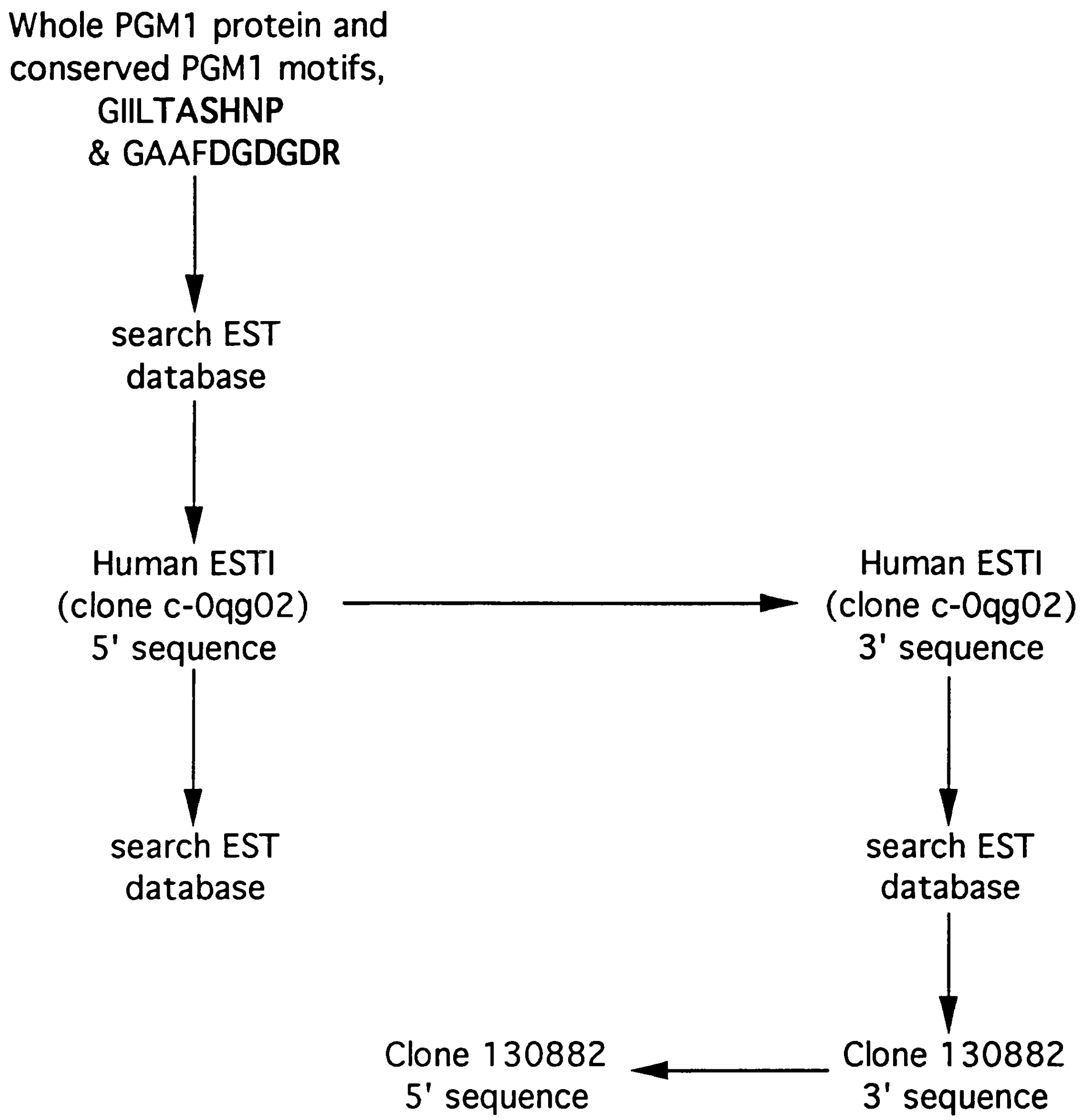
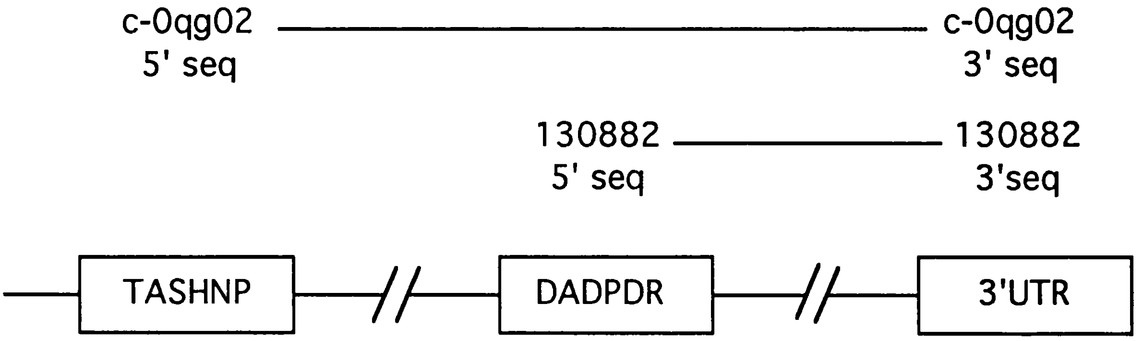


Figure 9.4 Partial cDNA sequence assembled for human ESTI.



primers, whilst cryptic sequencing errors, that coincide with the 3' end of the primer, will be significant.

Some sequencing errors will introduce stop codons into the sequence, as appears to be the case for the human EST homologous to yeast *sec53*, the gene encoding PMM, which is distinct from the other cloned phosphohexomutases. The human PMM EST clone was sequenced from both ends providing both a 5' and 3' sequence. The 5' sequence showed 57% identity with the *sec53* gene over 124bp (bestfit), whilst the 3' sequence was more highly diverged, as would be expected in the 3' UTR of homologous genes, showing 68% identity over only 28bp (bestfit). Translation of the 5' sequence revealed a stop codon in frame with the putative PMM protein. Since none of the six frames of the EST were open, it was concluded that sequencing errors must have occurred.

A third problem is contamination of the databases with vector and other spurious sequences, such as yeast and bacteria. Whilst every precaution was taken to filter out contaminants, usually by rigorous database searching with the new ESTs, inevitably some contaminants escape detection. This is exemplified by a comment included with the data output for human ESTII from dbEST; "Computer analyses of the total data set derived from this library [T-lymphoblastoid cell line ATCC-CCL119] indicate a significant proportion of sequences of yeast and bacterial origin."

In summary, the strategy to identify EST sequences which show amino acid conservation with human PGM1 has proved successful, with the identification of three PGM-related sequences and the assembly of a partial cDNA sequence for human ESTI. As the EST databases expand, there may be instances when it becomes possible to assemble an entire cDNA sequence using only the computer. The EST strategy has proved to be the most efficient of the four approaches investigated to identify PGM-related sequences.

The EST approach is well suited for the identification of genes in which other members of the gene family, or a partial amino acid sequence, is known, although the primary purpose of the EST projects is the identification of genes and their position in the genome. Precise mapping of all the characterized sequences in the integrated genetic, cytogenetic and physical maps of the human genome would provide a comprehensive resource for the positional candidate approach (Collins, 1995), which would supercede positional cloning of genes. At present, however, only a small proportion of the EST clones have been mapped, and therefore this is a limiting step in gene identification, (Houlgatte et al, 1995), particularly when a map position is used as a method to search the EST databases.

The coverage of human genes represented by ESTs is difficult to estimate since it is not known how many genes are in the human genome, nor how many genes are matched by ESTs. However, in a recent assessment of 87,983 non-overlapping ESTs, 10,214 ESTs matched 2,947 known genes, providing a coverage of 72% of the sequences in a non-redundant dataset compiled from Genbank (Adams et al, 1995). Allowing for the bias towards abundantly transcribed genes, and the sample size of the Genbank data set (4,100 genes), it was estimated that the non-overlapping ESTs described represented as many as 50% of the genes in the human genome. Complete sequences and precise mapping data in addition to detailed expression and functional studies of these ESTs will provided a more accurate estimate.

9.3 NOVEL HUMAN PGM-RELATED SEQUENCES

Two novel PGM-related sequences were identified. The first sequence, *hyhbf*, was obtained by using the degenerate primer PCR strategy whilst the second, human ESTI, was found by searching the EST databases. The origin of the *hyhbf* sequence and the possibility that human ESTI represents a candidate for *PGM2* are discussed below.

9.3.1 *HYHBF* - A NOVEL HUMAN PGM-RELATED SEQUENCE?

The novel PGM-related sequence, *hyhbf*, represents a partial cDNA encoding an ORF of 149 amino acids. Comparative sequence analysis of eukaryotic and prokaryotic phosphohexomutases identify conservation of key residues in addition to those of the active site and magnesium binding loop, suggesting *hyhbf* is a member of the PGM gene family. However, the molecular characterization of the sequence was inconclusive. Amplification of *hyhbf* was independent of the presence of reverse transcriptase in RT-PCR experiments; since the same size band was produced from genomic DNA, it was thought that the sequence originates from human genomic DNA, rather than RNA. However, Southern blot analysis to confirm this hypothesis was prevented by the high G+C content of the sequence.

Hyhbf showed a 64.6% identity with the *yhbf* gene from *E.coli* which is higher than expected from organisms of such diverse origins. The G+C ratio was

0.61, which is much higher than the average 0.4 for coding DNA in man. These data may suggest that the sequence originated from bacterial contamination. It was thus critical to attempt to determine the source of *hyhbf.*

9.3.1.1 Extraneous DNA in the Initial PCR

The initial RT-PCR of JG total RNA from which *hyhbf* was first isolated may have been contaminated, perhaps by a laboratory strain of bacteria such as *E.coli* RR1. However, this was thought to be unlikely since the reaction mixes are set up in one room, the PCR machines are located in another, and all work involving bacterial cultures is carried out in a third. In addition, no-DNA control PCR reactions set up simultaneously did not show any prominent PCR products, suggesting contamination had not occurred.

Alternatively, the JG RNA preparation may have been contaminated prior to the PCR. However, RT-PCR experiments with JG RNA samples prepared at different times both from cultured B cells and whole blood showed the expected 260bp PCR product. Amplification of the expected size band from whole blood rules out the possibility of contaminating mycoplasma (in the cell line) as a source.

9.3.1.2 Accidental Introduction of Extraneous DNA to the PCR Reactions

Following the initial amplification, *hyhbf* was cloned and primers designed to carry out RT-PCR and standard genomic DNA PCR. The results of these experiments were not reproducible, which might indicate the chance introduction of the clone, or subsequent PCR products. However, since RT-PCR required the use of nested primers to amplify *hyhbf*, yet genomic DNA PCR did not, this seems unlikely.

The addition to the reaction mix of first round PCR products in nested RT-PCR and genomic DNA in standard PCR was carried out in the laboratory. To determine if contamination was caused by aerosols of extraneous DNA in the laboratory, standard PCR reactions in which no DNA was added were exposed to the air for 5 seconds to allow for chance contamination to occur. In addition, the pipette that had been used to prepare the *hyhbf* sequencing reactions was used to add water to no-DNA controls. In both of these cases, no PCR products were observed.

Although there is no experimental evidence to suggest that the *hybhf* sequence was bacterial in origin, in view of the high identity it shows with *E.coli yhbf*, the high G+C content of the sequence and the inconclusive molecular characterization, this remains a possible explanation.

9.3.2 HUMAN ESTI - A NOVEL MEMBER OF THE PGM GENE FAMILY: A CANDIDATE FOR *PGM2*?

The EST databases were searched using the PGM1 amino acid sequence, and two homologous sequences were identified; one from human and the other from pig. The peptide encoded by the 5' sequence from these EST clones showed approximately 30% identity to human PGM1, including complete conservation of the active site motif TASHNP. Therefore, it was thought that these sequences represented further members of the PGM gene family. Molecular characterization of human ESTI was carried out at both the DNA and RNA level. RT-PCR using primers based on the 5' sequence amplified the expected size PCR product from human RNA. No band of this size was produced from genomic DNA, indicating that the sequence is transcribed. Northern blot and Southern blot analyses were also performed.

The RT-PCR product was used as a hybridization probe for Northern blot analysis and detected four transcripts of 4.5kb, 2.4kb 1.6kb and 1.35kb. Whilst all four were observed in heart, brain, liver, skeletal muscle, kidney and pancreas, only the 1.6kb and 1.35kb bands were present in placenta, and the 1.35kb and the 4.5kb transcript in lung and liver. These transcripts may represent alternative splicing or differential polyA addition of a single gene, or perhaps transcripts from related genes. The presence of a related sequence is supported by Southern blot analysis and preliminary mapping results (see below).

Differential processing provides a mechanism to increase the protein coding capacity from a single RNA transcript. Alternative transcripts may show tissue specificity, as may be observed here. Alternatively, the transcripts may confer different substrate specificities to their products. This has been observed in the fibroblast growth factor receptors (FGFR) FGFR1 and FGFR2, where alternate splicing of either exon IIIb or exon IIIc confers different binding specificities to the three fibroblast growth factors (Johnson & Williams, 1993). Furthermore, alternatively spliced forms may carry out distinct protein functions. For example, in rats, the same primary transcript, when expressed in the thyroid gland encodes the calcium regulating hormone calcitonin, but in neurones in the pituitary gland encodes the calcitonin gene-related peptide which is thought to function in taste (Bovenberg et al, 1988). Thus characterization of the four transcripts and the gene(s) encoding them is required to investigate further the role of alternate splicing.

Preliminary mapping data obtained from a panel of human-rodent somatic cell hybrids supports the presence of two related genes. An intense PCR product was amplified from the chromosome 4-only hybrid whilst a low intensity product was amplified from the chromosome 7-only hybrid. Since these experiments were carried out using primers based on the 3' UTR sequence, it seems that the two genes are highly conserved. It is interesting that the homologues of these sequences were not amplified from the rodent parent DNA. Generally, coding regions of homologous genes in closely related species show greater conservation than paralogous genes within a species, as exemplified by human, rabbit and rat *PGM1* and human *PGMRP*. Therefore, this could signify that duplication of the sequences located on chromosomes 4 and 7 is a relatively recent event, perhaps confined to primates, and the human ESTI homologue in rodents is more highly diverged.

Thus, the human ESTI sequence was localized to chromosome 4 (and the related sequence to chromosome 7) using a local somatic cell hybrid panel. This finding is supported by the independent mapping of the ESTI-related clone, 130882, to chromosome 4. Therefore the sequence represented by these ESTs is a strong candidate for *PGM2*. Other lines of evidence support this possibility. The peptides encoded by the 5' sequences of both clones show the highest identity with two mycoplasma PMM proteins, encoded by *cpsg* in *M.genitalium* and by *pmm* in *M.pirum* (Fraser et al, 1995; Tham et al, 1993). These genes have been designated to encode PMM due to their homology with other prokaryotic phosphohexomutases. However, both genes are located adjacent to genes involved in the salvage pathway of nucleosides and therefore Tham and collegues suggested that the gene may represent a phosphodeoxyribomutase. Although the human PGM2 isozyme possess both phosphoglucomutase and phosphomannomutase activities, it shows greater phosphodeoxyribomutase and phosphoribomutase activity than PGM1. Hence it is thought to be the true phosphopentomutase, with PGM1 the true phosphoglucomutase (Quick et al, 1972). Therefore, the mycoplasma PMMs and human ESTI sequences are suggested to represent phosphopentomutases.

The localization of *PGM2*, based on somatic cell hybrids, is 4p14-4q12 (McAlpine et al, 1990) and the precise chromosomal localization of the human

ESTI sequence to this region would provide additional evidence that the sequence is *PGM2*. Proof may be obtained by expression studies of the full cDNA, in which PGM, PMM and phosphopentomutase activity could be demonstrated, and by immunoblot detection. Antibodies raised against the protein could be used to detect antigen following both starch gel electrophoresis and isoelectric focusing, allowing direct comparison of variant PGM2 isozymes with activity stained gels.

9.4 EVOLUTION OF THE *PGM1* GENE

9.4.1 *PGM1* - A HIGHLY CONSERVED GENE IN MAMMALS

PGM1 is highly polymorphic at the protein level, with the ten commonest phenotypes arising from four alleles: *PGM1**1, *PGM1**2, *PGM1**+ and *PGM1**-. At the nucleotide level, however, *PGM1* is highly conserved. Sequencing of the entire coding region of the gene in 11 unrelated individuals and 6 lymphoblastoid cell lines identified only two nucleotide substitutions which, together with intragenic recombination, were found to give rise to the four alleles (March et al, 1993a).

Sequencing of the *PGM1* cDNA in a further 27 individuals identified that the rare *PGM1**3 and *PGM1**7 alleles are due to a further single nucleotide substitution in exon 1A (Takahashi et al, 1993). Again, no additional mutations were observed other than those which give rise to the most common PGM1 protein alleles. *PGM1*, therefore, appears to be unusual in two aspects: first, intragenic recombination generates protein variation, and second, the nucleotide sequence appears to be depauperate in nucleotide polymorphisms. Are these features specific to *PGM1* in man?

9.4.1.1 *PGM1* in the Hominoidea

The nucleotide sequence of five of the eleven *PGM1* exons was investigated in chimpanzee (5), gorilla (2) and orangutan (5) as representatives of great apes in the hominoidea superfamily. No nucleotide substitutions leading to missense mutations were observed in these exons (1A, 4, 5, 8 and 11). All of the samples carried a C at nt 723 in exon 4 and T at nt 1320 in exon 8, which are characteristic of the PGM1*1+ protein phenotype observed in man. This provides support for the PGM1*1+ as the ancestral allele in man.

Only four nucleotide changes were observed between man and the great apes, all of which were synonymous: one in gorilla in exon 1A at nt 269, and three in orangutan, two in exon 4 at nt 647 and nt 707 and one in exon 5 at nt 815. The exon 4 change at nt 707 was polymorphic in orangutans. Thus, the amino acid sequence is completely conserved between these three species and man, and the nucleotide sequence appears to be resistant to nucleotide substitutions in the hominidae chimpanzee and gorilla.

Two PGM1 isozymes have been observed in chimpanzee and gorilla, one of which corresponds in electrophoretic mobility to the PGM1*1 isozyme of man following starch gel electrophoresis (Schmitt et al, 1970). In gorilla the second isozyme, PGM1*Go, has a slower rate of migration than PGM1*1, producing a more cathodal band. In chimpanzee the second isozyme, PGM1*Pan, has a slower rate of migration than PGM1*1, but is faster than the PGM1*Go isozyme. These variant isozymes have not been demonstrated by IEF (Carter et al, 1979), and therefore it is not possible to determine if the they represent the additional cathodal bands observed by IEF in gorilla (Daniel) and chimpanzee (Halfpenny) in this study. The cathodal bands could alternatively represent variants of the PGM1*1 isozyme, in a similar way to that of man subdividing into + and - forms on IEF.

9.4.1.2 *PGM1* in Lemur, Rabbit, Rat and Mouse

Evolution of the *PGM1* gene was also investigated by sequence analysis of exons 1A and 5 in the lemur. Lemurs belong to the suborder prosimii, which is estimated to have diverged from the suborder anthropodiea, of which man, chimpanzee, gorilla and orangutan belong to, between 56 and 65 million years ago. *PGM1* sequence from exons 1A and 5 in the rabbit and rat, and from exon 5 in the mouse was also available (Whitehouse et al, 1992; Auger et al, 1994; Friedman, personal communication). Comparative sequence analysis at both the nucleotide and amino acid level revealed an unexpected observation in exon 5.

Four nucleotide changes were identified in exon 5 of lemur *PGM1*, two of which are missense mutations. Although this data alone may not be unexpected, in rabbits, rat and mouse up to 13 nucleotide changes have occurred yet not one gives rise to a change in the amino acid sequence. The two missense mutations in the lemur at nt 900 and nt 911, affecting codons 279 and 282, occur at the 3' end of exon 5, 24bp and 13bp respectively, upstream of the sequence encoding the magnesium binding loop. The location of the amino acid substitutions were analyzed with respect to the 3-D structure of PGM1. The Glu^{279} to Gln^{279} substitution occurs in a loop region between an α-helix and a region of β-sheeting preceding the magnesium binding loop. In contrast, the Phe^{282} to Leu^{282} substitution occurs in the region of β-sheeting. Although these amino acids are not structurally similar, (Phe is aromatic whilst Leu is aliphatic) both are non-polar, and a single nucleotide change in third base of the codon would account for the substitution (TTY->TTR). However, it is still a little surprising that such an amino acid change has become established at a position involved in the secondary structure of the protein.

In contrast to exon 5, in exon 1A, only one of the five nucleotide changes gives rise to a missense mutation, and this is more consistent with the expected level of evolutionary divergence between lemur and man. Thus, *PGM1* may be subject to different evolutionary pressures along the length of the protein. It would be necessary to analyze the complete lemur PGM1 protein sequence to determine if the level of amino acid substitutions encoded by exon 5 are exceptional or characteristic of lemur PGM1, and determine the extent of the region in which this occurs. Differential selection pressures within a gene have previously been demonstrated. In the major histocompatibility complex loci I and II, the peptide-binding region (PBR) of the molecules show a far greater number of missense mutations than synonymous (Klein et al, 1993). In contrast, the non-PBR shows a significantly higher number of synonymous mutations than missense mutations. It is not certain why a similar observation should occur in lemur PGM1, however it may be associated with an alternative role for PGM1 in which these mutations may have a selective advantage.

9.4.2 PGM1 - IS IT A MULTIFUNCTIONAL PROTEIN?

The high level of amino acid sequence conservation suggests a great selection pressure is acting to maintain the mammalian PGM1 protein. An extreme case of amino acid conservation is found in the calmodulin gene, which encodes a member of a family of eukaryotic proteins controlling calcium levels within cells (Creighton, 1993). Human and chicken calmodulins have identical amino acid sequences and vary from eel by a single amino acid substitution. It is thought that this is due to the requirementof the protein to bind calcium, change shape and interact with other proteins. Thus the amino acid sequence is constrained to maintain a structure which can carry out these multiple functions. Recognition of PGM1 as an isozyme marker is dependent on its ability to catalyze the interconversion of Glc-1-P and Glc-6-P. However, recent investigations suggest additional roles for PGM1.

9.4.2.1 PGM1 as a Phosphoprotein in Sarcoplasmic Reticulum

PGM1 is found as a phosphoprotein associated with the membrane in rabbit skeletal muscle sarcoplasmic reticulum (SR) (Lee et al, 1992a). It is phosphorylated by a calcium/calmodulin-dependent protein kinase, and PGM1 possesses five putative sites. Investigations of the phosphotransferase activity of PGM1 showed it was significantly reduced when associated with the membrane. However, loss of activity was not irreversible. PGM activity was recovered following incubation with 1M guanidine HCl, suggesting hydrophobic interactions between PGM1 and phospholipids and/or proteins located in the SR. The authors proposed that if the interaction occurred with the N-terminal hydrophobic region, which includes the active site residue Ser^{116}, this would explain the loss of PGM activity in the membrane associated protein.

The phosphoprotein is thought to be involved in the regulation of calcium from the SR, since an increase in phosphorylation corresponded to a reduction in calcium release from the SR (Kim & Ikemoto, 1986). However, the precise role of PGM1 as a phosphoprotein in calcium regulation remains to be elucidated.

9.4.2.2 PGM1 as a Glucose-Phosphotransferase Acceptor Protein

Intracellular glycoproteins usually occur within the lumen of the endoplasmic reticulum or Golgi apparatus, sometimes transversing the membranes of these organelles. However, glycoproteins have more recently been found in the nucleus and cytoplasm, and are the result of novel glycosylation events (Hayes & Hart, 1994). One such event is the transfer by UDP-glucose:glycoprotein glucose-1-phosphotransferase (Glc-phosphotransferase) of glucose-1-phosphate (Glc-1-P) from UDP-glucose (UDP-Glc) to mannose residues on an acceptor glycoprotein (Koro & Marchase, 1982). The predominant acceptor was identified as a 62,000mw protein (Marchase et al, 1987).

The acceptor glycoproteins from both *S.cerevisiae* and rat were identified as PGM1. The glycoproteins were purified and partial amino acid sequencing revealed a high homology to rabbit PGM1 (Marchase et al, 1993; Auger et al, 1993). The rat *PGM1* cDNA was subsequently obtained and subcloned into a eukaryotic expression vector (Rivera et al, 1993). The transfected cells showed a significant increase in PGM activity, and also overproduction of a protein metabolically-labelled with glucose and mannose (Veyna et al, 1994). The glycosylation of rat PGM by Glc-1-P was at a site distinct from the Ser^{116} (Marchase et al, 1993).

Although the acceptor glycoprotein was cytosolic, when high calcium concentrations were maintained during fractionation studies, the glycoprotein associated with the microsomal pellet, suggesting a calcium dependent binding to the internal membranes (Srisomsap et al, 1988). Further investigations found that upon depolarization of the membrane, an influx of calcium ions was associated with an increase in glycosylation of PGM (Veyna et al, 1994). Thus, the function of PGM1 as a Glc-phosphotransferase acceptor may complement the observation by Lee et al, (1992a) that it is involved in calcium regulation. Although the mechanism of this regulation is unknown, it is evident that cytoplasmic glycosylation, as exemplified here, is important in regulating an enzyme's function in the same way as phosphorylation and methylation.

9.4.3 HIGH CONSERVATION IN OTHER SPECIES

Amino acid sequence comparisons of PGM1 from man, primates, rabbit and rodents indicate it is a highly conserved protein in mammals. Previous studies show that the anti-rabbit PGM polyclonal antibody identifies further species in which the amino acid sequence is highly conserved. Immunoblot detection carried out on *Xenopus* skeletal muscle and guppy extracts identified a distinct PGM1 homologue in both species that was identical in electrophoretic mobility to the PGM isozyme detected by enzyme activity staining (unpublished data). Thus the PGM protein in these species contain epitopes in common with those of mammalian PGM1. However, these epitopes are not conserved in *Alvinella pompejana*, a deep sea hydrothermal polychaete, nor in *A.tumefaciens* PGM, which is 56.0% identical to rabbit PGM1 at the amino acid level.

9.4.4 AN EXAMPLE OF TRANS-KINGDOM HORIZONTAL GENE TRANSFER?

The *A.tumefaciens* PGM is distinctive among prokaryotic phosphohexomutases due to the high level of sequence conservation shared with mammalian PGMs. The nucleotide sequence shows greater identity to human *PGM1* (61%) than either of the paralogous yeast genes (both 58%). Phylogenetic analysis carried out in Chapter Eight, using both parsimony and neighbour-joining methods, places the *A.tumefaciens* PGM protein closer to human PGM1, than the yeast proteins. Therefore, this may be an example of trans-kingdom horizontal gene transfer, from a eukaryote to *A.tumefaciens.*

A number of putative horizontal gene transfer events have been reported, with the transfer occuring in both directions; from eukaryotes to prokaryotes and prokaryotes to eukaryotes. In both cases, examples include species belonging to the Rhizobiaceae, a family of plant bacteria of which *A.tumefaciens* is a member. They are characterized by the ability to transfer DNA into the plant cell and therefore this may provide a possible mechanism for trans-kingdom horizontal gene transfer.

9.4.4.1 Eukaryotic to Prokaryotic Transfer

The most quoted example of horizontal gene transfer involves the acquisition of a second glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) gene by *E.coli* from a eukaryote. The first *GAPDH* gene reflects the expected prokaryotic ancestry. The second, however, shows approximately 60% identity to the eukaryotic protein sequences, but only 40% to those which are prokaryotic. Phylogenetic analysis places this second *GAPDH* in the eukaryotic branch of the tree, alongside single celled organisms, trypanosomes and yeast (Doolittle et al, 1990).

An example involving the Rhizobiaceae is glutamine synthetase II (GSII). The GSII from the symbiont *Bradyrhizobium japonicum* is distinct from other known prokaryotic glutamine synthetases, including its own *GSI* gene, with respect to its structure, immunoreactivity and sequence. Since the two alternative forms of glutamine synthetase have only been found in the Rhizobiaceae, and the protein sequence of this GSII was approximately 43% identical to plant GSII, it was suggested the bacterial gene was of a eukaryotic origin (Carlson & Chelm, 1986). Further phylogenetic studies, however, suggest that the divergence of the GSII sequences are comparable to the divergence of the species involved. The bacterial GSII showed a similar level of identity at the amino acid level with mammalian GSII as plant GSII (Shatters & Kahn, 1989). Thus, if transfer had occurred from a plant, the bacterial GSII would be expected to show a higher identity with the plant GSII.

Another possible example of eukaryote to prokaryote horizontal gene transfer involves the Rhizobiaceae species *Vitreoscilla*. This bacteria expresses a haemoglobin-like protein, which is 24% identical to lupin leghaemoglobin, and the protein aligns to show secondary structure conservation of the helical regions in several animal and plant globins (Wakabayashi et al, 1986). Since this was a unique example of a bacterial haemoglobin, it was proposed that the bacteria acquired the gene via horizontal gene transfer.

9.4.4.2 Prokaryotic to Eukaryotic Transfer

The most likely example of prokaryotic to eukaryotic transfer involves the Fe superoxide dismutase (Fe-SOD) of *Entamoebae histolytica*. This sequence is approximately 60% identical to the prokaryotic Fe-SODs, but only 38% to the SOD sequences of other eukaryotes (Smith et al, 1992). The fact that *E.histolytica* engulfs bacteria may provide the opportunity for horizontal gene transfer to occur.

A more recently published example suggests the transfer of the *Agrobacterium rhizogenes rolC* gene into *Nicotiana tabacum* (Meyer et al, 1995). This gene is sited in *A.rhizogenes* on the Ri-plasmid, and is therefore transferred into the plant cell. The tobacco homologue shows between 54% and 89% identity at the amino acid level with the *rolC* genes from the different strains of Ri-plasmids. Although regions of homology with *rolC* have been identified in plants by Southern blot analysis, these were thought to be due to relics of an ancient infection. This is the first expressed homologue of a bacterial gene to be identified. As with the bacterial haemoglobin gene, the uniqueness of the sequence is the primary evidence for genetic transfer.

9.4.4.3 Unequal Rates of Change and Convergent Evolution

In addition to horizontal gene transfer, two other possiblities may account for an unconventional phylogeny. The first possibility is that the sequences show unequal rates of amino acid changes along different lineages. This is exemplified by the calmodulin-like gene in the chicken. Unlike all the previously cloned vertebrate calmodulin genes, which showed extremely high conservation, the sequence was highly diverged. It was proposed that the chicken had acquired the gene by horizontal gene transfer, posssibly involving viral-mediated retrotransposition, since the gene was intron-less (Gruskin et al, 1987). However, a human homologue was identified and both this and the chicken calmodulin-like genes showed a higher rate of evolutionary change, compared to the other vertebrate calmodulins (Syvanen, 1994). Thus, the case for horizontal transfer was disproved.

The second possibility is convergent evolution. However, although functional, mechanistic and structural convergence have been demonstrated, sequence convergence has not (Doolittle, 1994). Convergence is defined as adaptive changes occurring in two proteins such that they appear more related than they are. Functional convergence is demonstrated by many enzymes, including the superoxide dismutases (section 1.1.3), aldolases and sugar kinases. Mechanistic convergence is demonstrated by the catalytic triad of His, Asp and Ser. This is found in two distinct lineages of the serine proteases, the eukaryotic chymotrypsin and the bacterial subtilisin, despite having completely different three dimensional protein structures.

Apparent convergence of secondary structure, involving features such as β-barrels, is found in numerous diverse proteins, including fibronectin type III and immunoglobulin domains (Doolittle, 1994). Although this may represent descent from a common ancestor, it seems more likely that there is a general convergence due to the ease of formation and intrinsic stability of this structure. In all of the examples of structural convergence, the sequences remain distinct and convergence has not been demonstrated. During multiple sequence alignments of PGM reported in both Chapter Five and Chapter Eight, in addition to the conserved protein motifs, key amino acids, such as glycines and prolines which are important for secondary structure formation, were also found to be highly conserved. Therefore, this suggests common ancestry rather than chance occurrence of these amino acids at these locations.

9.4.4.4 *A.tumefaciens* and PGM: An example of trans-kingdom horizontal gene transfer?

The phylogenetic placing of *A.tumefaciens* PGM with the eukaryotic sequences is not thought to be due to unequal rates of change, since this would be expected to promote divergence, not homology. The possibility of sequence convergence is unlikely, due to the exceptionally high level of sequence conservation observed with mammalian PGM1. Furthermore, in cases of sequence convergence, it would be expected for amino acid identity to be greater than nucleotide identity, but this is not the case: the amino acid identity is 56% and the nucleotide identity is 61%, between human PGM1 and *A.tumefaciens* PGM. This reflects nucleotide changes in the first and second positions of the codon, rather than the third.

The *A.tumefaciens* PGM shows only 26% identity with the true bacterial PGMs, *Acetobacter xylinum* PGM and *E.coli* PGM, as identified by the phylogeny presented in Chapter Eight. This may also be supporting evidence for horizontal gene transfer, since this is far less than the GAPDH and Fe-SOD examples, in which the candidate for horizontal gene transfer showed approximately 40% identity with its expected homologues. Finally, although six

PGM-like genes have been identified from prokaryotes which appear to have diverged along distinct evolutionary pathways, it seems unlikely that a seventh pathway would represent evolution of the *A.tumefaciens PGM*. Extensive studies have been carried out in a wide range of bacteria, and this is the sole example of a highly conserved eukaryotic-like gene.

In conclusion, it would appear that the *A.tumefaciens PGM* is an example of horizontal gene transfer. Its position in the eukaryotic branch is supported by both parsimony and neighbour-joining methods of phylogenetic analysis, with the topology of the tree reflecting the expected evolution of the *PGM1* homologues and paralogues from man, protist, yeast and prokaryotes. Since the result cannot be explained by either unequal rates of evolution or convergent evolution, horizontal gene transfer appears to be the simplest explanation.

9.5 EVOLUTION OF THE PHOSPHOHEXOMUTASES

The phylogenetic analysis presented in Chapter Eight was carried out to identify alternative pathways in the evolution of the ancestral gene. Of the eight distinct branches found, two were thought to represent the evolution of the PGM2 and PGM3 isozymes. The first branch, of which the sole representative is the *M.pirum* PMM protein, was suspected to depict the evolution of PGM2. This theory has subsequently received support from work with the human ESTI sequence. The peptides encoded by the 5' sequence of both human ESTI and clone 130882 (which includes a putative magnesium binding loop) show greater identity with *M.pirum* PMM than any other sequence in the database, and both human ESTI and clone 130882 have independently been mapped to chromosome 4 (section 9.3.2). The second alternative pathway identified by the phylogenetic analysis, which was thought to depict the evolution of PGM3, was represented by the yeast AGM protein.

9.5.1 IS PGM3 THE HUMAN HOMOLOGUE OF YEAST AGM?

The *S.cerevisiae* N-acetylglucosamine phosphate mutase (AGM) protein is a member of a family of hexosephosphate mutases which exhibit overlapping substrate specificites; AGM, PGM1, PGM2 and PMM (encoded by *sec53*) all possess phosphoglucomutase activity (Boles et al, 1994; Hofmann et al, 1994). Since the AGM protein contains the conserved active site motif TASHNP, and a magnesium binding loop motif FDGDADR, it was thought that if there was a homologous gene in humans, it may encode a PGM isozyme. AGM exhibits only a low level of PGM activity and therefore, it was thought that the human homologue may be PGM3. This was supported by molecular weight data: the yeast AGM is 62,000mw which is comparable with the estimate of 65,000mw for PGM3.

Degenerate primer PCR was carried out in an attempt to identify the human homologue, but no candidate sequences were obtained. As discussed earlier (section 9.2.3), this probably reflects the design of the primers, which were based on peptides subsequently found to be less conserved than anticipated.

AGM activity corresponding to the PGM isozymes was investigated by activity staining following starch gel electrophoresis (Marenah, 1973). No AGM activity was observed. AGM had previously been demonstrated as an activity distinct from PGM (Carlson,1966) and has been purified from pig submaxillary gland (Fernandez-Sorenson & Carlson, 1971). The protein was found to require magnesium ions and a biphosphate cofactor, as PGM does. Studies of the mechanism of revealed it resembled the two-step reaction of phosphoglucomutase (Cheng & Carlson, 1979):

(i)
GlcNAc-1-P + phosphoenzyme = GlcNAc-1,6-P + dephosphoenzyme
(ii)
GlcNAc-1,6-P + dephosphoenzyme = GlcNAc-6-P + phosphoenzyme

Thus, although the biochemical data from mammals may be explained by functional convergence between AGM and PGM1, in combination with the genetic data from yeast, the simplist explanation is that the human *AGM* homologue is a highly diverged PGM-related gene.

9.5.2 EVIDENCE FOR A PGM-RELATED GENE FAMILY IN MAN

Recent evidence supports the existence of a PGM superfamily in man, with the genes encoding a diverse range of functions. In addition to *PGM1*, members include *PGMRP*, which encodes a structural protein located in adherens-type cellular junctions (Moiseeva et al, 1996), and human ESTI, a candidate for *PGM2*, which may represent a phosphopentomutase. A further member may be the human homologue of *PFUS*, which encodes the *Paramecium tetraurelia* protein parafusin (Wyroba et al, 1995).

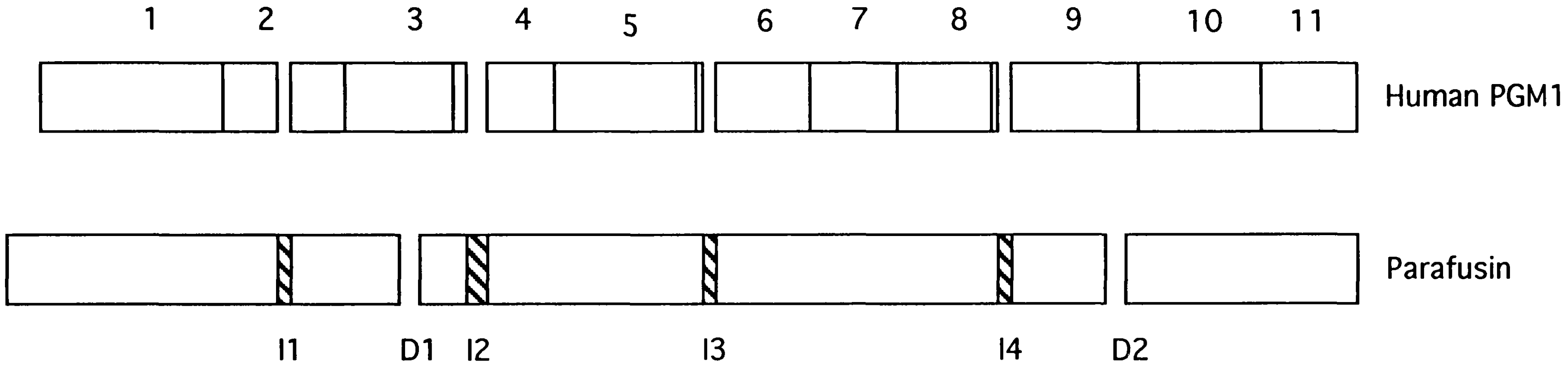
Parafusin is a glucose phosphotransferase acceptor protein, which undergoes rapid dephosphorylation upon stimulation of secretion, and is therefore thought to be involved in the regulation of exocytosis (Satir et al, 1990). However, there was contradictory evidence as to whether this protein is paramecium PGM, which has evolved to serve dual roles, (gene sharing), or whether it is distinct protein. Evidence for gene sharing is suggested by PGM and parafusin being co-eluted by chromatography, sharing the same molecular mass, being phosphorylated by same two protein kinases, sharing the same pI values of isoforms and showing cross reactivity of polyclonal antibodies raised against rabbit PGM to PGM in paramecium (Treptau et al, 1995).

However, the cloning of *PFUS* identified 4 insertions and 2 deletions compared to human *PGM1* (Figure 9.5). The inserted and deleted regions comprise five to ten amino acids and none of them correlate with intron/exon boundaries found in human *PGM1*, thereby ruling out the possiblity of alternate splicing of the parafusin *PGM* gene. Southern blot analysis using a probe designed to insertion-3 showed bands distinct from those obtained using a PGM specific probe designed from deletion-2 (Subramanian et al, 1994). Further, a specific peptide antibody generated from the insertion-4 region recognized parafusin, but did not show immunoreactivity with either commercially available rabbit PGM, nor paramecium PGM (Andersen et al, 1994). Finally, the paramecium PGM enriched fractions were shown to be enzymatically active, whilst those for parafusin were not.

The presence of a human homologue to parafusin was demonstrated in human pancreas extracts by immunoblot analysis using parafusin specific antibodies (Wyroba et al, 1995). These were raised against insertion 4 and insertion 2 of the parafusin protein and showed immunoreactivity with a 63,000mw protein, but not purified rabbit muscle PGM. PGM-specific antibodies, raised against deletion 2, were shown to detect a band of slightly higher molecular weight in the extracts. Southern blot analysis using probes specific to *PFUS* and *PGM1*, based on insertion 3 and deletion 2 respectively, showed different and distinct patterns of hybridization in rat DNA. Thus, in both human and rat, and therefore most mammals, *PFUS* and *PGM1* are distinct entities.

The human parafusin homologue is unlikely to represent one of the PGM isozymes. PGM2 and PGM3 isozymes have a greater molecular size than PGM1, whereas the parafusin homologue is smaller. In addition, parafusin does not show PGM activity and it is therefore expected for the mammalian homologue to also be enzymatically inactive.

Figure 9.5 Comparison of human PGM1 and *P.tetraurelia* parafusin proteins to illustrate the location of the insertions (I1 - I4) and deletions (D1 - D2) in parafusin. The location corresponding to the intron/exon boundaries are illustrated on human PGM1.



9.5.3 EVIDENCE FOR CONVERGENT EVOLUTION OF PMM IN MAN?

The PGM1, PGM2 and PGM3 isozymes are generally thought to be members of an diverged gene family, and this view will be supported if the human ESTI sequence is found to be PGM2. However, the possibility remains that the isozymes may also be the result of convergent evolution.

An alternative structural framework which catalyzes the PGM/PMM reaction was found in yeasts, with the cloning of the *sec53* from *Saccharomyces cerevisiae* and *pmm* from *Candida albicans* (Bernstein et al, 1985; Smith et al, 1992). The 29,000mw protein shows none of the characteristic motifs found in the other phosphohexomutases. A human EST clone was identifed which was homologous to *sec53*. However, preliminary RT-PCR experiments failed to amplify the sequence from RNA (section 6.3). The absence of an open reading frame in the 5' nucleotide sequence questioned both the reliability of the sequence data, and its status as an expressed sequence. Therefore, the presence of a *sec53* homologue in man remained a possibility. Southern blot analysis of human genomic DNA, using the *sec53* gene as probe, showed faint hybridization signals at low stringency, suggesting that a homologous sequence was present in the genome (data not presented).

There is recent evidence of a PMM in man which is distinct from PGM (Van Schaftingen & Jaeken, 1995). Carbohydrate-deficient glycoprotein (CDG) syndromes are multisystemic genetic disorders characterized by defective N-glycosylation of serum and cellular proteins. Deficiency of PMM activity was found in 6 patients with CDG syndrome type I, whilst PGM activity, as well as other enzymes involved in the conversion of glucose to mannose-1-phosphate, were normal. PMM activity was also normal in patients with CDGII, which is due to a deficiency in N-acetylglucosaminyltransferase II. Thus, the deficiency of PMM appears to be the cause of CDGI. The gene for CDGI has been mapped by linkage studies to chromosome 16p13.3-p13.12 (Martinsson et al, 1994). Therefore, the PMM which is deficient in these patients is likely to be distinct from the PGM2 and PGM3 isozymes, which map to chromosomes 4 and 6 respectively. However, this gene may represent the human *sec53* homologue.

9.5.4 FINAL CONCLUSIONS

The molecular investigations of the PGM gene family reported in this thesis have shown that *PGM2* and *PGM3* are not closely related to *PGM1*. Of the strategies investigated to identify the genes for *PGM2* and *PGM3*, searching the EST databases was the most efficient and successful. The data suggests the partial cDNA represented by the human ESTI clone is *PGM2*, although further characterization is required for confirmation. If it does encode *PGM2*, it will support the long standing belief that the three isozymes are members of an ancient gene family. In addition, it indicates the usefulness of phylogeny construction to investigate alternative pathways of evolution. The human ESTI clone is most similar to PMM of *Mycoplasma pirum*, which was identified as a probable evolutionary pathway for PGM2. Another alternative pathway identified, which I believe depicts the evolution of the PGM3 isozyme, is that represented by AGM in yeast. Further studies to pursue the molecular characterization of *PGM3* should progress in this direction, by perhaps using the full length yeast *AGM* genes to search the EST databases.

Molecular characterization of the gene represented by human ESTI will allow comparative studies with *PGM1* to be carried out. *PGM1* shows remarkable nucleotide conservation, with the nucleotide substitutions which underlie the protein alleles the only changes seen so far between individuals. It will be interesting to determine if the nucleotide sequence is as highly conserved in the human ESTI gene, especially if it is *PGM2*, since PGM2 variant protein alleles are rarely found. Amino acid sequence comparisons and molecular modelling will allow a prediction of the protein's structure. It will be interesting to see how it compares with PGM1 and what features underlie the change in specificity. If the gene is *PGM2*, the comparison will identify where the extra amino acids and/or domains lie, since PGM2 is an estimated 10,000mw larger than PGM1.

Appendix A Multiple sequence alignment file of the 28 PGM, PMM and PGM-related sequences included in the phylogenetic analysis.

```
                  GapWeight: 3.000
            GapLengthWeight: 0.100

Name: l24077           Len: 848  Check: 3502  Weight: 1.00
Name: u08369           Len: 848  Check: 9385  Weight: 1.00
Name: x72016           Len: 848  Check: 4575  Weight: 1.00
Name: x74823           Len: 848  Check: 5193  Weight: 1.00
Name: pgmrp            Len: 848  Check: 5294  Weight: 1.00
Name: m83088           Len: 848  Check: 9503  Weight: 1.00
Name: l24117           Len: 848  Check: 4510  Weight: 1.00
Name: pfus             Len: 848  Check: 2376  Weight: 1.00
Name: x57132           Len: 848  Check: 4595  Weight: 1.00
Name: l12968           Len: 848  Check: 9119  Weight: 1.00
Name: l27632           Len: 848  Check: 1812  Weight: 1.00
Name: l27646           Len: 848  Check: 6699  Weight: 1.00
Name: d13231           Len: 848  Check: 6983  Weight: 1.00
Name: x59886           Len: 848  Check: 7562  Weight: 1.00
Name: m84642           Len: 848  Check: 6129  Weight: 1.00
Name: m77127           Len: 848  Check: 9989  Weight: 1.00
Name: l11721           Len: 848  Check: 9989  Weight: 1.00
Name: l04596           Len: 848  Check: 6233  Weight: 1.00
Name: m83231           Len: 848  Check: 3438  Weight: 1.00
Name: u02489           Len: 848  Check: 6324  Weight: 1.00
Name: u02490           Len: 848  Check: 6941  Weight: 1.00
Name: m60873           Len: 848  Check: 2258  Weight: 1.00
Name: x79075           Len: 848  Check: 7623  Weight: 1.00
Name: u20583           Len: 848  Check:  923  Weight: 1.00
Name: x75898           Len: 848  Check: 4978  Weight: 1.00
Name: l13289           Len: 848  Check: 2513  Weight: 1.00
Name: x56793           Len: 848  Check: 8570  Weight: 1.00
Name: x75816           Len: 848  Check: 5924  Weight: 1.00

        1                                                   50
l24077  .......... .......... .......... .......... ..........
u08369  .......... .......... .......... .......... ..........
x72016  .......... .......... .......... .......... ..........
x74823  .......... .......... .......... .......... ..........
 pgmrp  .......... .......... .......... .......... ..........
m83088  .......... .......... .......... .......... ..........
l24117  .......... .......... .......... .......... ..........
  pfus  .......... .......... .......... .......... ..........
x57132  .......... .......... .......... .......... ..........
l12968  .......... .......... .......... .......... ..........
l27632  .......... .......... .......... .......... ..........
l27646  .......... .......... .......... .......... ..........
d13231  .......... .......... .......... .......... ..........
x59886  .......... .......... .......... .......... ..........
m84642  .......... .......... .......... .......... ..........
m77127  .......... .......... .......... .......... ..........
l11721  .......... .......... .......... .......... ..........
l04596  .......... .......... .......... .......... ..........
m83231  .......... .......... .......... .......... ..........
u02489  .......... .......... .......... .......... ..........
u02490  .......... .......... .......... .......... ..........
m60873  .......... .......... .......... .......... ..........
x79075  .......... .......... .......... .......... ..........
u20583  LLPQGHAGPG AHRPAGWQPV RIIETVVDIN DKRKKQMGGG IVAACGIGVR
x75898  .......... .......... .......... .......... ..........
l13289  .......... .......... .......... .......... ..........
x56793  .......... .......... .......... .......... ..........
x75816  .......... .......... .......... .......... ..........
```

```
          51                                                 100
 124077   .......... .......... .......... .......... ..........
 u08369   .......... .......... .......... .......... ..........
 x72016   .......... .......... .......... .......... ..........
 x74823   .......... .......... .......... .......... ..........
  pgmrp   .......... .......... .......... .......... ..........
 m83088   .......... .......... .......... .......... ..........
 124117   .......... .......... .......... .......... ..........
   pfus   .......... .......... .......... .......... ..........
 x57132   .......... .......... .......... .......... ..........
 112968   .......... .......... .......... .......... ..........
 127632   .......... .......... .......... .......... ..........
 127646   .......... .......... .......... .......... ..........
 d13231   .......... .......... .......... .......... ..........
 x59886   .......... .......... .......... .......... ..........
 m84642   .......... .......... .......... .......... ..........
 m77127   .......... .......... .......... .......... ..........
 111721   .......... .......... .......... .......... ..........
 104596   .......... .......... .......... .......... ..........
 m83231   .......... .......... .......... .......... ..........
 u02489   .......... .......... .......... .......... ..........
 u02490   .......... .......... .......... .......... ..........
 m60873   .......... .......... .......... .......... ..........
 x79075   .......... .......... .......... .......... ..........
 u20583   QDRRRPGRHL QTQHRRHARR PVSRHRPGAP GRRRHRAGLR PRRHARGREA
 x75898   .......... .......... .......... .......... ..........
 113289   .......... .......... .......... .......... ..........
 x56793   .......... .......... .......... .......... ..........
 x75816   .......... .......... .......... .......... ..........

          101                                                150
 124077   .......... .......... .......... .......... ..........
 u08369   .......... .......... .......... .......... ..........
 x72016   .......... .......... .......... .......... ..........
 x74823   .......... .......... .......... .......... ..........
  pgmrp   .......... .......... .......... .......... ..........
 m83088   .......... .......... .......... .......... ..........
 124117   .......... .......... .......... .......... ..........
   pfus   .......... .......... .......... .......... ..........
 x57132   .......... .......... .......... .......... ..........
 112968   .......... .......... .......... .......... ..........
 127632   .......... .......... .......... .......... ..........
 127646   .......... .......... .......... .......... ..........
 d13231   .......... .......... .......... .......... ..........
 x59886   .......... .......... .......... .......... ..........
 m84642   .......... .......... .......... .......... ..........
 m77127   .......... .......... .......... .......... ..........
 111721   .......... .......... .......... .......... ..........
 104596   .......... .......... .......... .......... ..........
 m83231   .......... .......... .......... .......... ..........
 u02489   .......... .......... .......... .......... ..........
 u02490   .......... .......... .......... .......... ..........
 m60873   .......... .......... .......... .......... ..........
 x79075   .......... .......... .......... .......... ..........
 u20583   AARRGLGEGR LRPLDGADCV AILTEWNEFR ALDLKRVKTL LKEPVMIDLR
 x75898   .......... .......... .......... .......... ..........
 113289   .......... .......... .......... .......... ..........
 x56793   .......... .......... .......... .......... ..........
 x75816   .......... .......... .......... .......... ..........
```

```
        151                                                200
l24077  .......... .......... .......... .......... ..MPSISPFA
u08369  .......... .......... .......... .......... ...MAIHNRA
x72016  .......... .......... .......... .......... ..........
x74823  .......... .......... .......... .......... ..........
 pgmrp  .......... .......... .......... .......... ..........
m83088  .......... .......... .......... .......... ..........
l24117  .......... .......... .......... .......... ..........
  pfus  .......... .......... .......... .......... ..........
x57132  .......... .......... .......... .......... .......MKI
l12968  .......... .......... .......... .......... ....MSNRKY
l27632  .......... .......... .......... .......... ....MTQLTC
l27646  .......... .......... .......... .......... ....MTQLTC
d13231  .......... .......... .......... ..MVVANFFG TKRRMTQLTC
x59886  .......... .......... .......... .......... ....MTKLTC
m84642  .......... .......... .......... .......... ....MNNLTC
m77127  .......... .......... .......... .......... ....MKKLTC
l11721  .......... .......... .......... .......... ....MKKLTC
l04596  .......... .......... .......... .......... ......MLTC
m83231  .......... .......... .......... .......... .....MTLPA
u02489  .......... .......... .......... .......... ..MASITRDI
u02490  .......... .......... .......... .......... ..MASIARDI
m60873  .......... .......... .......... .......MST VKAPTLPASI
x79075  .......... .......... .......... .......... .....VPATL
u20583  NVYNPGDMAA AVRLLLHRTA HRTVPERLKR SSPRRTVETM SEAHTFHPTV
x75898  .......... .......... .......... ....MPAIFV RSSSSSSSTF
l13289  .......... .......... .......... ...MNNEIVK KWLSSDNVPQ
x56793  .......... .......... .......... .......... ..........
x75816  .......... .......... .......... .......... ..........

        201                                                250
l24077  GKPVDPDRLV NIDALLDAYY TRKPDPAIAT QRVAFGTSGH RGS....SLT
u08369  GQPAQQSDLI NVAQLTAQYY VLKPEAGNAE HAVKFGTSGH RGS....AAR
x72016  .......... .......MSL LIDSVPTVAY KDQKPGTSGL RKKTKVFMDE
x74823  .......... .......MSF QIETVPTKPY EDQKPGTSGL RKKTKVFKDE
 pgmrp  .......... .......... .......... .......... ..........
m83088  .......... ........MV KIVTVKTQAY QDQKPGTSGL RKRVKVFQSS
l24117  .......... .......... MIKTIKTTPY QDQKPGTSGL RKKVPVF.AQ
  pfus  .MVVLFLLPL RLGHNLWRIE APRVQVTQPY AGQKPGTSGL RKKVSE.ATQ
x57132  FGTDGV.... .......... ......RGKA G.V.....KL TPMFV.MRLG
l12968  FGTDGI.... .......... ......RGRV GDA.....PI TPDFV.LKLG
l27632  FKAYDI.... .......... ......RGEL .GE.....EL NEDIA.YRIG
l27646  FKAYDI.... .......... ......RGEL .GE.....EL NEDIA.YRIG
d13231  FKAYDI.... .......... ......RGEL .GE.....EL NEDIA.YRIG
x59886  FKAYDI.... .......... ......RGRL .GE.....EL NEDIA.WRIG
m84642  FKAYDI.... .......... ......RGRL .GE.....EL NEDIA.WRIG
m77127  FKAYDI.... .......... ......RGKL .GE.....EL NEDIA.WRIG
l11721  FKAYDI.... .......... ......RGKL .GE.....EL NEDIA.WRIG
l04596  FKAYDI.... .......... ......RGKL .GE.....EL NEDIA.WRIG
m83231  FKAYDI.... .......... ......RGRV .PD.....EL NEDLA.RRIG
u02489  FKAYDI.... .......... ......RG.I VGK.....TL TDDAA.YFIG
u02490  FKAYDI.... .......... ......RG.I VGK.....TL TDEAA.YLIG
m60873  FRAYDI.... .......... ......R.RV VGD.....TL TAETA.YWIG
x79075  FRAYDI.... .......... ......RGPV TSE.....AL TPGLA.YAVG
u20583  LREYDI.... .......... ......RG.I VGS.....TL TAADA.RAVG
x75898  ISTTDVFTND DDIERIKRLQ NGSDVSRVAL EGEKGREVDL TPPAV.EAIA
l13289  TDKDIISKMK NEELELAFSN APLSFGTAGI RAKMAPGTQF LNKITYYQMA
x56793  .......... .......... ...MNVVNNS RDVIYSSGIV FGTSGARGLV
x75816  .......... .......... .......... ......MKVD YEQLCKLYDD
```

```
         251                                                300
124077   TSFNENHILS ISQAIADYRK GAGITGPLFI GIDTHALSRP ALKSALEVFA
u08369   HSFNEPHILA IAQAIAEERA KNGITGPCYV GKDTHALSEP AFISVLEVLA
x72016   PHYTENFIQA TMQSI....P NGSEGTTLVV GGDGRFYNDV IMNKIAAVGA
x74823   PNYTENFIQS IMEAI....P EGSKGATLVV GGDGRYYNDV ILHKIAAIGA
 pgmrp   .......... .......... .......MVV GSDGRYFSRT AIEIVVQMAA
m83088   ANYAENFIQS IISTV...EP AQRQEATLVV GGDGRFYMKE AIQLIARIAA
124117   ENYAENFIQS IFDAL...EG FEGQ..TLVI GGDGRYYNRE VIQKAIKMAA
  pfus   PNYLENFVQS IFNTL...RK .DELKNVLFV GGDGRYFNRQ AIFSIIRLAY
x57132   IAAGLYF... .........K KHSQTNKILI GKDTRKSGYM VENALVSALT
112968   WAAGKVL... .........A RHG.SRKIII GKDTRISGYM LESALEAGLA
127632   RAYGEFL... .......... KPG...KIVV GGDVRLTSES LKLALARGLM
127646   RAYGEFL... .......... KPG...KIVV GGDVRLTSES LKLALARGLM
d13231   RAYGEFL... .......... KPG...KIVV GGDVRLTSES LKLALARGLM
x59886   RAYGEYL... .......... KPK...TVVL GGDVRLTSEA LNVALAKGLQ
m84642   RAYGEYL... .......... KPK...TIVL GGDVRLTSEA LKLALAKGLQ
m77127   RAYGEFL... .......... KPK...TIVL GGDVRLTSET LKLALAKGLQ
111721   RAYGEFL... .......... KPK...TIVL GGDVRLTSET LKLALAKGLQ
104596   RAYGEFL... .......... KPK...TIVL GGDVRLTSET LKLALAKGLQ
m83231   VALAAQL... .......... DQG...PVVL GHDVRLASPA LQEALSAGLR
u02489   RAIAAKA... .......... AEKGIARIAL GRDGRLSGPE LMEHIQRGLT
u02490   KAIAAKA... .......... AEKGITRIAL GRDGRLSGPE LMEHIRRGFT
m60873   RAIGSES... .......... LARGEPCVAV GRDGRLSGPE LVKQLIQGLV
x79075   LSIGSEA... .......... REQGQKAIVV GRDGRLSGPK LTAALIQGLC
u20583   RLRHRGR... .......... PAAVRKTVCV GYDGRLSSPE LEAAMVDGLV
x75898   ESFGEWLIAK LRDDDDYKEK QGVDVVKVSL GKDPRVTGAK LSVAVFSGLA
113289   TGYGKFLKNK F......... .SNQNISVIV AHDNRNNGID FSIDVTNILT
x56793   KDFTPQVCAA FTVSFVAVMQ EHFSFDTVAL AIDNRPSSYG MAQACAAALA
x75816   MCRTKNVQFS YGTAGFRTLA KNLDTVMFST GILAVLRSLK LQGQYVGVMI

         301                                                350
124077   ANGVEVRIDA QDGYTPTPVI SHAILTYNRD RSSDLADGVV ITPSHNPP..
u08369   ANGVDVIVQE NNGFTPTPAV SNAILVHNK. KGGPLADGIV ITPSHNPP..
x72016   ANGVRKLVIG QGGLLSTPAA SHIIRTYE.E KC..TGGGII LTASHNPGGP
x74823   ANGIKKLVIG QHGLLSTPAA SHIMRTYE.E KC..T.GGII LTASHNPGGP
 pgmrp   ANGIGRLIIG QNGILSTPAV SCIIRKI... ...KAAGGII LTASHCPGGP
m83088   ANGIGRLVIG QNGILSTPAV SCIIRKI... ...KAIGGII LTASHNPGGP
124117   AAGFGKVLVG QGGILSTPAA SNVIRKY... ...KAFGGIV LSASHNPGGP
  pfus   ANDISEVHVG QAGLMSTPAS SHYIRKVN.E EVGNCIGGII LTASHNPGGK
x57132   SIGYNVI... QIGPMPTPAI AFLTE..DMR ....CDAGIM ISASHNPFED
112968   AAGLSAL... FTGPMPTPAV AYLTR..TFR ....AEAGIV ISASHNPFYD
127632   DAGTDVL... DIGLSGTEEI YFATF..HLG ....VDGGIE VTASHNPMNY
127646   DAGTDVL... DIGLSGTEEI YFATF..HLG ....VDGGIE VTASHNPMNY
d13231   DAGTDVL... DIGLSGTEEI YFATF..HLG ....VDGGIE VTASHNPMNY
x59886   DAGVDVL... DIGMSGTEEI YFATF..HLG ....VDGGIE VTASHNPMDY
m84642   DAGVDVL... DIGMSGTEEI YFATF..HLG ....VDGGIE VTASHNPMDY
m77127   DAGVDVL... DIGMSGTEEI YFATF..HLG ....VDGGIE VTASHNPMDY
111721   DAGVDVL... DIGMSGTEEI YFATF..HLG ....VDGGIE VTASHNPMDY
104596   DAGVDVL... DIGMSGTEEI YFATF..HLG ....VDGGIE VTASHNPMDY
m83231   ASGREVI... DIGLCGTEEV YFQTD..HLK ....AAGGVM VTASHNPMDY
u02489   DSGISVL... NVGMVTTPML YFAAV..NEC ....GGSGVM ITGSHNPPDY
u02490   DSGINVL... NVGMVATPML YFAAV..NEC ....GGSGVM ITGSHNPPDY
m60873   DCGCQVS... DVGMVPTPVL YYAAN..VLE ....GKSGVM LTGSHNPPDY
x79075   ETGLAVL... NVGLVPTPLV YFATN..RLE ....TNSGVM VTASHNPGHH
u20583   ACGLHVL... RIGLGPTPML YFATR..DRE ....AAAGIM ITGSHNPPDY
x75898   RAGCLAF... DMGLATTPAC FMSTVFPHFS ....YHGSIM MTASHLPYTR
113289   SLELEFICLK IINLLLRQLF SYAI..RKLN ....AQGAVI VTASHNPKED
x56793   DKGVNCIF.. .YGVVPTPAL AFQSMSDNM. ......PAIM VTGSHIPFER
x75816   TASHNPYQDN GVKIVEPDGS MLLATWEPYA MQLANAASFA TNFEEFRVEL
```

```
        351                                               400
124077  ..EDGGYKYN PPHGGPADTD ITKVVETAAN DY........ ...MAKKMEG
u08369  ..EDGGIKYN PPNGGPADTN VTKVVEDRAN AL........ ...LADGLKG
x72016  E.NDLGIKYN LPNGGPAPES VTNAIWEASK KLTHYKIIK. ......NFPK
x74823  E.NDMGIKYN LSNGGPAPES VTNAIWEISK KLTSYKIIK. ......DFPE
 pgmrp  G.GEFGVKFN VANGGPAPDV VSDKIYQISK TIEEYAICP. ..DLRIDLSR
m83088  N.GDFGIKFN ISNGGPAPEA ITDKIFQISK TIEEYAVCP. ..DLKVDLGV
124117  T.EDFGIKYN IGNGGPAPEK ITDAIYARSK VIDSYKISD. ..AADIDLDK
  pfus  EHGDFGIKFN VRTGAPAPED FTDQIYTHTT KIKEYLTVDY EFEKHINLDQ
x57132  N....GIKFF NSYGYKLKEE EE.RAIEE.. .......... ....IFH..D
112968  N....GIKFF SIDGTKLPDA VE.EAIEA.. .......... ....EME..K
127632  N....GMKLV RENAKPISGD TGLRDIQR.. .......... ....LAE..E
127646  N....GMKLV RENAKPISGD TGLRDIQR.. .......... ....LAE..E
d13231  N....GMKLV RENAKPISGD TGLRDIQR.. .......... ....LAE..E
x59886  N....GMKLV REGARPISGD TGLRDVQR.. .......... ....LAE..A
m84642  N....GMKLV REGARPISGD TGLRDVQR.. .......... ....LAE..A
m77127  N....GMKLV REGARPISGD TGLRDVQR.. .......... ....LAE..A
111721  N....GMKLV REGARPISGD TGLRDVQR.. .......... ....LAE..A
104596  N....GMKLV REGARPISGD TGLRDVQR.. .......... ....LAE..A
m83231  N....GMKLV REQARPISSD TGLFAIRD.. .......... ....TVA..A
u02489  N....GFKMM LGGDTLAGEA ..IQELLA.. .......... ....IVE..K
u02490  N....GFKMM LGGDTLAGEA ..IQELLS.. .......... ....IIE..K
m60873  N....GFKIV VAGETLANEQ ..IQALRE.. .......... ....RIE..K
x79075  N....GFKIV LNGKTLRSEE ..IATIRT.. .......... ....RIL..E
u20583  N....GIKMM LGKGPVYGRQ ..ILDIGA.. .......... ....IAS..K
x75898  N....GLKFF TRRGGLTS.. ...LEVEE.. .......... ....ICDRAA
113289  N....GFKIY NETGAQVLPD DGLKVVEL.. .......... ....MPNVF.
x56793  N....GLKFY RPDG...... .......... .......... ..........
x75816  AKLIEHEKID LNTTVVPHIV VGRDSRESSP YLLRCLTSSM ASVFHAQVLD

        401                                               450
124077  VKRVSFEDAL KAPTTKRHDY ITP...YVDD LAAVVDM..D VIRE......
u08369  VKRISLDEAM ASGHVKEQDL VQP...FVEG LADIVDM..A AIQK......
x72016  LNLNKLGKNQ KYGPLL.VDI IDPAKAYVQF LKEIFDF..D LIKSFLAKQR
x74823  LDLGTIGKNK KYGPLL.VDI IDITKDYVNF LKEIFDF..D LIKKFIDNQR
 pgmrp  LGRQEFDLEN KFKP.FRVEI VDPVDIYLNL LRTIFDF..H AIKGLLTG..
m83088  LGKQQFDLEN KFKP.FTVEI VDSVEAYATM LRSIFDF..S ALKELLSG..
124117  IG......SF KVDE.LTVDV IDPVADYAAL MEELFDF..G AIRSLIAG..
  pfus  IGVYKFEGTR LEKSHFEVKV VDTVQDYTQL MQKLFDF..D LLKGLFSN..
x57132  EGLLHSSYKV GESVGSAKRI DDVIGRYIAH LKHSF..... ......PKHL
112968  E....ISCVD SAELGKASRI VDAAGRYIEF CKATF..... ......PNEL
127632  NQFPPVDPAR RGTL....RQ ISVLKEYVDH LMGYV..... ......DLAN
127646  NQFPPVDPAR RGTL....RQ ISVLKEYVDH LMGYV..... ......DLAN
d13231  NQFPPVDPAR RGTL....RQ ISVLKEYVDH LMGYV..... ......DLAN
x59886  GDFPPVNEAA RGSY....RQ ISLRDAYIDH LLGYI..... ......SVNN
m84642  GDFPPVNDAA RGSY....RQ ISLRDAYIDH LLAYI..... ......SVNN
m77127  NDFPPVDETK RGRY....QQ INLRDAYVDH LFGYI..... ......NVKN
111721  NDFPPVDETK RGRY....QQ INLRDAYVDH LFGYI..... ......NVKN
104596  NDFPPVDETK RGRY....QQ INLRDAYVDH LFGYI..... ......NVKN
m83231  DTAAAGEPTA AEHS....R. .TDKTAYLEH LLSYV..... ......DRST
u02489  DGF..VAADK QGSV....TE KDISGAYHDH IVGHV..... ......K...
u02490  DGF..ARAGK QGSV....TE KDISGEYLKH ITGHI..... ......R...
m60873  NDL..ASG.. VGSV....EQ VDILPRYFKQ IRDDI..... ......A...
x79075  RRF..VKG.. HGAV....VD VDIIEDYESY ITKHI..... ......Q...
u20583  ADY..VSG.. EGSS....EQ LDIKDAYVER LLRDD..... ......D...
x75898  RKYANRQAKV SLTLINPPTK VNLMSAYANH LRDIIKE..R .INHPTNYDT
113289  .EMIDLKVAN DDSLITYLNE DIFRQYYEDC KQALIKT..N .I.......N
x56793  .EITKHDEAA ILSVEDTCSH LELKELIVSE MAAVNYI..S .RYTSLFSTP
x75816  LGCVTTPQLH YITDLSNRRK LEGDTAPVAT ERDYYSFFIG AFNELFATYQ
```

```
        451                                                   500
124077  ..SGV...SI GIDPLGGAAV DYWQPI.IDK YGINATIVSK EVDPTFRFMT
u08369  ..AGL...TL GVDPLGGSGI EYWKRI.GEY YNLNLTIVND QVDQTFRFMH
x72016  KDKGW...KL LFDSLNGITG PYGKAIFVDE FGLPAEEVLQ NWHPLPDFGG
x74823  STKNW...KL LFDSMNGVTG PYGKAIFVDE FGLPADEVLQ NWHPSPDFGG
 pgmrp  .PSQL...KI RIDAMHGVMG PYVRKVLCDE LGAPANSAI. NCVPLEDFGG
m83088  .PNRL...KI CIDAMHGVVG PYVKKILCEE LGAPANSAV. NCVPLEDFGG
124117  .G..F...KV VVDSMSAVTG PYAVEILEKR LGAPKGSVR. NATPLPDFGG
  pfus  ..KDF...SF RFDGMHGVAG PYAKHIFGTL LGCSKESLL. NCDPSEDFGG
x57132  NLQSL...RI VLDTANGAAY KVAPVVFSEL GA.....DVL VINDEPN..G
112968  SLSEL...KI VVDCANGATY HIAPNVLREL GA.....NVI AIGCEPN..G
127632  FTRPL...KL VVNSGNGAAG HVIDEVEKRF AAAGVPVTFI KVHHQPD..G
127646  FTRPL...KL VVNSGNGAAG HVIDEVEKRF AAAGVPVTFI KVHHQPD..G
d13231  FTRPL...KL VVNSGNGAAG HVIDEVEKRF AAAGAPVTFI KVHHQPD..G
x59886  LT.PL...KL VFNAGNGAAG PVIDAIEARL KALGAPVEFI KIHNTPD..G
m84642  LT.PL...KL VVNSGNGAAG PVIDAIEARL KALGAPVEFI KIHNTPD..G
m77127  LT.PL...KL VINSGNGAAG PVVDAIEARF KALGAPVELI KVHNTPD..G
111721  LT.PL...KL VINSGNGAAG PVVDAIEARF KALGAPVELI KVHNTPD..G
104596  LT.PL...KL VINSGNGAAG PVVDAIEARF KALGAPVELI KVHNTPD..G
m83231  L.KPL...KL VVNAGNGGAG LIVDLLAPHL .....PFEFV RVFHEPD..G
u02489  LKRPI...NI AIDAGNGVGG AFAGKLYKGL .....GNEVT ELFCEVD..G
u02490  LKRPM...NI AIDAGNGVGG AFAGKLYKGL .....GNEVT ELFCDVD..G
m60873  MAKPM...KV VVDCGNGVAG VIAPQLIEAL .....GCSVI PLYCEVD..G
x79075  LDRPL...KV VVDCGNGIAG KVAPALYRKL .....GCEVV ELFCEVD..G
u20583  GTRDL...TI AWDAGNGASG EDPAPPDREV .....PGKHV LLFDEID..G
x75898  PLQGF...QI IVNAGNGSGG FFTWDVLDKL GA....DTFG SLYLNPD..G
113289  ESKEF...SI VFSGQHGTAC KRLPEFLKLL GYKN..IILV EEQCIFD..G
x56793  FLKNK...RI GIYEHSSAGR DLYKPLFIAL GAEVVSLGRS DNFVPID..T
x75816  LEKRLSVPKL FIDTANGIGG PQLKKLLASE DWDVPAEQVE VINDRSDVPE

        501                                                   550
124077  ADWDGQIRMD CSSPYAMARL VGMK..DKFD .IAFANDTDA DRHGI....V
u08369  LDKDGAIRMD CSSECAMAGL LALR..DKFD .LAFANDPDY DRHGI....V
x72016  LHPDPNLTYA RTLVDRVDR. ......EKIA .FGAASDGDG DRNMI....Y
x74823  MHPDPNLTYA SSLVKRVDR. ......EKIE .FGAASDGDG DRNMI....Y
 pgmrp  QHPDPNLTYA TTLLEAM... ...KGGE.YG .FGAAFDADG DRYMI....L
m83088  HHPDPNLTYA ADLVETM... ...KSGE.HD .FGAAFDGDG DRNMI....L
124117  HHPDPNLVHA KELYDDV... ...MSPEGPD .FGAASDGDG DRNMV....V
  pfus  GHPDPNLTYA HDLVELLDIH KKKDVGTVPQ .FGAACDGDA DRNMI....L
x57132  CNIN.EQC.G ALHP...NQL SQEVKKYRAD .LGFAFDGDA DRLVV....V
112968  VNIN.AEV.G ATDV...RAL QARVLAEKAD .LGIAFDGDG DRVIM....V
127632  HFPN.GIP.N PLLPECRQDT ADAVREHQAD .MGIAFDGDF DRCFL....F
127646  HFPN.GIP.N PLLPECRQDT ADAVREHQAD .MGIAFDGDF DRCFL....F
d13231  HFPN.GIP.N PLLPECRQDT ADAVRAHQAD .MGIAFDGDF DRCFL....F
x59886  TFPN.GIP.N PLLPECRDDT RKAVIEHGAD .MGIAFDGDF DRCFL....F
m84642  TFPN.GIP.N PLLPECRDDT RKAVIEHGAD .MGIAFDGDF DRCFL....F
m77127  NFPN.GIP.N PLLPECRDDT RNAVIKHGAD .MGIAFDGDF DRCFL....F
111721  NFPN.GIP.N PLLPECRDDT RNAVIKHGAD .MGIAFDGDF DRCFL....F
104596  NFPN.GIP.N PLLPECRDDT RNAVIKHGAD .MGIAFDGDF DRCFL....F
m83231  NFPN.GIP.N PLLQENRDAT AKAVKEHGAD .FGIAWDGDF DRCFF....F
u02489  NFPN.HHP.D PSKPENLQDL IAALKNGDAE .IGLAFDGDA DRLGV....V
u02490  TFPN.HHP.D PSKPKNLQDL IAALKNGDAE .IGLAFDGDA DRLGV....V
m60873  NFPN.HHP.D PGKPENLKDL IAKVKAENAD .LGLAFDGDG DRVGV....V
x79075  HFPN.HHP.D PTIPANLTDL IHKVKETQAD .LGLAFDGDA DRLGI....V
u20583  NFPN.HHP.D PTVEKNLVDL KAAVAEHGCD .IGIGFDGDG DRIGA....I
x75898  MFPN.HIP.N PEDKKAMALT RAAVLENSAD .LGIVFDTDV DRSGV....V
113289  NFSNTPTP.N PENRAAWDLS IEYADKNNAN .VIIQVDPDA DRFAL.GVRY
x56793  .........E AVSKEDREKA RSWAKEFDLD .AIFSTDGDG DRPLI.A...
x75816  LL...NFECG ADYVKTNQRL PKGLSPSSFD SLYCSFDGDA DRVVFYYVDS
```

```
        551                                                600
l24077  SGKYGLMNPN HYLAVAIEYL FNNRENW... .NASA.GVGK TVVSSSMIDR
u08369  TPA.GLMNPN HYLAVAINYL FQHRPQW... .GKDV.AVGK TLVSSAMIDR
x72016  GYGPAFVSPG DSVAIIAEYA PEIPYFA... .KQGIYGLAR SFPTSSAIDR
x74823  GYGPSFVSPG DSVAIIAEYA AEIPYFA... .KQGIYGLAR SFPTSGAIDR
 pgmrp  GQNGFFVSPS DSLAIIAANL SCIPYFR... .QMGVRGFGR SMPTSMALDR
m83088  GKHGFFVNPS DSVAVIAANI FSIPYFQ... .QTGVRGFAR SMPTSGALDR
l24117  GK.GMFVTPS DSLAIIAANA KLAPGY.... .AAGISGIAR SMPTSAAADR
  pfus  GRQ.FFVTPS DSLAVIAANA NLI...F... .KNGLLGAAR SMPTSGALDK
x57132  DNLGNIVHGD KLLGVLGVYQ KSKNALS... .SQAI..VAT NMSNLALKE.
l12968  DHEGNKVDGD QIMYIIAREG LRQGQLR... .GGA...VGT LMSNMGLEL.
l27632  DDEASFIEGY YIVGLLAEAF LQKQP..... .GAKI..IHD PRLTWNTVD.
l27646  DDEASFIEGY YIVGLLAEAF LQKQP..... .GAKI..IHD PRLTWNTVD.
d13231  DDEASFIEGY YIVGLLAEAF LQKQP..... .GAKI..IHD PRLTWNTVD.
x59886  DEKGQFIEGY YIVGLLAEAF LEKHP..... .GAKI..IHD PRLTWNTEA.
m84642  DEKGQFIEGY YIVGLLAEAF LEKHP..... .GAKI..IHD PRLTWNTEA.
m77127  DEKGQFIEGY YIVGLLAEAF LEKNP..... .GAKI..IHD PRLSWNTVD.
l11721  DEKGQFIEGY YIVGLLAEAF LEKNP..... .GAKI..IHD PRLSWNTVD.
l04596  DEKGQFIEGY YIVGLLAEAF LEKNP..... .GAKI..IHD PRLSWNTVD.
m83231  DHTGRFIEGY YLVGLLAQAI LAKQP..... .GGKV..VHD PRLTWNTVE.
u02489  TKDGNIIYPD RQLMLFAQDV LNRNP..... .GAKV..IFD VKSTRLLAP.
u02490  TKDGNIIYPD RQLMLFAQDV LNRNP..... .GAKA..IFD VESTRLVAP.
m60873  TNTGTIIYPD RLLMLFAKDV VSRNP..... .GADI..IFD VKCTRRLIA.
x79075  TDKGEIIWPD RQMMLFSMDV LSRLP..... .GSDI..VFD VKCSRSLAE.
u20583  DHLGRVVWGD QLVAIYAADV LKSHP..... .GATI..IAD VKASQTLFD.
x75898  DNKGNPINGD KLIALMSSIV LKEHP..... .ETTI..VTD ARTSIGLSR.
l13289  KNSWRFLSGN QMGIIYTDYI LKNKTFT... .KKPY..IVS SYVSTNLIDR
x56793  DEAGEWLRGD .ILGLLCSLA LDAEAVA... .IPVS..CNS IISSGRFFKH
x75816  GSKFHLLDGD KISTLFAKFL SKQLELAHLE HSLKIGVVQT AYANGSSTAY

        601                                                650
l24077  VAKEIGRKLV EVPVGFKWFV D......GLY NGTLGFGGEE SAGASFLRRA
u08369  VVNDLGRKLV EVPVGFKWFV D......GLF DGSFGFGGEE SAGASFLRFD
x72016  VAAKKGLRCY EVPTGWKFFC A......LFD AKKLSICGEE SFGT......
x74823  VAKAHGLNCY EVPTGWKFFC A......LFD AKKLSICGEE SFGT......
 pgmrp  VAKSMKVPVY ETPAGWRFFS N......LMD SGRCNLCGEE SFGT......
m83088  VASATKIALY ETPTGWKFFG N......LMD ASKLSLCGEE SFGT......
l24117  VAEKLGLGMY ETPTGWKFFG N......LMD AGKVTICGEE SFGT......
  pfus  VAAKNGIKLF ETPTGWKFFG N......LMD AGLINLCGEE SFGT......
x57132  YLKSQDLELK HCAIGDKFVS EC.....MRL NKAN.FGGEQ S...GHIIFS
l12968  ALKQLGIPFA RAKVGDRYVL EK.....MQE KGWR.IGAEN S...GHVILL
l27632  IVTRNGGQPV MSKTGHAFIK ER.....MRQ EDAI.YGGEM S...AHHYFR
l27646  IVTRNGGQPV MSKTGHAFIK ER.....MRQ EDAI.YGGEM S...AHHYFR
d13231  IVTRSGGQPV MSKTGHAFIK ER.....MRQ EDAI.YGGEM S...AHHYFR
x59886  VVTAAGGTPV MSKTGHAFIK ER.....MRT EDAI.YGGEM S...AHHYFR
m84642  VVTAAGGTPV MSKTGHAFIK ER.....MRT EDAI.YGGEM S...AHHYFR
m77127  VVTAAGGTPV MSKTGHAFIK ER.....MRK EDAI.YGGEM S...AHHYFR
l11721  VVTAAGGTPV MSKTGHAFIK ER.....MRK EDAI.YGGEM S...AHHYFR
l04596  VVTAA.GTPV MSKTGHAFIK ER.....MRK EDAI.YGGEM S...AHHYFR
m83231  MVEDAGGIPV LCKSGHAFIK EK.....MRS ENAV.YGGEM S...AHHYFR
u02489  WIKEHGGEAI MEKTGHSFIK SA.....MKK TGAL.VAGEM S...GHVFFK
u02490  WIKEHGGKAI MEKTGHSFIK SA.....MKE TGAP.VAGEM S...GHIFFK
m60873  LISGYGGRPV MWKTGHSLIK KK.....MKE TGAL.LAGEM S...GHVFFK
x79075  IIKKYGGNPV MWRTGHSILK AK.....LFE IGAP.LAGEM S...GHIFFK
u20583  EIARLGGNPL MWKTGHSLLK AK.....MAE TGSP.LAGEM S...GHIFFA
x75898  FITNRGGKHC LYRVGYRNVI DKGVQLNEDD IETH.LMMET S...GHGALK
l13289  IIKEYHGEVY RVGTGFKWVG DKINKIKDSE EFVVGFEEAV G...ALNSTI
x56793  VKLTKIGSPY VIEAFNELSR SYSRIVGFEA NGGFLLGSDI C.......IN
x75816  IKNTLHCPVS CTKTGVKHL. HHEAATQYDI GIYFEANGHG TIIFSGKFHR
```

```
          651                                                700
 l24077   GTVWSTDKDG IILGLLAAEI TARTKRT... ........PG AAYEDMTRRL
 u08369   GTPWSTDKDG IIMCLLAAEI TAVTGKN... ........PQ EHYNELAKRF
 x72016   GSNHIREKDG LWAIIAWLNI LAIYHRRNPE KEASIKTIQD EFWNEYGRTF
 x74823   GSNHVREKDG VWAIMAWLNI LAIYNKHHPE NEASIKTIQN EFWAKYGRTF
  pgmrp   GSDHLREKDG LWAVLVWLSI IAA.RKQ... ...SVEEIVR DHWAKFGRHY
 m83088   GSDHIREKDG LWAVLAWLSI LAT.RKQ... ...SVEDILK DHWQKHGRNF
 l24117   GSNHVREKDG LWAVLYWLNI VAA.RKE... ...SVKDIVT KHWAEYGRNY
   pfus   GSNHIREKDG IWAVLAWLTI LAH.KNKNTD HFVTVEEIVT QYWQQFGRNY
 x57132   D..YAKTGDG LVCALQVSAL VLESKLVSSV RLNPF..EL. YPQNL.VNL.
 l12968   D..KTTTGDG IVAGLQVLAA MARNHMSLHD LCSGM..KM. FPQIL.VNVR
 l27632   D..FAYCDSG MIPWLLVAEL LCLKNSSLKS LVADRQAAF. PASGE.INRK
 l27646   D..FAYCDSG MIPWLLVAEL LCLKNSSLKS LVADRQAAF. PASGE.INRK
 d13231   D..FAYCDSG MIPWLLVAEL LCLKNSSLKS LVADRQAAF. PASGE.INRK
 x59886   D..FAYCDSG MIPWLLVAEL VCLKRQSLGE LVRDRMAAF. PASGE.INSR
 m84642   D..FAYCDSG MIPWLLVAEL VCLKGQSLGE LVRDRMAAF. PASGE.INSR
 m77127   D..FAYCDSG MIPWLLVAEL VCLKDKTLGE LVRDRMAAF. PASGE.INSK
 l11721   D..FAYCDSG MIPWLLVAEL VCLKDKTLGE LVRDRMAAF. PASGE.INSK
 l04596   D..FAYCDTG MIPWLLVAEL VCLKGKTLGE LVRDRMAAF. PASGE.INSK
 m83231   E..FAYADSG MIPWLLIAEL VSQSGRSLAD LVEARMQKF. PCSGE.INFK
 u02489   ERWFGF.DDG LYAGARLLEI LSASDNPS.E VL.DNLPQS. ISTPE.LNIS
 u02490   ERWFGF.DDG LYAGARLLEI LSASDNPT.E VL.NNLPQS. ISTPE.LNIA
 m60873   ERWFGF.DDG IYSAARLLEI LSQDQRDSEH VF.SAFPSD. ISTPE.INIT
 x79075   DEWFGF.DDG IYVGARLLRI ISQTNQRTSE IF.AELPDS. VNTPE.LKLP
 u20583   DKWYGF.DDA LYCAVRLIGL VSKLNQPLSE LR.DRLPDV. VNTPE.TRFQ
 x75898   ENYF.LDDGA YMVVKIIIEM VRMRLSGSSE GIGNLIEDL. EDPVESVELR
 l13289   NRDKDAYQAA ALALEIYNEC LKNNINIIDH LEKNIYGKYG IIHNDTISFT
 x56793   EQNL....HA LPTRDAVLPA IMLLYKSRNT SISALVNELP TRYTHSDRLQ
 x75816   TIKSELSKSK LNGDTLALRT LKCFSELINQ TVGDAISDML AVLATLAILK

          701                                                750
 l24077   GTPY.YARID APADPEQKAI LKNLSPEQIG MTELAGEPIL STLTNAPGNG
 u08369   GAPS.YNRLQ AAATSAQKAA LSKLSPEMVS ASTLAGDPIT ARLTAAPGNG
 x72016   FTRYDYEHIE CEQAEKVVAL LSE....FVS RPNVCGSHFP ADESLTVIDC
 x74823   FTRYDFEKVE TEKANKIVDQ LRA....YVT KSGVVNSAFP ADESLKVTDC
  pgmrp   YCRFDYEGLD PKTTYY...I MRDLEALVTD KSFIGQQFAV GSHVYSVAKT
 m83088   FTRYDYEEVE AEGANK...M MKDLEALMFD RSFVGKQFSA NDKVYTVEKA
 l24117   YSRHDYEEVD SDAANTLVAI LREKLATLPG TSYGN..... ....LKVAAA
   pfus   YSRYDYEQVD SAGANKMMEH LKTKFQYF.. .......... .EQLKQGNKA
 x57132   .N.......V QKKPPL..ES LKGYNALLKE LDKL...... ..........
 l12968   YT.......A GSGDPLEHES VKAVTAEVEA ALGN...... ..........
 l27632   LG.......N A......AEA IRRIRAQYEP AAAHIDTTDG ISIEYP....
 l27646   LG.......N A......AEA IARIRAQYEP AAAHIDTTDG ISIEYP....
 d13231   LG.......N A......AEA IARIRAQYEP AAAHIDTTDG ISIEYP....
 x59886   LA.......E P......AAA IARVEAHFAE EAQAVDRTDG LSMSFA....
 m84642   LA.......E P......AAA IARVEAHFAE EAQAVDRTDG LSMSFA....
 m77127   LA.......Q P......VEA INRVEQHFSR EALAVDRTDG ISMTFA....
 l11721   LA.......Q P......VEA INRVEQHFSR EALAVDRTDG ISMTFA....
 l04596   LA.......H P......VEA INRVEQHFSR EALAVDRTDG ISMTFA....
 m83231   VD.......D A......KAA VARVMAHYGD QSPELDYTDG ISADFG....
 u02489   LP.......E GSNGHQVIEE LAAKAE..FE GATEIITIDG LRVEFP....
 u02490   LP.......E GSNGHQVIDE LAAKAE..FE GATEIITIDG LRVEFP....
 m60873   VT.......E DSK.FAIIEA LQRDAQ..W. GEGNITTLDG VRVDYP....
 x79075   MT.......E EKK.QPFMQA LLKKAD..F. GNAKLITIDG LRVEFE....
 u20583   VS.......E ERK.FQVVQE VEGRSSRLMA EGADVNDIDG VRVKDA....
 x75898   MDVISEPRYA KTKAVEVIDT FRRYVEEDKL EGWMLDSCGD CWVGEGCL.V
 l13289   FVENNWKELV KKSLDKILKY SEKTIGNRTI TSIKYNEVGG ..........
 x56793   GITTDKSQSL ISMGRENLSN LLSYIGLENE GAISTDMTDG MRITLRDG.C
 x75816   MSPMDWDEEY TDLPNKLVKC IVPDRSIFQT TDQERKLLNP VGLQ......
```

```
        751                                              800
124077  AAIG...... .......GLK VSAKDG.WFA ARPSGTENV. ..YKIYAESF
u08369  ASIG...... .......GLK VMTDNG.WFA ARPSGTEDA. ..YKIYCESF
x72016  GDFSYRD.LD GSISENQGLF VKFSNGTKFV LRLSGTGSSG ATIRLYVEKY
x74823  GDFSYTD.LD GSVSDHQGLY VKLSNGARFV LRLSGTGSSG ATIRLYIEKY
 pgmrp  DSFEYVDPVD GTVTKKQGLR IIFSDASRLI FRLSSSSGVR ATLRLYAESY
m83088  DNFEYSDPVD GSISRNQGLR LIFTDGSRIV FRLSGTGSAG ATIRLYIDSY
124117  DDFAYHDPVD QSVSKNQGIR ILFEGGSRIV LRLSGTGTAG ATLRLYVERY
  pfus  DIYDYVDPVD QSVSKNQGVR FVFGDGSRII FRLSGTGSVG ATIRIYFEQF
x57132  .......... .......... .....EIRHL IRYSGTEN.. ....KLRILL
112968  .......... .......... .....RGRVL LRKSGTEP.. ....LIRVMV
127632  .......... .......... .....EWRFN LRTSNTEP.. ......VVRL
127646  .......... .......... .....EWRFN LRTSNTEP.. ......VVRL
d13231  .......... .......... .....EWRFN LRTSNTEP.. ......VVRL
x59886  .......... .......... .....DWRFN LRSSNTEP.. ......VVRL
m84642  .......... .......... .....DWRFN LRSSNTEP.. ......VVRL
m77127  .......... .......... .....DWRFN LRTSNTEP.. ......VVRL
111721  .......... .......... .....DWRFN LRTSNTEP.. ......VVRL
104596  .......... .......... .....DWRFN LRLLNTEP.. ......VVRL
m83231  .......... .......... .....QWRFN LRSSNTEP.. ......LLRL
u02489  .......... .......... .....DGFGL MRASNTTP.. ....ILVLRF
u02490  .......... .......... .....DGFGL MRASNTTP.. ....ILVLRF
m60873  .......... .......... .....KGWGL VRASNTTP.. ....VLVLRF
x79075  .......... .......... .....DGWGL IRPSNTSP.. ....YLILRF
u20583  .......... .......... .....DGWWL LRASNTQD.. ....VLVARA
x75898  DLNENPTAID AHMYRVKVLD NEQNEHGWVH LRQSVHNP.. ....NIAVNM
113289  .......... .......CYD WILDGDSWLR FRMSGTEP.. ....KFKVYY
x56793  I......... .......... ........VH LRASGNAP.. ....ELRCYA
x75816  .......... ....DKIDLV VAKYPMGRSF VRASGTED.. ......AVRV

        801                                            848
124077  KSAAHLKAIQ TEAQDAISAL FAKAAQKNAG *......... ........
u08369  LGEEHRKQIE KEAVEIVSEV LKNA*..... .......... ........
x72016  ..TDKKENYG QTADVFLKPV INSIVKFLRF KEILGTDEPT VRT*....
x74823  ..CDDKSQYQ KTAEEYLKPI INSVIKFLNF KQVLGTEEPT VRT*....
 pgmrp  ..ERDPSGHD QEPQAVLSPL IAIALKISQI HERTGRRGPT VIT.....
m83088  ..EKDVAKIN QDPQVMLAPL ISIALKVSQL QERTGRTAPT VIT*....
124117  ..EPDAARHG IETQSALADL ISVADTIAGI KAHTADSEPT VIT*....
  pfus  ..EQQ..QIQ HETATALANI IKLGLEISDI AQFTGRNEPT VIT*....
x57132  EAKDEK.... ..LLESKMQE LKEFFEGHLC *......... ........
112968  EGEDEA.... ..QVTEFAHR IADAVKAV*. .......... ........
127632  NVESRA.... ..DVALMNEK TTELLNLLKE ELL*...... ........
127646  NVESRA.... ..DTALMNAK TEEILALLK* .......... ........
d13231  NVESRA.... ..DTALMNEK TAELLNLLKE ESL*...... ........
x59886  NVESRG.... ..DIPLMEAR TRTLLALLNQ *......... ........
m84642  NVESRG.... ..DIPLMEAR TKEILQLLNS *......... ........
m77127  NVESRG.... ..DVPLMEAR TRTLLTLLNE *......... ........
111721  NVESRG.... ..DVPLMEAR TRTLLTLLNE *......... ........
104596  NVESRG.... ..DVPLMEEK TKLILELLNK *......... ........
m83231  NVETRG.... ..DAALLETR TQEISNLLRG *......... ........
u02489  EADTQA.... ..AIERIQNR FKA...VIES NPHLIWPL*. ........
u02490  EADTQE.... ..AIERIQNQ FKA...VIES NPNLIWPL*. ........
m60873  EADPEE.... ..ELERIKTV FRNQLKAVDS SLPVPF*... ........
x79075  EADTEE.... ..KLKRIQEI FRTQLRMIDN ALELPF*... ........
u20583  ESGTRR.... ..SWERLKGM VVAHWKPPAS RPFLRGRRQL H*......
x75898  QSSIPG.... ..GCRSMTEI FKDKFLFASG LDKVVDTSQI EQYVKEH*
113289  NLYGENLNAL SQEAKTINDQ IKTLLNL*.. .......... ........
x56793  EANLLNRAQD LVNTTLANIK KRCLL*.... .......... ........
x75816  YAECKDSSKL GQFCDEVVEH VKASA*.... .......... ........
```

Appendix B Pairwise genetic distance scores between PGM, PMM and PGM-related sequences calculated for construction of the phylogenetic tree based on the neighbour-joining distance method of Saitou and Nei (1987). Below the diagonal are shown the absolute distances, above are shown the mean distances.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1 124077	-	0.420	0.740	0.757	0.757	0.762	0.748	0.760	0.827	0.827	0.845	0.843	0.841	0.838	0.831	0.826	0.826	0.832	0.840	0.824	0.827	0.843	0.842	0.844	0.864	0.863	0.861	0.893
2 u08369	230	-	0.733	0.755	0.762	0.759	0.759	0.789	0.829	0.813	0.819	0.823	0.821	0.833	0.828	0.816	0.816	0.814	0.822	0.840	0.845	0.830	0.825	0.821	0.849	0.868	0.862	0.878
3 x72016	372	363	-	0.209	0.575	0.487	0.489	0.564	0.828	0.800	0.819	0.815	0.817	0.799	0.796	0.799	0.799	0.798	0.836	0.831	0.835	0.818	0.818	0.807	0.869	0.856	0.873	0.873
4 x74823	380	373	119	-	0.558	0.477	0.476	0.551	0.825	0.792	0.805	0.805	0.805	0.798	0.798	0.791	0.791	0.793	0.841	0.825	0.825	0.818	0.811	0.802	0.861	0.856	0.884	0.869
5 pgmrp	339	336	283	274	-	0.324	0.485	0.504	0.836	0.772	0.833	0.834	0.830	0.831	0.836	0.826	0.826	0.826	0.823	0.837	0.837	0.826	0.833	0.824	0.855	0.860	0.877	0.882
6 m83088	381	374	267	261	164	-	0.438	0.462	0.812	0.780	0.818	0.816	0.815	0.809	0.814	0.811	0.811	0.811	0.818	0.806	0.821	0.812	0.829	0.823	0.853	0.861	0.870	0.871
7 124117	359	359	258	251	236	236	-	0.465	0.810	0.777	0.808	0.809	0.810	0.807	0.804	0.804	0.804	0.809	0.821	0.809	0.811	0.803	0.815	0.811	0.858	0.872	0.882	0.876
8 pfus	390	399	307	299	245	250	245	-	0.823	0.820	0.842	0.841	0.845	0.839	0.839	0.839	0.839	0.833	0.841	0.816	0.819	0.811	0.807	0.838	0.887	0.854	0.890	0.896
9 x57132	345	339	342	340	312	329	319	340	-	0.571	0.759	0.759	0.759	0.756	0.749	0.763	0.763	0.761	0.768	0.748	0.758	0.748	0.737	0.757	0.774	0.851	0.797	0.850
10 112968	345	334	328	324	284	312	306	336	250	-	0.750	0.752	0.748	0.761	0.752	0.775	0.775	0.783	0.754	0.759	0.780	0.728	0.745	0.746	0.761	0.840	0.825	0.854
11 127632	361	343	349	342	324	341	324	353	325	321	-	0.015	0.015	0.252	0.245	0.236	0.236	0.242	0.425	0.691	0.711	0.692	0.689	0.703	0.761	0.811	0.834	0.865
12 127646	359	345	344	339	321	337	321	349	324	322	7	-	0.018	0.246	0.237	0.239	0.239	0.243	0.422	0.686	0.706	0.691	0.691	0.700	0.763	0.816	0.834	0.867
13 d13231	361	345	348	342	323	340	325	354	325	320	7	8	-	0.249	0.243	0.236	0.236	0.242	0.425	0.690	0.710	0.694	0.694	0.713	0.768	0.816	0.837	0.865
14 x59886	357	348	337	336	320	334	321	348	323	325	115	112	114	-	0.033	0.109	0.109	0.126	0.397	0.672	0.688	0.686	0.691	0.677	0.738	0.815	0.814	0.860
15 m84642	354	346	336	336	322	336	320	348	320	321	112	108	111	15	-	0.114	0.114	0.115	0.393	0.670	0.686	0.686	0.691	0.680	0.740	0.818	0.821	0.858
16 m77127	352	341	337	333	318	335	320	348	326	331	108	109	108	50	52	-	0.000	0.029	0.420	0.674	0.700	0.691	0.698	0.684	0.749	0.813	0.819	0.862
17 111721	352	341	337	333	318	335	320	348	326	331	108	109	108	50	52	0	-	0.029	0.420	0.674	0.700	0.691	0.698	0.684	0.749	0.813	0.819	0.862
18 104596	352	338	336	333	317	334	321	345	324	332	110	110	110	57	52	13	13	-	0.422	0.679	0.702	0.696	0.703	0.686	0.748	0.817	0.823	0.862
19 m83231	352	338	347	348	311	332	320	344	325	318	191	189	191	178	176	188	188	188	-	0.686	0.709	0.687	0.697	0.693	0.758	0.815	0.833	0.874
20 u02489	352	352	353	350	324	337	326	342	315	321	304	299	305	293	292	294	294	294	297	-	0.082	0.443	0.493	0.592	0.772	0.842	0.825	0.869
21 u02490	353	354	355	350	324	343	327	343	319	330	313	308	314	300	299	305	305	304	307	38	-	0.449	0.487	0.605	0.772	0.845	0.825	0.867
22 m60873	359	346	346	345	317	337	321	339	315	308	305	302	311	300	300	302	302	302	298	201	204	-	0.445	0.594	0.771	0.828	0.832	0.865
23 x79075	357	343	347	343	320	345	327	338	311	315	304	302	306	302	302	305	305	306	303	223	220	203	-	0.590	0.743	0.817	0.835	0.862
24 u20583	362	345	347	344	323	348	331	356	321	318	312	308	325	298	299	301	301	300	303	271	277	275	269	-	0.762	0.853	0.819	0.858
25 x75898	438	423	446	441	383	430	416	461	340	334	343	341	354	330	331	335	335	332	335	352	352	354	336	358	-	0.843	0.848	0.883
26 113289	430	427	410	409	361	404	395	414	370	367	357	359	368	358	359	357	357	356	352	369	370	367	356	384	436	-	0.870	0.905
27 x56793	377	376	393	398	341	383	372	388	311	321	332	332	333	323	326	325	325	326	325	325	325	328	330	325	390	389	-	0.897
28 x75816	436	426	439	436	403	431	418	447	358	358	372	373	372	369	368	370	370	369	369	366	365	364	363	364	429	438	409	-

REFERENCES

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., Sutton, G., Blake, J.A., Brandon, R.C., Chiu, M-W., Clayton, R.A., Cline, R.T., Cotton, M.D., Earle-Hughes, J., Fine, L.D., FitzGerald, L.M., FitzHugh W.M., Fritchman, J.L., Geoghagen, N.S.M., Glodek, A., Gnehm, C.L., Hanna, M.C., Hedblom, E., Hinkle Jr., P.S., Kelley, J.M., Klimek, K.M., Kelley, J.C., Liu, L-I., Marmaros, E., Merrick, J.M., Moreno-Palanques, R.F., McDonald, L.A., Nguyen, D.T., Pellegrino, S.M., Philips, C.A., Ryder, S.E., Scott, J.L., Saudek, D.M., Shirley, R., Small, K.V., Spriggs, T.A., Utterback, T.R., Weidman, J.F., Li, Y., Barthlow, R., Bednarik, D.P., Cao, L., Cepeda, M.A., Coleman, T.A., Collins, E-J., Dimke, D., Feng, P., Ferrie, A., Fischer, C., Hastings, P.A., He, W-W., Hu, J-S., Huddleston, K.A., Greene, J.M., Gruber, J., Hudson, P., Kim, A., Kozak, D.L., Kunsch, C., Ji, H., Li, H., Meissner, P.S., Olsen, H., Raymond, L., Wei, Y-F., Wing, J., Xu, C., Yu, G-L., Ruben, S.M., Dillon, P.J., Fannon, M.R., Rosen, C.A., Haseltine, W.A., Fields, C., Fraser, C.M. & Venter, J.C. (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature.* **377**, Supplement, 3-174.

Ajmar, F., Garre, C., Sessarego, M., Ravazzolo, R., Barresi, R., Scarra, G.B. & Lituania, M. (1983). Expression of erythroid acetylcholinesterase in the K562 leukemia cell line. *Cancer Res.* **43**, 5560-5563.

Andersen, A.P., Wyroba, E., Reichman, M., Zhao, H. & Satir, B.H. (1994). The activity of parafusin is distinct from that of phosphoglucomutase in the unicelllular eukaryote *Paramecium. Biochem. Biophys. Res. Comm.* **200**, 1353-1358.

Anderson, L. & Jolles, G.R. (1957). A study of the linkage of phosphorus to protein in phosphoglucomutase. *Arch. Biochem. Biophys.* **70**, 121-128.

Andersson, L.C., Nilsson, K. & Gahmberg, C.G. (1979). K562 - A human erythroleukemic cell line. *Int. J. Cancer.* **23**, 143-147.

Aoyama, K., Haase, A.M. & Reeves, P.R. (1994). Evidence for effect of randon genetic drift on G+C content after lateral transfer of fucose pathway genes to *Escherichia coli* K12. *Mol. Biol. Evol.* **11**, 829-838.

Auger, D., Bounelis, P. & Marchase, R.B. (1993). Phosphoglucomutase is a cytoplasmic glycoprotein implicated in the regulated secretory pathway. *Molecular Mechanisms of Membrane Traffic.* Ed. Moore, D.J., Bergeron, J.J.M. & Howell, K.M. Springer, New York. 289-292.

Bao, J., Sifers, R.N., Kidd, V.J., Ledley, F.D. & Woo, S.L.C. (1987). Molecular evolution of serpins: homologous structure of the human α_1-antichymotrypsin and α_1-antitrypsin genes. *Biochemistry.* **26**, 7755-7759.

Bark, J.E., Harris, M.J. & Firth, M. (1976). Typing of the common phosphoglucomutase variants using isoelectric focusing - A new interpretation of the phosphoglucomutase system. *J. Foren. Sci. Soc.* **16**, 115-120.

Belkin, A.M., Klimanskaya, I.V., Lukashev, M.E., Lilley, K., Critchley, D.R. & Koteliansky, V.E. (1994). A novel phosphoglucomutase-related protein is concentrated in adherens junctions of muscle and non-muscle cells. *J. Cell Sci.* **107**, 159-173.

Belkin, A.M. & Burridge, K. (1995). Association of aciculin with dystrophin and utrophin. *J. Biol. Chem.* **270**, 6328-6337.

Bernstein, M., Hoffmann, W., Ammerer, G. & Schekman, R. (1985). Characterization of a gene product (sec53p) required for protein assembly in yeast endoplasmic recticulum. *J. Cell Biol.* **101**, 2374-2382.

Bevan, P. & Douglas, H.C. (1969). Genetic control of phosphoglucomutase variants in *Saccharomyces cerevisiae. J Bact.* **98**, 532-535.

Blake, N.M. & Omoto, K. (1975). Phosphoglucomutase types in the Asian-Pacific area: a critical review including new phenotypes. *Ann. Hum. Genet.* **38**, 251-273.

Boguski, M.S., Lowe, T.M.J. & Tolstoshev, C.M. (1993). dbEST - database for "expressed sequence tags". *Nature Genet.* **4**, 332-333.

Boles, E., Liebetrau, W., Hofmann, M. & Zimmermann, F.K. (1994). A family of hexosephosphate mutases in *Saccharomyces cerevisiae. Eur. J. Biochem.* **220**, 83-96.

Bovenberg, R.A.L., Adema, G.J., Jansz, H.S. & Bass, P.D. (1988). Model for tissue specific calcitonin/CGRP-1 RNA processing from *in vitro* experiments. *NAR.* **16**, 7867-7883.

Brautaset, T., Standal, R., Fjærvik, E. & Valla, S. (1994). Nucleotide sequence and expression analysis of the *Acetobacter xylinum* phosphoglucomutase gene. *Microbiology.* **140**, 1183-1188

Brown, P.K., Romana, L.K. & Reeves, P.R. (1992). Molecular analysis of the *rfb* gene cluster of *Salmonella* serovar muenchen (strain M67): the genetic basis of the polymorphism between groups C2 and B. *Mol. Microbiol.* **6**, 1385-1394.

Bruns, G.A.P. & Sherman, S.L. (1989). Report of the committee on the genetic constitution of chromosome 1. *Cyto. Cell Genet.* **51**, 67-90.

Cantu, J.M. & Ibarra, B. (1982). Phosphoglucomutase: evidence for a new locus expressed in human milk. *Science.* **216**, 639-640.

Carlson, D.M. (1966). Phosphoacetylglucosamine mutase from pig submaxillary gland. *Methods in Enzymology.* **8**, 179-182.

Carlson, T.A. & Chelm, B.K. (1986). Apparent eukaryotic origin of glutamine synthetase II from the bacterium *Bradyrhizobium japonicum. Nature.* **322**, 568-570.

Carrell, R.W., Aulak, K.s. & Owen, M.C. (1989). The molecular pathology of the serpins. *Mol. Biol. Med.* **6**, 35-42.

Carter, N.D., West, C.M., Emes, E., Parkin, B. & Marshall, W.H. (1979). Phosphoglucomutase polymorphism detected by isoelectric focusing: gene frequencies, evolution and linkage. *Ann. Hum. Biol.* **6**, 221-230.

Cheng, P.-W. & Carlson, D.M. (1979). Mechanism of phosphoacetylglucosamine mutase. *J. Biol. Chem.* **254**, 8353-8357.

Collins, F.S. (1995). Positional cloning moves from perditional to traditional. *Nature Genet.* **9**, 347-350.

Corbo, R.M., Palmarino, R., Spennati, G.F., Pascone, R. & Lucarelli, P. (1980). Human placental phosphoglucomutase locus 3 studies in the Italian population. *Jpn. J. Hum. Genet.* **25**, 325-328.

Cori, G.T. & Cori C.F. (1937). Formation of glucose-1-phosphoric acid in muscle extract. *Proc. Soc. Exp. Biol. Med.* **36**, 119-122.

Cori, G.T., Colowick, S.P. & Cori C.F. (1938a). The formation of glucose-1-phosphoric acid in extracts of mammalian tissues and of yeast. *J. Biol. Chem.* **123**, 375-380.

Cori, G.T., Colowick, S.P. & Cori C.F. (1938b). The enzymatic conversion of glucose-1-phosphoric ester to 6-ester in tissue extracts. *J. Biol. Chem.* **124**, 543-555.

Coyne, M.J., Russell, K.S., Coyle, C.L. & Goldberg, J.B. (1994). The *Pseudomonas areoginsoa algC* gene encodes phosphoglucomutase, required for the synthesis of a complete lipopolysaccharide core. *J. Bact.* **176**, 3500-3507.

Creighton, T.E. (1993). *Proteins: Structures and Molecular Properties.* 2nd Ed. W.H. Freeman and Company, New York.

Dai, J.B., Lui, Y., Ray, W.J. & Konno, M. (1992). The crystal structure of muscle phosphoglucomutase refined at 2.7 angstrom resolution. *J. Biol. Chem.* **267**, 6322-6337.

Dallas, W.S., Dev, I.K. & Ray, P.H. (1993). The dihydropteroate synthase gene, *folP*, is near the leucine tRNA gene, *leuU*, on the *Escherichia coli* chromosome. *J. Bact.* **175**, 7743-7744.

Dayhoff, M.O. (1978). *Atlas of Protein Sequence and Structure.* **5**, Suppl. 3. National Biomedical Research Foundation, Silver Springs, ND, USA.

De Ley, J., Mannheim, W., Mutters, R., Piechulla, K., Tytgat, R., Segers, P., Bisgaard, M., Frederiksen, W., Hinz, K.-H. & Vanhoucke, M. (1990). Inter- and intrafamilial similarities of rRNA cistrons of the *Pasteurellaceae. Int. J. Syst. Bact.* **40**, 126-137.

Don, R.H., Cox, P.T, Wainwright, B.J., Baker, K. & Mattick, J.S. (1991). 'Touchdown' PCR to circumvent spurious priming during gene amplification. *NAR.* **19**, 4008.

Doolittle, R.F. (1985). The genealogy of some recently evolved vertebrate proteins. *TIBS.* **10**, 233-237.

Doolittle, R.F., Feng, D.F., Anderson, K.L. & Alberro, M.R. (1990). A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. *J. Mol. Evol.* **31**, 383-388.

Doolittle, R.F. (1994). Convergent evolution: the need to be explicit. *TIBS.* **19** 15-18.

Dracopoli, N.C., Stanger, B.Z., Ito, C.Y., Call, K.M., Lincoln, S.E., Lander, E.S. & Housman, D.E. (1988). A genetic linkage map of 27 loci from PND to FY on the short arm of human chromosome 1. *Am. J. Hum. Genet.* **43**, 462-470.

Drago, G.A., Hopkinson, D.A., Westwood, S.A. & Whitehouse, D.B. (1991). Antigenic analysis of the major human phosphoglucomutase isozymes: PGM1, PGM2, PGM3 and PGM4. *Ann. Hum. Genet.* **55**, 263-271.

Drago, G.A. (1992). The development of ultrasensitive immunological methods for the detection of protein polymorphisms. PhD Thesis. (University of London).

Dykes, D.D., Kuhnl, P. & Martin, W. (1985). PGM1 system. Report on the international workshop, October 10-11, 1983, Munich, West Germany. *Am. J. Hum. Genet.* **37**, 1225-1231.

Endow, S.A. & Hatsumi, M. (1991). A multimember kinesin gene family in *Drosophila. PNAS USA.* **88**, 4424-4427.

Fernandez-Sorensen, A. & Carlson, D.M. (1971). Purification and properties of phosphoacetylglucosamine mutase. *J. Biol. Chem.* **246**, 3485-3493.

Fisher, R.A. & Harris, H. (1972). 'Secondary' isozymes derived from the three PGM loci. *Ann. Hum. Genet.* **36**, 69-77.

Fitch, W.M. & Margoliash, E. Construction of phylogenetic trees. *Science.* **155**, 279-284.

Fjærvik, E., Frydenlund, K., Valla, S., Huggirat, Y. & Benziman, M. (1991). Complementation of cellulose-negative mutants of *Acetobacter xylinum* by the cloned structural gene for phosphoglucomutase. *FEMS Microbiol. Lett.* **77**, 325-330.

Fong, A.M. & Santoro, S.A. (1994). Transcriptional regulation of α_{IIb} integrin gene expression during megakaryocytic differentiation of K562 cells. *J. Biol. Chemi* **269**, 18441-18447.

Fox, M., Tomkins, J., Whitehouse, D.B. & Parrington, J. (1996). Cytogenetic analysis of the erythroleukaemic cell line, K562, using fluorescence in situ hybridization (FISH) and chromosome-specific paint probes. *Eur. J. Hum. Genet.* 4, suppl.1, 4.033.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M.m Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.S., Merrick, J.M., Tomb. J.-F., Dougherty, B.A., Bott, K.F., Hu, P.-C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison, C.A. & Venter, J.C. (1995). The minimal gene complement of *Mycoplasma genitalium. Science.* **270**, 397-403.

Goldman, D., Goldin, L.R., Rathnagiri, P., O'Brien, S.J., Egeland, J.A. & Merril, C.R. (1985). Twenty-seven protein polymorphisms by two-dimensional electrophoresis of serum, erythrocytes and fibroblasts in two pedigrees. *Am. J. Hum. Genet.* **37**, 898-911.

Griffin, L.D., MacGregor, G.R., Muzny, D.M., Harter, J., Cook, R.G. & McCabe, E.R.B. (1988). Synthesis of hexokinase 1 (HK1) cDNA probes by mixed oligonucleotide primed amplification of cDNA (MOPAC) using primer mixtures of high complexity. *Am. J. Hum. Genet.* **43**, A185.

Gruskin, K.D., Smith, T.F. & Goodman, M. (1987). Possible origin of a calmodulin gene that lacks intervening sequences. *PNAS USA.* **84**, 1605-1608.

Hanabusa, K., Dougherty, H.W., del Rio, C., Hashimoto, T. & Handler, P. (1966). Phosphoglucomutase II: preparation and properties of phosphoglucomutases from *Micrococcus lysodeikticus* and *Bacillus cereus*. *J. Biol. Chem.* **241**, 3930-3939.

Hayes B.K. & Hart, G.W. (1994). Novel forms of protein glycosylation. *Curr. Opin. Struct. Biol.* **4**, 692-696.

Harding, N.E., Raffo, S., Raimondi, A., Cleary, J.M. & Ielpi, L. (1993). Identification, genetic and biochemical analysis of genes involved in synthesis of sugar nucleotide precursors of xanthan gum. *J. Gen. Microbiol.* **139**, 447-457.

Harris, H. & Hopkinson, D.A. (1976). *Handbook of Enzyme Electrophoresis in Human Genetics.* North Holland Publishing Company, Amsterdam.

Hashimoto, T. & Handler, P. (1966). Phosphoglucomutase III: purification and properties of phosphoglucomutases from flounder and shark. *J. Biol. Chem.* **241**, 3940-3948.

Hashimoto, T., Del Rio, C. & Handler, P. (1966). Comparative structure and function of phosphoglucomutase. *Fed. Proc.* **25**, 408.

He, M., Liu, H., Wang, Y. & Austen, B. (1992). Optimized centrifugation for rapid elution of DNA from agarose gels. *GATA.* **9**, 31-33.

Hofmann, M., Boles, E. & Zimmermann, F. (1994). Characterization of the essential yeast gene encoding N-acetylglucosamine-phosphate mutase. *Eur. J. Biochem.* **221**, 741-747.

Hollyoake, M., Putt, W., Edwards, Y.H. & Whitehouse, D.B. (1992). Two *TaqI* polymorphisms at the human PGM1 locus. *Hum. Mol. Genet.* **1**, 354.

Hopkinson, D.A. & Harris, H. (1965). Evidence for a second 'structural' locus determining human phosphoglucomutase. *Nature.* **208**, 410-412.

Hopkinson, D.A. & Harris, H. (1966). Rare phosphoglucomutase phenotypes. *Ann. Hum. Genet.* **30**, 167-178.

Hopkinson, D.A. & Harris, H. (1968). A third phosphoglucomutase locus in man. *Ann. Hum. Genet.* **31**, 359-367.

Houlgatte, R., Mariage-Samson, R., Duprat, S., Tessier, A., Bentolila, S., Lamy, B. & Auffray, C. (1995). The genexpress index: a resource for gene discovery and the genic map of the human genome. *Genome Res.* **5**, 272-304.

Ives, J. (1995). Structural studies of the PGM1 gene and a search for PGM3 and PGM4. PhD Thesis. (University of London).

Jagannathan, V. & Luck, J.M. (1949). Phosphoglucomutase II: mechanism of action. *J. Biol. Chem.* **179**, 569-575.

Jayaratne, P., Bronner, D., MacLachlan, P.R., Dodgson, C., Kido, N. & Whitfield, C. (1994). Cloning and analysis of duplicated *rfbM* and *rfbK* genes involved in the formation of GDP-mannose in *Escherichia coli* 09:K30 and participation of the group I K30 capsular polysaccharide. *J. Bact.* **176**, 3126-3139.

Jiang, X-M., Neal, B., Santiago, F., Lee, S.J., Romana, L.K. and Reeves, P.R. (1991). Structure and sequence of the *rfb* (O antigen) gene cluster of *Salmonella* serovar typhimurium (strain LT2). *Mol. Microbiol.* **5**, 695-713.

Jones, D.S.C. & Schofield, J.P. (1990). A rapid method for isolating high quality plasmid DNA suitable for DNA sequencing. *NAR.* **18**, 7463-7464.

Jongsma, A., van Someren, H., Westerveld, A., Hagemeijer, A. & Pearson, P. (1973). Localization of genes on human chromosomes by studies of human-chinese hamster somatic cell hybrids: assignment of PGM3 to chromosome C6 and regional mapping of the PGD, PGM1 and Pep-C genes of chromosome A1. *Humangenetik.* **20**. 195-202.

Johnson, D.E. & Williams, L.T. (1993). Structural and functional diversity in the FGF receptor multigene family. *Adv. Cancer. Res.* **60**, 1-41.

Joshi, J.G., & Handler, P. (1964). Phosphoglucomutase I: purification and properties of phosphoglucomutase from *Escherichia coli.* *J. Biol. Chem.* **239**, 2741-2751.

Joshi, J.G., Hooper, J., Kuwaki, T., Sakurada, T., Swanson, J.R. & Handler, P. (1967). Phosphoglucomutase V: multiple forms of phosphoglucomutase. *PNAS USA.* **57**, 1482-1489.

Karlsson, S., Swallow, D.M., Griffiths, B., Corney, G., Hopkinson, D.A., Dawnay, A. & Cartron, J.P. (1983). A genetic polymorphism of a human urinary mucin. *Ann. Hum. Genet.* **47**, 263-269.

Kepes, F. & Schekman, R. (1988). The yeast *sec53* gene encocdes phosphomannomutase. *J. Biol. Chem.* **263**, 9155-9161.

Kim, D.H. & Ikemoto, N. (1986). Involvement of 60-kilodalton phosphoprotein in the regulation of calcium release from skeletal muscle sarcoplasmic recticulum. *J. Biol. Chem.* **261**, 11674-11679.

Klein, J., Satta, Y., O'hUigin, C. & Takahata, N. (1993). The molecular descent of the major histocompatibility complex. *Ann. Rev. Immunol.* **11**, 269-295.

Koplin, R., Arnold, W., Hotte, B., Simon, R., Wang, G.E. & Puhler, A. (1992). Genetics of xanthan production in *Xanthamonas campestris*: the *xanA* and *xanB* genes are involved in UDP-glucose and GDP-mannose biosynthesis. *J. Bact.* **174**, 191-199.

Koro, L.A. & Marchase, R.B. (1982). A UDP-glucose:glycoprotein glucose-1-phosphotransferase in embryonic chicken neural retina. *Cell.* **31**, 739-748.

Kuhnl, P., Schmidtmann, U. & Spielmann, W. (1977). Evidence for two additonal common alleles at the PGM1 locus (Phosphoglucomutase - E.C.: 2.7.5.1). *Hum. Genet.* **35**, 219-223.

Labigne, A., Cussac, V. & Courcoux, P. (1991). Shuttle cloning and nucleotide sequences of *Helicobacter pylori* genes responsible for urease activity. *J. Bact.* **173**, 1920-1931.

Lalley, P.A., Francke, U. & Minna, J.D. (1978). Homologous genes for enolase, phosphogluconate dehydrogenase, phosphoglucomutase and adenylate kinase are syntenic on mouse chromosome 4 and human chromosome 1p. *PNAS USA.* **75**, 2382-2386.

Lamm, L.U., Kissmeyer-Nielsen, F. & Henningsen, K. (1970). Linkage and association studies of two phosphoglucomutase loci (PGM1 and PGM3) to eighteen other markers. Analysis of the segregation at the marker loci. *Hum. Hered.* **20**, 305-318.

Lamm, L.U., Jorgensen, F. & Kissmeyer-Nielsen, F. (1981). On the mapping of PGM3 in relation to HLA. *Tissue Antigens.* **17**, 245-246.

Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature.* **227**, 680-685.

Lee, C.C., Wu, X., Gibbs, R.A., Cook, R.G., Muzny, D.M. & Caskey, C.T. (1988). Generation of cDNA probes directed by amino acid sequence: cloning of urate oxidase. *Science.* **239**, 1288-1291.

Lee, C.C. & Caskey, C.T. (1990). cDNA cloning using degenerate primers, in *PCR Protocols: a guide to methods and applications.* Academic Press, New York. 46-53.

Lee, S.J., Romana, L.K. & Reeves, P.R. (1992b). Cloning and structure of group C1 O antigen (*rfb* gene cluster) from *Salmonella enterica* serovar *montevideo. J. Gen. Microbiol.* **138**, 305-312.

Lee, S.Y., Marks, A.R., Gureckas, N., Lacro, R., Nadal-Ginard, B. & Kim, D.H. (1992a). Purification, characterization and molecular cloning of a 60 kDa phosphoprotein in rabbit skeletal sarcoplasmic recticulum which is an isoform of phosphoglucomutase. *J. Biol. Chem.* **267**, 21080-21088.

Leloir, L.F., Trucco, R.E., Cardini, C.E., Paladini, A. & Caputto, R. (1948). The coenzyme of phosphoglucomutase. *Arch. Biochem. Biophys.* **19**, 339-340.

Logan, N.A. (1994). *Bacterial Systematics.* Blackwell Scientific Publications, Oxford.

Lowe, N., Brady, J.M., Barlow, J.H., Sowden, J.C., Edwards, M. & Butterworth, P.H.W. (1990). Structure and methylation patterns of the gene encoding human carbonic anhydrase I. *Gene.* **93**, 277-283.

Lozzio, C.B. & Lozzio, B.B. (1975). Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood.* **45,** 321-334.

Lu, M. & Kleckner, N. (1994). Molecular cloning and characterization of the *pgm* gene encoding phosphoglucomutase of *Escherichia coli*. *J. Bact.* **176**, 5847-5851.

Maeda, N. & Smithies, O. (1986). The evolution of multigene families: human haptoglobin genes. *Ann. Rev. Genet.* **20**, 81-108.

March, R.E., Putt, W., Hollyoake, M., Ives, J.H., Lovegrove, J.U., Hopkinson, D.A., Edwards, Y.H, & Whitehouse, D.B. (1993a). The classical human phosphoglucomutase (PGM1) isozyme polymorphism is generated by intragenic recombination. *PNAS USA.* **90**, 10730-10733.

March, R.E., Hollyoake, M., Putt, W., Hopkinson, D.A., Edwards, Y.H. & Whitehouse, D.B. (1993b). Genetic polymorphism in the 3' untranslated region of human phosphoglucomutase 1. *Ann. Hum. Genet.* **57**, 1-8.

Marchase, R.B., Saunders, A.M., Rivera, A.A. & Cook, J.M. (1987). The β-phosphoro[^{35}S]thioate analogue of UDP-Glc is efficiently utilized by the glucose phosphotransferase and is relatively resistent to hydrolytic degradation. *Biochim. Biophys. Acta.* **916**, 157-162.

Marchase, R.B., Bounelis, P., Brumley, L.M., Dey, N., Browne, B., Auger, D., Fritz, T.A., Kulesza, P. & Bedwell, D.M. (1993). Phosphoglucomutase in *Saccharomyces cerevisiae* is a cytoplasmic glycoprotein and the acceptor for a Glc-phosphotransferase. *J. Biol. Chem.* **268**, 8341-8349.

Marchuk, D., Drumm, M., Saulino, A. & Collins, F.S. (1991). Construction of T-vectors, a rapid and general system for direct cloning of unmodified PCR products. *NAR.* **19**, 1154.

Marenah, C.B. (1973). An investigation of the biochemical properties of the human phosphoglucomutase isozymes determined by the PGM1, PGM2 and PGM3 loci. PhD Thesis. (University of London).

Marie, J.P., Izaguirre, C.A., Civin, C.I., Mirro, J. & McCulloch, E.A. (1981). The presence within single K562 cells of erythropoietic and granulopoietic differentiation markers. *Blood.* **58**, 708-711.

Marolda C.L. & Valvano, M.A. (1993). Identification, expression and DNA sequence of the GDP-mannose biosynthesis genes encoded by the 07 *rfb* gene cluster of strain VW187 (*Escherichia coli* 07:K1). *J. Bact.* **175**, 148-158.

Martinsson, T., Bjursell, C., Stibler, H., Kristiansson, B., Skovby, F., Jaeken, J., Blennow G., Stromme, P., Hanefeld, F. & Wahlstrom, J. (1994). Linkage of a locus for carbohydrate-deficient glycoprotein syndrome type I (CDG1) to chromosome 16p, and linkage disequilibrium to microsatellite marker D16S406. *Hum. Mol. Genet.* **3**, 2037-2042.

McAlpine, P.J., Hopkinson, D.A. & Harris, H. (1970a). The relative activities attributable to three phosphoglucomutase loci (*PGM1, PGM2, PGM3*) in human tissues. *Ann. Hum. Genet.* **34**, 169-173.

McAlpine, P.J., Hopkinson, D.A. & Harris, H. (1970b). Thermostability studies on the isozymes of human phophoglucomutase. *Ann. Hum. Genet.* **34**, 61-71.

McAlpine, P.J., Hopkinson, D.A. & Harris, H. (1970c). Molecular size estimates of the human phosphoglucomutase isozymes by gel filtration chromatography. *Ann. Hum. Genet.* **34**, 177-185.

McAlpine, P.J., Mohandas, T. & Hamerton, J.L. (1975). Isozyme analysis of somatic cell hybrids: assignment of the phosphoglucomutase 2 (*PGM2*) gene locus to chromosome 4 in man with data on the molecular structure and human chromosome assignments of six additional markers. *Isozymes IV: Genetics and Evolution.* Academic Press, New York. 149-167.

McAlpine, P.J., Stranc, L.C., Boucheix C. & Shows, T.B. (1990). The 1990 catalogue of mapped genes and report of the nomenculture committee. Human gene mapping 10.5 (1990): update to the tenth international workshop on human gene mapping. *Cyto. Cell Genet.* **55**, 5-76.

McCoy, E.E. & Najjar, V.A. (1959). The purification and mechanism of action of yeast phosphoglucomutase. *J. Biol. Chem.* **234**, 3017-3021.

Meera Khan, P., Hagemeijer, A., Wijnen, L.M.M. & v.d.Goes, R.G.M. (1984). PGM and ME1 are probably in the 6pter-q/12 region. *Cyto. Cell Genet.* **37**, 537.

Meyer, A.D., Ichikawa, T. & Meins, F. (1995). Horizontal gene transfer: regulated expression of a tobacco homologue of the *Agrobacterium rhizogenes rolC* gene. *Mol. Gen. Genet.* **249**, 265-273.

Milstein, C. (1961). The amino acid sequence around serine phosphate in phosphoglucomutase from different origins. *Biochem. J.* **79**, 26P.

Milstein, C. & Sanger, F. (1961). An amino acid sequence in the active centre of phosphoglucomutase. *Biochem. J.* **79**, 456-469.

Milstein, C.P. & Milstein, C. (1968). A tryptic peptide containing a unique serine phosphate residue in rabbit phosphoglucomutase. *Biochem. J.* **109**, 93-99.

Moiseeva, E.P., Belkin, A.M., Spurr, N.K., Koteliansky, V.E. & Critchely, D.R. (1996) A novel dystrophin/utrophin-associated protein is an enzymatically inactive member of the phosphoglucomutase superfamily. *Eur. J. Biochem.* **235**, 103-113.

Monn, E. (1969). Chromatographic studies on human red cell phosphoglucomutase. *Int. J. Protein Res.* **1**, 73-80.

Muhlbach, H. & Schnarrenberger, C. (1978). Properties and intracellular distribution of two phosphoglucomutases from spinach leaves. *Planta.* **141**, 65-70.

Nadeau, J.H., Kompf, J., Siebert, G. & Taylor, B.A. (1981). Linkage of PGM3 in the house mouse and homologies of three phosphoglucomutase loci in mouse and man. *Biochem. Genet.* **19**, 465-474.

Najjar, V.A. & Pullman, M.E. (1954). The occurence of a group transfer involving enzyme (phosphoglucomutase) and substrate. *Science.* **119**, 631-634.

Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K. & Sekiya, T. (1989a). Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *PNAS USA.* **86**, 2766-2770.

Orita, M., Suzuki, Y., Sekiya, T. & Hayashi, K. (1989b). Rapid and sensitive detecton of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics.* **5**, 874-879.

Penger, A., Pelzer-Reith, B. & Schnarrenberger, C. (1994). cDNA sequence for the plastidic phosphoglucomutase from *Spinacia oleracea* (L). *Plant Physiol.* **105**, 1439-1440.

Piatigorsky, J. & Wistow, G. (1991). The recruitment of crystallins: new functions precede gene duplication. *Science.* **252**, 1078-1079.

Povey, S., Jeremiah, S., Arthur, E., Steel, M. & Klein, G. (1980). Differences in genetic stability between human cell lines from patients with and without lymphoreticular malignancy. *Ann. Hum. Genet.* **44**, 119-133.

Putt, W., Ives, J.H., Hollyoake, M., Hopkinson, D.A., Whitehouse, D.B. & Edwards, Y.E. (1993). Phosphoglucomutase 1: a gene with two promoters and a duplicated first exon. *Biochem. J.* **296**, 417-422.

Quick, C.B., Fisher, R.A. & Harris, H. (1972). Differntiation of the *PGM2* locus isozymes from those of *PGM1* and *PGM3* in terms of phosphopentomutase activity. *Ann. Hum. Genet.* **35**, 445-454.

Ravazzolo, R., Sessarego, M., Barresi, R., Garre, C., Scarra, G.B. & Ajmar, F. (1985). Demonstration of phosphoglucomutase 1 in a subclone of the K562 cell line. *Cancer Res.* **45**, 1296-1299.

Ray, W.J. & Peck, E.J. (1972). Phosphomutases in *The Enzymes.* VI, 3rd Edition. Ed Boyer, P.D. Academic Press, New York & London. 407-477.

Ray, W.J., Hermodson, M.A., Puvathingal, K.M. & Mahoney, W.C. (1983). The complete amino acid sequence of rabbit muscle phosphoglucomutase. *J. Biol. Chem.* **258**, 9166-9174.

Reeves, P. (1993). Evolution of *Salmonella* O antigen variation by interspecific gene transfer on a large scale. *TIGS.* **9**, 17-22.

Rivera, A.A., Elton, T.S., Dey, N.B., Bounelis, P. & Marchase, R.B. (1993). Isolation and expression of a rat liver cDNA encoding phosphoglucomutase. *Gene.* **133**, 261-266.

Ruddle, F., Ricciuti, F., McMorris, F.A., Tischfield, J., Creagan, R., Darlington, G. & Chen, T. (1972). Somatic cell genetic assignment of peptidase C and the Rh linkage group to chromosme A-1 in man. *Science.* **176**, 1429-1431.

Saitou, N. & Nei, M. (1987). The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425.

Salvucci, M.E., Drake, R.R., Broadbent, K.P., Haley, B.E., Hanson, K.R. & McHale, N.A. (1990). Identification of the 64 kilodalton chloroplast stromal phosphoprotein as phosphoglucomutase. *Plant Physiol.* **93**, 105-109.

Sambrook, J., Fritsch, E.F. & Maniatis, T. (1989). *Molecular cloning: A laboratory manual.* 2nd Ed. Coldspring Harbour Laboratory Press, Plainview, New York.

Sandlin, R.C. & Stein, D.C. (1994). Role of phosphoglucomutase in lipooligosaccharide biosynthesis in *Neisseria gonorrhoeae. J. Bact.* **176**, 2930-2937.

Sanger, F., Nicklen, S. & Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *PNAS USA.* **74**, 5463-5467.

Santachiara, A.S.B. (1969). Ultracentrifuge studies of red cell phosphoglucomutase. *Nature.* **223**, 625-626.

Satir, B.H., Reichman, M., Srisomsap, C. & Marchase, R.B. (1988). Parafusin, a stimulus sensitive phosphoprotein from *Paramecium* is a substrate for glucosephosphotransferase. *J. Cell. Biol.* **107**, 404a.

Satir, B.H., Srisomsap, C., Reichman, M. & Marchase, R.B. (1990). Parafusin, an exocytic-sensitive phosphoprotein, is the primary acceptor for the glucosylphosphotransferase in *Paramecium tetraurelia* and rat liver. *J. Cell. Biol.* **111**, 901-907.

Scharf, S.J., Horn, G.T. & Erlich, H.A. (1986). Direct cloning and sequence analysis of enzymatically amplified genomic sequences. *Science.* **233**, 1076-1078.

Schmitt, J., Lichte, K.H. & Fuhrmann, W. (1970). Red cell enzymes of the Pongidae. *Humangenetik.* **10**, 138-144.

Selden, J.R., Emanuel, B.S., Wang, E., Cannizzaro, L., Palumbo, A., Erikson, J., Nowell, P.C., Rovera, G. & Croce, C.M. (1983). Amplified C_λ and c-abl genes are on the same marker chromosome in K562 leukaemia cells. *PNAS USA.* **80**, 7289-7292.

Singh, R. and Green, M.R. (1993). Sequence-specific binding of transfer RNA by glyceraldehyde-3-phosphate dehydrogenase. *Science.* **259**, 365-368.

Shatters, R.G. & Kahn, M.L. (1989). Glutamine synthetase II in *Rhizobium*: reexamination of the proposed horizontal transfer of DNA from eukaryotes to prokaryotes. *J. Mol. Evol.* **29**, 422-428.

Sheppard, P.O., Grant, F.J., Oort, P.J., Sprecher, C.A., Foster, D.A., Hagen, F.S., Upshall, A., McKnight, G.L. & O'Hara, P.J. (1994). The use of conserved cellulase family-specific sequences to clone cellulase homologue cDNAs from *Fusarium oxysporum. Gene.* **150**, 163-167.

Smith, D.J., Cooper, M., DeTiani, M., Losberger, C. & Payton, M.A. (1992). The *Candida albicans PMM1* gene encoding phosphomannomutase complements a *Saccharomyces cerevisiae sec53-6* mutation. *Curr. Genet.* **22**, 501-503.

Smith, M.W. & Doolittle, R.F. (1992). A comparison of evolutionary rates of the two major kinds of superoxide dismutase. *J. Mol. Evol.* **34**, 175-184.

Smith, M.W., Feng, D.F. & Doolittle, R.F. (1992) Evolution by acquisition: the case for horizontal gene transfers. *TIBS.* **17**, 489-493.

Spencer, N., Hopkinson, D.A. & Harris, H. (1964). Phosphoglucomutase polymorphism in man. *Nature* . **204**, 742-745.

Sprague, G.F. (1991). Genetic exchange between kingdoms. *Curr. Opin. Genet. Dev.* **1**, 530-533.

Srisomsap, C., Richardson, K.L., Jay, J.C. & Marchase, R.B. (1988). Localization of the glucose phosphotransferase to a cytoplasmically accessible site on intracellular membranes. *J. Biol. Chem.* **263**, 17792-17797.

Stevenson, G., Lee, S.J., Romana, L.K. & Reeves, P.R. (1991). The *cps* gene cluster of *Salmonella* strain LT2 includes a second mannose pathway: sequence of two genes and relationship to genes in the *rbf* gene cluster. *Mol. Gen. Genet.* **227**, 173-180.

Subramanian, S.V., Wyroba, E., Andersen, A.P. & Satir, B.H. (1994). Cloning and sequencing of parafusin, a calcium-dependent exocytosis-related phosphoglycoprotein. *PNAS USA.* **91**, 9832-9836.

Sugiyama, T., Kido, N., Komatsu, T., Ohta, M., Jann, K., Jann, B., Saeki, A. & Kato, N. (1994). Genetic analysis of *Escherichia coli* 09 *rfb*: identification and DNA sequence of phosphomannomutase and GDP-mannose pyrophosphorylase genes. *Microbiology.* **140**, 59-71.

Swofford, D.L. (1980). PAUP: phylogenetic analysis using parsimony. Illinois Natural History Survey, Champaign, Illinois, USA.

Syvanen, M. (1994). Horizontal gene transfer: evidence and possible consequences. *Ann. Rev. Genet.* **28**, 237-261.

Takahashi, N., Neel, J.V., Satoh, C., Nishizaki, J. & Masunari, N. (1982). A phylogeny for the principal alleles of the human phosphoglucomutase-1 locus. *PNAS USA* . **79**, 6636-6640.

Takahashi, N. & Neel, J.V. (1993). Intragenic recombination at the human phosphoglucomutase 1 locus: predictions fulfilled. *PNAS USA.* **90**, 10725-10729.

Tashian, R.E. (1989). The carbonic anhydrases: widening perspectives on their evolution, expression and function. *BioEssays.* **10**, 186-192.

Tham, T.N., Ferris, S., Kovacic, R., Montagnier, L. & Blanchard, A. (1993). Identification of *Mycoplasma pirum* genes involved in the salvage pathways for nucleosides. *J. Bact.* **175**, 5281-5285.

Treptau, T., Kissmehl, R., Wissmann, J.-D. & Plattner, H. (1995). A 63 kDA phosphoprotein undergoing rapid dephosphorylation during exocytosis in *Paramecium* cells shares biochemical characteristics with phosphoglucomutase. *Biochem. J.* **309**, 557-567.

Tsoi, A. & Douglas, H.C. (1964). The effect of mutation on two forms of phosphoglucomutase in *Saccharomyces*. *Biochim. Biophys. Acta.* **92**, 513-520.

Uttaro, A.D., Cangelosi, G.A., Geremia, R.A., Nester, E.W. & Ugalde, R.A. (1990). Biochemical characterization of avirulent *exoC* mutants of *Agrobacterium tumefaciens*. *J. Bact.* **172**, 1640-1646.

Uttaro, A.D. & Ugalde, R.A. (1994). A chromosomal cluster of genes encoding ADP-glucose synthetase, glycogen synthase and phosphoglucomutase in *Agrobacterium tumefaciens*. *Gene.* **150**, 117-122.

Uttaro, A.D. & Ugalde, R.A. (1995). A chromosomal cluster of genes encoding ADP-glucose synthetase, glycogen synthase and phosphoglucomutase in *Agrobacterium tumefaciens*. *Gene.* **151**, 141-143.

Van Schaftingen, E. & Jaeken, J. (1995). Phosphomannomutase deficiency is a cause of carbohydrate-deficient glycoprotein syndrome type I. *FEBS Letters.* **377**, 318-320.

van Someren, H., Westerveld, A, Hagemeijer, A., Mees, J.R., Meera Khan, P. & Zaalberg, O.B. (1974). Human antigen and enzyme markers in man-chinese hamster somatic cell hybrids: evidence for synteny between the *HL-A, PGM3, ME-1* and *IPO-B* loci. *PNAS USA.* **71**, 962-965.

Verma, N.K., Quigley, N.B. & Reeves, P.R. (1988). O-Antigen variation in *Salmonella* spp.: *rfb* gene clusters of three strains. *J. Bact.* **170**, 103-107.

Veyna, N.A., Jay, J.C., Srisomsap, C., Bounelis, P. & Marchase, R.B. (1994). The addition of glucose-1-phosphate to the cytoplamic glycoprotein phosphoglucomutase is modulated by intracellular calcium in PC12 cells and rat cortical synaptosomes. *J. Neurochem.* **62**, 456-464.

Wakabayashi, S.W., Matsubara, H. & Webster, D.A. (1986). Primary sequence of a dimeric bacterial haemoglobin from *Vitreoscilla*. *Nature.* **322**, 481-482.

Wang, L., Romana, L.K. & Reeves, P.R. (1992). Molecular analysis of a *Salmonella enterica* group E1 *rfb* gene cluster: O antigen and the genetic basis of the major polymorphism. *Genetics.* **130**, 429-443.

Ward, L.J., Elston, R.C., Keats, B.J.B. & Graham, J.B. (1985). *PGM1* null allele detected in a caucasian mother-son pair. *Hum. Hered.* **35**, 178-181.

Whitehouse, D.B., Putt, W., Lovegrove, J.U., Morrison, K., Hollyoake, M., Fox, M.F., Hopkinson, D.A. & Edwards Y.H. (1992). Phosphoglucomutase 1: complete human and rabbit mRNA sequences and direct mapping of this highly polymorphic marker on human chromosome 1. *PNAS USA.* **89**, 411-415.

Wilks, A.F. (1989). Two putative protein-tyrosine kinases identified by application of the polymerase chain reaction. *PNAS USA.* **86**, 1603-1607.

Wistow, G.J. & Piatigorsky, J. (1990). Gene conversion and splice -site slippage in the argininosuccinate lyases/δ-crystallins of the duck lens: members of a enzyme superfamily. *Gene.* **96**, 263-270.

Wyroba, E., Hoyer, A.W., Storgaard, P. & Satir, B.H. (1995). Mammalian homologue of the calcium-sensitive phosphoglycoprotein, parafusin. *Eur. J. Cell. Biol.* **68**, 419-426.

Ye, R.W., Zielenski, N.A. & Chakrabarty, A.M. (1994). Purification and characterization of phosphoglucomutase from *Pseudomonas aeroginosa* involved in biosynthesis of both alginate and lipopolysaccharide. *J. Bact.* **176**, 4851-4857.

Zhou, D., Stephens, D.S., Gibson, B.W., Engstrom, J.J., McAllister, C.F., Lee, F.K.N. & Apicella, M.A. (1994). Lipooligocaccharide biosynthesis in pathogenic *Neisseria. J. Biol. Chem.* **269**, 11162-11169.

Zielenski, N.A., Chakrabarty, A.M. & Berry, A. (1991). Characterization and regulation of the *Pseudomonas aeruginosa algC* gene encoding phosphomannomutase. *J. Biol. Chem.* 266, 9754-9763.