Molecular evolution of the human *Pgk-2* retroposon

John R. McCarrey

Division of Reproductive Biology, The Johns Hopkins University, School of Hygiene and Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA

Received October 9, 1989; Revised and Accepted November 30, 1989

ABSTRACT

The human phosphoglycerate kinase (*Pgk*) gene family includes the functional, intronless *Pgk-2* gene and the intronless ψ *hPgk-1* pseudogene, both of which are retroposons of the intron-containing *Pgk-1* gene. The divergence of the *Pgk-2* retroposon from *Pgk-1* is compared with that of the ψ *hPgk-1* retroposon from *Pgk-1* to reveal nucleotide characteristics diagnostic of functional genes. A comparison of the human and mouse *Pgk* genes indicates that *Pgk-2* has evolved more rapidly than *Pgk-1* since the two genes diverged early in mammalian evolution, but that the lack of introns in *Pgk-2* may have diminished inter-exon variation. The hypothesis that codon bias is related to expression level is shown not to hold for the *Pgk* genes; however, the idea that a deficiency of TA and CG dinucleotides and an excess of TG and CT dinucleotides contributes to codon bias is supported. Finally, the hypothesis that the *Pgk-2* retroposon initially included a copy of the *Pgk-1* 'housekeeping' promoter and subsequently evolved a tissue-specific promoter is examined and supported. It is concluded that this process involved the loss of the 5' CpG island present in the *Pgk-1* gene, and that selection for cell type-specific expression of *Pgk-2* at high levels has driven the divergence of this retroposon from its progenitor, *Pgk-1*.

INTRODUCTION

Previous studies of human and mouse phosphoglycerate kinase (*Pgk*) genes indicate the mammalian *Pgk* gene family includes two functional loci, of which one (*Pgk-1*) is intron-containing and ubiquitously expressed, and the other (*Pgk-2*) is intronless and expressed only in spermatogenic cells (1–4). The *Pgk-2* gene is a retroposon that arose by reverse transcriptase-mediated processing of a transcript from the *Pgk-1* gene (1,3,5). In the human genome an X-linked intronless pseudogene (ψ *hPgk-1*) has also been identified as a retroposon of the functional *Pgk-1* locus (6). The human *Pgk* gene family provides a novel system in which to compare the molecular evolution of genes that have diverged from a common progenitor and produce a functionally similar protein product, yet differ in whether or not they contain introns, whether or not they are expressed, and/or whether they are expressed at high or low levels.

The *Pgk-1* gene is expressed at low levels in all mammalian cells that contain a transcriptionally active X chromosome,

including somatic cells, oogenic cells and premeiotic spermatogenic cells (7–9). In eutherian mammals, the *Pgk-2* gene is specifically expressed at high levels in the one metabolically active cell type that lacks an active X chromosome, the meiotic/postmeiotic spermatogenic cell (1,8). The protein products of the two *Pgk* genes are functionally and structurally similar [10 & unpublished results (JRM)]. Thus these two genes present a natural experiment in which the effects of the presence or absence of introns on the divergence of closely related, functional gene sequences can be directly compared.

The ψ *hPgk-1* pseudogene is not expressed and has evolved stop codons within the putative coding sequence (6). Thus the functional *Pgk-2* and non-functional ψ *hPgk-1* retroposons originated by the same retroposition mechanism (Figure 1) and initially carried the same coding sequence, but have subsequently diverged from their common progenitor (*Pgk-1*) in the presence or absence, respectively, of selection at the protein level. A comparison of these sequences provides a direct test for characteristics at the nucleotide level that are diagnostic of divergence of functional versus non-functional genes.

The *Pgk-2* gene originated during mammalian evolution prior to the divergence between eutherians and metatherians, approximately 125 million years (Myr) ago (7). Nucleotide sequence characteristics indicate the *Pgk-2* retroposon included a copy of the endogenous *Pgk-1* 5'-regulatory sequences, which facilitated the initial expression of this retroposon (5). This implies that the *Pgk-2* gene initially carried a 'housekeeping' promoter sequence that directed ubiquitous low-level transcription, and has subsequently evolved a cell type-specific promoter that directs high-level transcription. Thus this system provides a model for the evolution of cell type-specific regulation of a gene. The analysis presented here suggests that selection for high-level, cell type-specific expression in spermatogenic cells has been a primary influence directing the divergence of the *Pgk-2* gene sequence from that of *Pgk-1*.

METHODS

Nucleotide sequence information for the mammalian *Pgk* genes has been previously published: human *Pgk-1* (*hPgk-1*) (2,11–13), human *Pgk-2* (*hPgk-2*) (1,5), human *Pgk-1* pseudogene (ψ *hPgk-1*) (6), mouse *Pgk-1* (*mPgk-1*) (4,14), mouse *Pgk-2* (*mPgk-2*) (3). For the analyses reported here, this sequence information was stored, manipulated and compared using the MicroGenie sequence analysis program, version 6.0 (Beckman). This program was also used for homology determinations, codon

FIGURE 1. The origin of human *Pgk* retroposons.

The proposed origin of the *Pgk*-2 (C) and ψ h*Pgk*-1 (B) intronless retroposons by gene processing from the intron-containing *Pgk*-1 gene (A) is schematically represented. The non-functional ψ h*Pgk*-1 pseudogene is shown originating by reverse transcriptase-mediated processing of a 'normal' spliced, polyadenylated transcript of *Pgk*-1 (6). This transcript initiated at the normal start point (solid arrowhead) in *Pgk*-1. The functional *Pgk*-2 gene is shown originating by similar processing of an aberrant transcript of *Pgk*-1 that initiated at an upstream start point (open arrowhead) and included the endogenous *Pgk*-1 5'-promoter sequence (5). (Boxes represent exon sequences; hnRNA—heteronuclear RNA; mRNA—messenger RNA).

usage summaries, determination of mononucleotide and dinucleotide frequencies, and G+C and A+T-richness assessments.

Statistical analyses of sequence information included a standard chi square test to estimate the significance of inter-exon variation within each *Pgk* gene (Table IV). The exon borders for the *hPgk*-1 gene have been established by Michelson et al. (2), and analogous 'exon regions' noted in the ψ h*Pgk*-1 (6) and *hPgk*-2 (1) retroposons. Chi square calculations were based on the actual number of similar and dissimilar nucleotides or amino acids, rather than the corresponding homology percentages.

A 'scaled chi square' analysis analogous to that described by Shields et al. (15) was used to calculate codon bias in each *Pgk* gene. A standard chi square statistic was calculated for variation within each set of synonymous codons, and that statistic was divided by the number of occurrences of the corresponding amino acid in each *Pgk* gene. Chi square calculations were based on the actual numbers of codons used in each case, rather than their relative frequency of use. The scaled chi square statistics were

then compared using a standard sign test to determine if a significant difference in overall codon bias could be demonstrated among the *Pgk* genes. This analysis was designed to directly test the hypothesis that genes expressed at higher levels should exhibit a greater codon bias, thus a one-tailed probability test was applied.

RESULTS & DISCUSSION

Sequence characteristics of functional versus non-functional genes

A comparison of the *hPgk*-2 and ψ h*Pgk*-1 retroposon sequences with that of their common progenitor locus, *hPgk*-1, is shown in Table I. The X-linked ψ h*Pgk*-1 retroposon apparently arose subsequent to the divergence between humans and mice, since there is no X-linked pseudogene of *Pgk* in the murine genome (16,17), and the difference between the *hPgk*-1 and ψ h*Pgk*-1 nucleotide sequences is only 6%. Nevertheless, the ψ h*Pgk*-1 locus has already evolved two stop codons within the putative coding sequence that would preclude its encoding a functional

TABLE I: DIVERGENCE OF HUMAN *Pgk* RETROPOSON CODING SEQUENCES FROM *hPgk-1*

	Open Reading Frame	XHomology with <i>hPgk-1</i>		XSubstitutions at codon position #3	XSynonymous/ nonsynonymous nucleotide substitutions	XTransitions/ XTransversions
		N.A. ¹	A.A. ²			
ψ <i>hPgk-1</i>	No	94	90	45	30.8/69.2	26.9/73.1
<i>hPgk-2</i>	Yes	85	87	70	69.5/30.5	39.1/60.9

¹N.A. - nucleic acid²A.A. - amino acidTABLE II: HUMAN *Pgk* GENE CODING/UNTRANSLATED SEQUENCE HOMOLOGIES

Gene	Region				XHomology with <i>hPgk-1</i> sequence
<i>hPgk-1</i>	5'-UT ¹	-60	CGGCAGTCGGCTCCCTCGTTGACCGAATCACCGACCTCTCTCCCAGCTGTATTTCCAAA	-1	--
ψ <i>hPgk-1</i>	5'-UT	-60	.A....C.....TAC..T.T.....T.....-----.....	-1	87
<i>hPgk-2</i>	5'-UT	-60	A.CA...T..T..TTCA.CATTAA..TC..GGTGT.AGC..ATGTCTT.A....CT...G	-1	40
<i>hPgk-1</i>	5'-CS ²	+1	ATGTCGCTTTCTAACAAGCTGACGCTGGACAAGCTGGACGTTAAAGGGAAGCGGGTCGTT	60	--
ψ <i>hPgk-1</i>	5'-CS	+1	A.....AT.....A.CA.....C.....C.....A..	60	87
<i>hPgk-2</i>	5'-CS	+1	T.....G...T...TT.A....A....T...G.....A...A.C	60	80
<i>hPgk-1</i>	3'-CS	1195	AGTTTGGAGCTCCTGGAAGGTAAAGTCCTTCCTGGGGTGGATGCTCTCAGCAATATTTAG	1254	--
ψ <i>hPgk-1</i>	3'-CS	1195	A.....CA.....G.....G.....	1254	93
<i>hPgk-2</i>	3'-CS	1195	...C.A....T.....A.....A...A...G...C.....C...G...G...	1254	83
<i>hPgk-1</i>	3'-UT	1255	TACTTTCTGCCTTTTAGTTCCTGTGCACAGCCCTAAGTCAACTTAGCATTCTTCTGCAT	1315	--
ψ <i>hPgk-1</i>	3'-UT	1255	T.....	1315	98
<i>hPgk-2</i>	3'-UT	1255	.TAA.ATAGTGT.AC.TCC.T....TTT.T.T..A.GGCC.TTAAGTCAGC..AA...T.	1315	35

¹UT - untranslated sequence.²CS - coding sequence.

phosphoglycerate kinase (6). The *hPgk-2* nucleotide sequence differs from that of *hPgk-1* by 15%. Yet, consistent with its functional status, it has retained an open reading frame encoding a protein of 417 amino acids, exactly the same length as that encoded by the *hPgk-1* gene. Other indications that the divergence of *hPgk-2* from *hPgk-1* has been restricted by selection operating on a functional protein include: 1) a higher homology between the *hPgk-2* and *hPgk-1* amino acid sequences than between the nucleotide sequences, 2) a disproportionately large percentage of the nucleotide sequence differences between *hPgk-2* and *hPgk-1* occurring at the third position in a codon, and 3) a significantly higher occurrence of synonymous versus nonsynonymous nucleotide substitutions (Table I). In each case, the divergence of the ψ *hPgk-1* pseudogene from the *hPgk-1* gene, which has occurred in the absence of any selection at the protein level, demonstrates the opposite pattern indicating that these characteristics are diagnostic of the divergence of functional versus non-functional genes. By contrast, the relative occurrence of transitions and transversions between either *hPgk-2* or ψ *hPgk-1* and *hPgk-1* is not diagnostically different.

The relative effect of selection at the protein level to restrict divergence of functional *Pgk* coding sequences is demonstrated in Table II. The overall homology of 85% is represented equally throughout the coding sequences of *hPgk-2* and *hPgk-1*, including in the first (5'-CS) and last (3'-CS) 60 bp of the coding sequence shown in Table II. Similarly the 94% homology between the putative coding sequence of ψ *hPgk-1* and the coding sequence of *hPgk-1* is consistent throughout. In the latter case, this same approximate degree of homology extends into the flanking untranslated regions (5'-UT and 3'-UT), indicating no differential selection on the divergence of the non-functional coding sequence and that of the flanking sequences. However, there is a dramatic drop in the degree of homology between the first 60 bp upstream of the AUG start codons (5'-UT) in *hPgk-2* and *hPgk-1*, as well as between the first 60 bp immediately downstream of the TAG stop codons (3'-UT). Thus, although both the *Pgk-2* coding sequence and the flanking untranslated sequences were duplicated from the *Pgk-1* gene in the same retroposition event, selection at the protein level has acted to restrict the divergence of the coding sequence relative to that of the untranslated sequences.

TABLE III: MAMMALIAN *Pgk* GENE REGION HOMOLOGIES (%)

Genes Compared	Promoter (200 bp)	5'-Untranslated (90 bp)	Coding Sequence (1251 bp)		3'-untranslated (200 bp)
			N.A. ¹	A.A. ²	
<i>hPgk-1</i> / <i>mPgk-1</i>	48.0	65.2	92.9	97.6	55.8
<i>hPgk-2</i> / <i>mPgk-2</i>	63.9	44.8 (67.2)*	84.9	86.1	56.0
<i>hPgk-1</i> / <i>hPgk-2</i>	44.1	46.2	85.3	87.1	55.9
<i>mPgk-1</i> / <i>mPgk-2</i>	36.1	24.7 (40.7)*	79.2	84.0	50.5
<i>hPgk-1</i> / Ψ <i>hPgk-1</i>	30.9	74.4	94.1	90.0	98.5

* Values in parentheses are based on maximized alignments with deletions excluded.

¹ N.A. - nucleic acid

² A.A. - amino acid

TABLE IV: MAMMALIAN *Pgk* GENE EXON HOMOLOGIES (%)

Exon	Genes Compared					
	N.A. ¹	A.A. ²	<i>hPgk-1/mPgk-1</i>		<i>hPgk-2/mPgk-2</i>	
			N.A.	A.A.	N.A.	A.A.
1	1- 66	1- 22	89.4	100.0	78.8	77.7
2	67- 117	23- 39	94.1	100.0	92.2	94.1
3	118- 273	40- 91	94.2	96.2	92.2	94.1
4	274- 417	92-139	94.4	95.8	84.0	83.3
5	418- 522	140-174	88.6	97.1	83.6	88.6
6	523- 642	175-214	90.0	100.0	79.4	90.0
7	643- 756	215-252	95.6	100.0	86.8	86.8
8	757- 936	253-312	94.4	98.3	86.1	80.0
9	937-1113	313-371	90.4	94.9	83.6	83.1
10	1114-1212	372-404	94.9	100.0	84.8	93.9
11	1213-1251	405-417	97.6	92.9	92.9	100.0
Chi square			12.12		8.03	

¹ N.A. - nucleic acid

² A.A. - amino acid

Divergence of *Pgk-1* and *Pgk-2*

With the availability of nucleotide sequence data for the mouse *Pgk* genes (3,4,14), comparisons can be made between the mouse and human *Pgk-1* and *Pgk-2* genes. This has been done on a gene region-specific basis in Table III. As predicted (18), there is a high degree of conservation (92.9%) between the coding sequences of the human and mouse X-linked *Pgk-1* genes, and a somewhat lower degree of conservation between the coding sequences of the autosomal *Pgk-2* genes from these same species (85.3%). However, this is due not merely to a general difference in divergence rates of X-linked and autosomal sequences since all four *Pgk* genes (*hPgk-1*, *mPgk-1*, *hPgk-2*, *mPgk-2*) share a

similar homology among their 5'- and 3'-untranslated regions. Indeed, the fact that the *hPgk-2* coding sequence has diverged as much from the *mPgk-2* coding sequence in approximately 75 Myr as it has from the *hPgk-1* coding sequence in over 125 Myr suggests specific selection favoring more rapid divergence of the *Pgk-2* genes. This may be related to the function of the *Pgk-2* gene product in spermatozoa versus that of the *Pgk-1* gene product in somatic cells. It is well established that spermatozoa demonstrate a degree of interspecific variability unparalleled in any other cell type. This is the case for many aspects of sperm structure and function (19), including the biochemistry of metabolic enzymes (20,21).

Effects of exon/intron structure on P_{gk} gene evolution

It has been suggested that exon/intron structure may have facilitated the evolution of many genes including *Pgk-1* by mechanisms such as exon shuffling (2,22,23). The mammalian *Pgk* genes provide an opportunity to assess the effect of exon/intron structure on gene divergence by comparing the variation between analogous exon sequences that are or are not separated by introns. A comparison of the divergence between exons of the *hPgk-1* and *mPgk-1* genes versus that between analogous exon-regions of the *hPgk-2* and *mPgk-2* genes (Table IV) indicates the inter-exon variation is not significantly different than random in either case when analyzed by chi square (X^2 at the .05 level = 18.3 for 10 degrees of freedom). However, the chi square statistic calculated for the *Pgk-1* genes (12.12) was higher than that calculated for the *Pgk-2* genes (8.03), indicating that in this case more inter-exon variation evolved in the *Pgk* genes that possess introns than in those that do not. This is surprising since the overall difference between the coding sequences of the *Pgk-1* genes is less than half that between the *Pgk-2* genes.

Codon bias and expression levels of P_{gk} genes

A summary of codon use in the *hPgk-1*, ψ *hPgk-1* and *hPgk-2* genes is presented in Table V. All three of these genes demonstrate a degree of codon bias. In bacteria, *Drosophila* and yeast a greater codon bias is found in genes expressed at high levels, and this bias conforms to that found in the charged tRNA pool, thus enhancing translational efficiency (15,24–29). If similar selection affects mammalian genes, the *hPgk-2* gene should demonstrate more codon bias than the *hPgk-1* gene, since it is expressed at significantly higher levels, and the ψ *hPgk-1* pseudogene should have less codon bias than the *hPgk-1* gene since the former is not expressed at all. However, the opposite is observed. Of 17 amino acids for which a difference was noted, the *hPgk-1* gene showed a greater codon bias than the *hPgk-2* gene for 10, and the ψ *hPgk-1* gene showed a greater bias than the *hPgk-1* gene for 12. This supports the suggestion that codon bias in genomes of vertebrates is not related to the relative expression levels of genes (27,30).

An alternate hypothesis to explain codon bias in vertebrate species has been proposed by Ohno (30,31). He suggests that a certain codon preference is the predictable consequence of the primordial periodic nature of coding sequences, and is not influenced by relative expression levels. Specifically, he has noted that this preference is based on an overall deficiency of TA and CG dinucleotides and a relative excess of TG and CT dinucleotides. Table VI shows the representation of each of the four individual bases, the A+T and G+C content, and the occurrence of each dinucleotide in each of the human *Pgk* genes. As predicted by Ohno's hypothesis, all three *hPgk* gene sequences contain a relatively high amount of the TG ($x = 10.1\%$) and CT ($x = 7.7\%$) dinucleotides, and a relative paucity of the TA ($x = 3.5\%$) and CG ($x = 1.1\%$) dinucleotides compared with the expected average of 6.25%.

Divergence of P_{gk} promoter sequences

Ohno (31) noted that one advantage of the deficiency of TA and CG dinucleotides in coding sequences is that they are effectively used in intergenic regulatory sequences. Indeed, the presence of CG dinucleotides in the promoter region of the *hPgk-1* gene has resulted in its categorization as a 'CpG-island' gene, a

characteristic of many ubiquitously expressed genes (32). It is interesting to note that although *mPgk-1* is also a CpG-island gene, the *hPgk-1* and *mPgk-1* promoter sequences are not otherwise highly conserved (48.0% – Table III). This would suggest that the CpG-island is the primary regulatory feature in both *Pgk-1* promoters and is sufficient to direct ubiquitous expression in both species. The *hPgk-2* gene is dramatically different in this regard, having apparently lost its CpG island and become TA-rich in the 5'-flanking region (5). This is the most striking difference between the *hPgk-1* and *hPgk-2* gene sequences, and appears to be a consequence of the evolution of cell type-specific expression of the *hPgk-2* gene.

The suggestion that the *Pgk-2* retroposon included a processed copy of the endogenous *Pgk-1* promoter sequence (Figure 1) is based primarily on the presence of a GC-box in the *hPgk-2* promoter, apparently in lieu of a TATA-box (5). Although this is a common feature of many ubiquitously expressed housekeeping genes, it is not generally found in the promoters of cell type-specific genes. This hypothesis predicts that the *Pgk-2* retroposon was originally expressed ubiquitously, and subsequently evolved cell type-specific expression. This is supported by the data in Table VII which show the relative tissue-specificity of *Pgk-2* gene expression in metatherian and eutherian mammals (7). Thus, since its origin prior to the divergence of metatherian and eutherian mammals, the expression of the *Pgk-2* retroposon has become increasingly tissue- and cell type-specific. This is presumably due to changes in the sequence of the promoter region, including the loss of the CpG island associated with the *Pgk-1* gene promoter. A recent survey of available vertebrate gene sequences showed that genes which lack a TATA-box and possess a 5' CpG island are invariably ubiquitously or widely expressed (32). Although there are other examples of TATA-less genes that are tissue-specific, there has been no report of a tissue-specific gene that both lacks a TATA-box and possesses a 5' CpG island (reviewed in 32). Thus, increasing selection during mammalian evolution for high-level expression of the *Pgk-2* gene specifically in spermatogenic cells has apparently resulted in the elimination of the 5' CpG island from this locus.

Consequences of P_{gk} promoter divergence

CpG islands actually represent the absence of a CpG deficiency typically found throughout most vertebrate genomes, and are maintained by hypomethylation (33). When methylated, cytosine tends to undergo deamination to thymine so that the CpG dinucleotide mutates to TpG and CpA (34). Thus a predicted consequence of the elimination of a CpG island would be a relative excess of the TG and CA dinucleotides, and a relative paucity of the CG dinucleotide. This is observed in a comparison of the first 150 bp upstream from the transcriptional start sites in the *hPgk-1* and *hPgk-2* gene promoters. The *hPgk-1* promoter contains 12.7% of the CG dinucleotide and 12.5% of the CA + TG dinucleotides combined. The *hPgk-2* promoter contains only 2.7% of the CG dinucleotide and 15.4% of the CA + TG dinucleotides.

The elimination of the CpG island from the 5' region of the *Pgk-2* gene may have influenced the coding region sequence as well. Ikemura et al. (27,35) have shown that genes surrounded by relatively G+C-rich sequences (typically housekeeping genes) tend to have a higher G+C percentage at the third base position in codons than do genes surrounded by G+C-depleted sequences (typically tissue-specific genes). In the CpG island-containing

TABLE V: HUMAN *Pgk* GENE CODON USAGE

Codon	AA ¹	Z Usage			Codon	AA	Z Usage			Codon	AA	Z Usage			Codon	AA	Z Usage		
		hPgk-1	ψPgk-1	hPgk-1			hPgk-1	ψPgk-1	hPgk-2			hPgk-1	ψPgk-1	hPgk-2			hPgk-1	ψPgk-1	hPgk-2
TTT	Phe	2.4	2.9	2.4	TCT	Ser	1.7	1.7	1.2	TAT	Tyr	0.5	0.2	0.7	TGT	Cys	0.7	1.0	1.0
TTC	Phe	1.4	1.7	1.2	TCC	Ser	1.0	1.0	1.9	TAC	Tyr	0.5	0.5	0.5	TGC	Cys	1.0	0.7	0.7
TTA	Leu	0.2	0.7	0.7	TCA	Ser	0.2	0.5	0.5	TAA	End	0.0	0.0	0.0	TGA	End	0.0	0.2	0.0
TTG	Leu	1.7	1.7	1.2	TCG	Ser	0.5	0.2	0.0	TAG	End	0.2	0.7	0.2	TGG	Trp	1.0	1.0	1.0
CTT	Leu	1.2	1.2	1.9	CCT	Pro	1.9	2.2	1.9	CAT	His	0.5	0.7	1.4	CGT	Arg	0.0	0.0	0.0
CTC	Leu	1.7	1.2	1.0	CCT	Pro	0.5	0.5	1.2	CAC	His	0.7	1.2	0.2	CGC	Arg	0.2	0.0	0.7
CTA	Leu	0.7	0.7	1.2	CCA	Pro	1.7	1.4	1.0	CAA	Gln	0.2	0.5	1.2	CGA	Arg	0.5	0.5	0.5
CTG	Leu	3.6	3.3	3.1	CCG	Pro	0.0	0.0	0.2	CAG	Gln	1.2	1.7	0.7	CGG	Arg	1.0	0.2	0.2
ATT	Ile	1.9	2.9	1.9	ACT	Thr	2.4	3.1	2.6	AAT	Asn	2.6	2.6	2.2	AGT	Ser	0.2	0.7	0.7
ATC	Ile	1.7	1.4	2.6	ACC	Thr	1.0	1.0	0.5	AAC	Asn	2.9	2.9	2.9	AGC	Ser	2.2	1.7	1.2
ATA	Ile	1.0	1.0	1.0	ACA	Thr	0.2	0.7	0.2	AAA	Lys	3.8	4.3	4.5	AGA	Arg	0.5	0.7	0.7
ATG	Met	3.3	2.9	3.6	ACG	Thr	0.5	0.5	0.0	AAG	Lys	6.2	4.5	5.7	AGG	Arg	0.5	0.7	0.7
GTT	Val	1.9	0.7	3.3	GCT	Ala	5.5	4.5	3.6	GAT	Asp	2.4	1.7	3.1	GGT	Gly	2.9	3.3	3.1
GTC	Val	3.3	3.3	1.7	GCC	Ala	3.8	3.6	3.6	GAC	Asp	3.1	2.9	2.9	GGC	Gly	2.6	2.6	1.7
GTA	Val	1.2	1.2	1.9	GCA	Ala	0.5	0.2	2.4	GAA	Glu	2.4	2.4	2.6	GGA	Gly	1.9	1.9	3.1
GTG	Val	2.9	3.6	1.9	GCG	Ala	0.0	0.0	0.0	GAG	Glu	4.1	3.8	2.9	GGG	Gly	2.2	1.7	1.4

¹ A.A. - amino acidTABLE VI: HUMAN *Pgk* GENE CODING SEQUENCE BASE COMPOSITION

Base(s)	Z Representation In			Base(s)	Z Representation In		
	hPgk-1	ψPgk-1	hPgk-2		hPgk-1	ψPgk-1	hPgk-2
A	25.8	26.7	28.1	CA	6.6	6.9	6.9
C	21.5	21.0	20.6	CC	5.4	5.4	5.4
G	28.9	27.3	26.3	CG	1.4	0.7	1.2
T	23.9	25.0	25.0	CT	8.1	7.9	7.1
				GA	8.0	7.9	8.1
A+T	49.7	51.7	53.1	GC	7.0	6.1	5.7
G+C	50.3	48.3	46.9	GG	8.5	7.7	7.2
				GT	5.3	5.4	5.2
AA	8.1	8.2	9.1	TA	3.0	3.5	4.0
AC	4.2	4.7	4.2	TC	4.9	4.6	5.3
AG	8.5	8.1	8.6	TG	10.5	10.5	9.3
AT	5.0	5.3	6.3	TT	5.6	6.1	6.4

TABLE VII: EVOLUTION OF SPERMATOGENESIS-SPECIFIC EXPRESSION OF THE *Pgk-2* RETROPOSON*

Infraclass/family	# Species studied	# (%) With somatic <i>Pgk-2</i> expression
Eutheria	22	2 (9%)
Metatheria/nonmacropididae	31	10 (32%)
Metatheria/macropididae	27	17 (63%)

* Data from VandeBerg (7)

hPgk-1 gene, 56% of the codons end in C or G. In the G+C-depleted *hPgk-2* gene, 48% of the codons end in G or C. In the unexpressed *ψhPgk-1* pseudogene, which has an overall G+C content intermediate between that of *hPgk-1* and *hPgk-2* (Table VI), 53% of the codons end in G or C.

Significance of *Pgk* gene divergence

The primary functional difference between the *Pgk-1* and *Pgk-2* genes in eutherian species is in the level and specificity of their

expression. Expression of the *Pgk-2* gene in spermatogenic cells appears to compensate for the depletion of *Pgk-1* gene product following inactivation of the single X chromosome in these cells. Inasmuch as spermatozoa require phosphoglycerate kinase, there must not be a similar depletion of *Pgk* gene product in spermatogenic cells of non-mammalian vertebrates or prototherian mammals which have only one functional *Pgk* gene. In these cases *Pgk* may not be X-linked, or X-chromosome inactivation may be less stringent in spermatogenic cells. However, with the

appearance of a second, autosomal *Pgk* gene in the genomes of metatherian and eutherian species, depletion of the X-linked *Pgk-1* gene product during spermatogenesis could be tolerated. Thus selection for high-level, cell type-specific expression of the *Pgk-2* gene could have come from a combination of increasing stringency of X-chromosome inactivation and/or more rapid depletion of *Pgk-1* gene product in spermatogenic cells, and the absence of any selection for *Pgk-2* expression in other cell types. An apparent consequence of this process was the loss of the CpG island structure in the *Pgk-2* promoter, which may also have predisposed a bias for codons ending in A or T. A major question that remains is of what benefit inactivation of the X chromosome is to the process of spermatogenesis.

ACKNOWLEDGEMENTS

This work was supported by grants HD 23126 and HD 00829 from the National Institutes of Health. I am grateful to Drs. M.Green, M.A.Handel, R.MacIntyre, S.Ohno and B.Osborne for their critical reading of this manuscript.

REFERENCES

- McCarrey, J.R. & Thomas, K. (1987) *Nature* 326:501–505.
- Michelson, A.M., Blake, C.C., Evans, S.T. & Orkin, S.H. (1985) *Proc. Natl. Acad. Sci. (USA)* 82:6965–6969.
- Boer, P.H., Adra, C.N., Lau, Y-F. & McBurney, M.W. (1987) *Mol. Cell. Biol.* 7:3107–3112.
- Adra, C.N., Boer, P.H. & McBurney, M.W. (1987) *Gene* 60: 65–74.
- McCarrey, J.R. (1987) *Gene* 61:291–298.
- Michelson, A.M., Bruns, G.A.P., Morton, C.C. & Orkin, S.H. (1985) *J. Biol. Chem.* 260:6982–6992.
- VandeBerg, J.L. (1985) *Isozymes: Current topics in biological and medical research* 12:133–187.
- Kramer, J.M. & Erickson, R.P. (1981) *Devel. Biol.* 87:37–45.
- Kozak, L.P., McLean, G.K. & Eicher, E.M. (1974) *Biochem. Genet.* 11:41–47.
- Pegoraro, B. & Lee, C-Y. (1978) *Biochim. Biophys. Acta* 52:423–433.
- Michelson, A.M., Markham, A.G. & Orkin, S.H. (1983) *Proc. Natl. Acad. Sci. (USA)* 80:472–476.
- Singer-Sam, J., Simmer, R.L., Keith, D.H., Shively, L., Teplitz, M., Itakura, K., Gartler, S.M., & Riggs, A.D. (1983) *Proc. Natl. Acad. Sci. (USA)* 80:802–806.
- Singer-Sam, J., Keith, D.H., Tani, K., Simmer, R.L., Shively, L., Lindsay, S., Yoshida, A. & Riggs, A.D. (1984) *Gene* 32: 409–417.
- Mori, N., Singer-Sam, J., Lee, C-Y. & Riggs, A.D. (1986) *Gene* 45:275–280.
- Shields, D.C., Sharpe, P.M., Higgins, D.G. & Wright, F. (1988) *Mol. Biol. Evol.* 5:704–716.
- Adra, C.N., Ellis, N.A. & McBurney, M.W. (1988) *Somatic Cell. Mol. Genet.* 14:69–81.
- Potten, H., Jendraschak, E., Hauck, S., Amar, L.C., Avner, P. & Mullhofer, G. (1988) *Gene* 71:461–471.
- Ohno, S. (1967) *Sex chromosomes and sex-linked genes*. Springer-Verlag, New York.
- Baccetti, B. & Afzelius, B.A. (1976) *The biology of the sperm cell*. Monographs in Devel. Biol., vol. 10, S. Karger, Basel.
- Guraya, S.S. (1987) *Biology of spermatogenesis and spermatozoa in mammals*. Springer-Verlag, Berlin.
- Mann, T. (1967) Sperm metabolism. In: *Fertilization: comparative morphology, biochemistry and immunology*. C. Metz & A. Monroy (eds.). Academic Press, New York.
- Gilbert, W. (1978) *Nature* 271:501.
- Blake, C.C.F. (1978) *Nature* 273:267.
- Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. (1981) *Nuc. Acids Res.* 9:r43–r74.
- Bennetzen, J.L. & Hall, B.D. (1981) *J. Biol. Chem.* 257:3026–3031.
- Grosjean, H. & Fiers, W. (1982) *Gene* 18:199–209.
- Ikemura, T. (1985) *Mol. Biol. Evol.* 2:13–34.
- Sharp, P.M., Tuohy, T.M.F. & Mosurski, K.R. (1986) *Nuc. Acids Res.* 14:5125–5143.
- Sharp, P.M. & Li, W-H. (1987) *Mol. Biol. Evol.* 4:222–230.
- Ohno, S. (1988) *Proc. Natl. Acad. Sci. (USA)* 85:4378–4382.
- Ohno, S. (1988) *Proc. Natl. Acad. Sci. (USA)* 85:9630–9634.
- Gardiner-Garden, M. & Frommer, M. (1987) *J. Mol. Biol.* 196:261–282.
- Bird, A.P. (1987) *Trends in Genet.* 3:342–347.
- Bird, A.P. (1980) *Nuc. Acids Res.* 8:1499–1504.
- Aota, S-I. & Ikemura, T. (1986) *Nuc. Acids Res.* 14:6345–6355.