Available online at www.sciencedirect.com

Clinical Biochemistry 37 (2004) 961–967

CLINICAL
BIOCHEMISTRY

Cloning of a kallikrein pseudogene

George M. Yousef[a], Carla A. Borgono[b,c], Iacovos P. Michael[b,c], Eleftherios P. Diamandis[b,c,*]

[a]*Discipline of Pathology, Health Care Corporation of St. John's, St. John's, Newfoundland, Canada*
[b]*Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Toronto, Ontario, Canada*
[c]*Department of Laboratory Medicine and Pathobiology, University of Toronto, Ontario, Canada*

Received 20 January 2004; received in revised form 16 June 2004; accepted 22 July 2004
Available online 22 September 2004

Abstract

Objectives: Kallikreins are a group of serine proteases clustered together on a small region of chromosome 19q13.4. Recent reports suggest that kallikreins are differentially expressed in malignancy and have potential as cancer biomarkers. The human kallikrein gene locus has now been fully characterized and 15 functional kallikreins were identified. Although many kallikrein pseudogenes have already been characterized in rodents, none have been identified in humans.

Methods and results: In the current study, we identified the first human kallikrein pseudogene named *ΨKLK1* and mapped it between the *KLK2* and *KLK4* genes. This pseudogene shares a moderate degree of similarity with the adjacent functional kallikreins. It has a conserved histidine residue of the catalytic triad of serine proteases and its surrounding motif, but lacks the aspartate and serine residues. Positions of some cysteine residues are also conserved in the pseudogene. This pseudogene lacks intronic sequences and should thus be classified as a processed pseudogene. EST and PCR analyses indicate that this pseudogene may be transcriptionally active, because mRNA was detected in many tissues including the prostate, testis, pituitary, and adrenal glands, as well as in tissues of the female genital organs.

Discussion: The mRNA sequence of the gene is, however, defective and is not predicted to code for a protein. Highly conserved sequences were found in the flanking region of the pseudogene, thus supporting the view that it evolved by retrotransposition. We also identified another serine protease fragment that has only the conserved histidine residue. The functional significance of the pseudogene and the other fragment is yet to be identified.

© 2004 The Canadian Society of Clinical Chemists. All rights reserved.

Keywords: Kallikreins; Cancer biomarkers; Pseudogene; Serine proteases

Introduction

The term "pseudogene" comprises a wide group of nonfunctional loci with a marked diversity of characteristics [1]. Vanin [2] defined a pseudogene by two major features: being related to one or more paralogous genes, and being defective in function. The lack of function results from either failure of transcription, translation, or production of a protein that does not have the same functional repertoire as that encoded by the normal paralog gene [3]. Most pseudogenes are created by one of two mechanisms: tandem duplication (nonprocessed pseudogenes) or retrotransposition from a functional gene (processed pseudogenes) [4].

Kallikreins are a subgroup of the serine protease family of enzymes [5,6]. In humans, this family consists primarily of the plasma kallikrein gene and tissue or glandular kallikreins. Plasma kallikrein is encoded by a single gene that is structurally different from the genes encoding tissue kallikreins. Tissue kallikreins comprise a large multigene family of enzymes in human and many other species [7] with a highly conserved sequence and tertiary structures [6]. Many members of the human kallikrein gene family were found to be differentially expressed in diverse disease states including diseases of the central nervous system [8], skin

Abbreviations: RT-PCR, reverse transcription-polymerase chain reaction; EST, expressed sequence tag; ψKLK1, kallikrein pseudogene-1.

* Corresponding author. Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, 600 University Avenue, Toronto, ON, Canada M5G 1X5. Fax: +1 416 586 8628.

E-mail address: ediamandis@mtsinai.on.ca (E.P. Diamandis).

0009-9120/$ - see front matter © 2004 The Canadian Society of Clinical Chemists. All rights reserved.
doi:10.1016/j.clinbiochem.2004.07.012

[9,10], and malignancy [11–13]. In our previous work, we characterized the human kallikrein gene locus on chromosome 19q13.4 [5,14]. Data from our laboratory and others indicate that no additional functional kallikreins are present in this locus [6,15,16].

In rodents, kallikreins are represented by large multigene families. In the mouse genome, at least 24 genes have been identified, and many others were recently found by bioinformatics tools [17]. A similar family of 15–20 kallikreins has been found in the rat genome [18]. The structural organization of the kallikrein genes includes five coding exons; this structure is highly conserved in all species studied thus far [19]. Interestingly, while many kallikrein pseudogenes were identified in rodents, none have been characterized in humans thus far. However, both Stephenson et al. [20] and Gan et al. [15] have discovered incomplete fragments or potential pseudogenes in the kallikrein locus.

The aim of this study was to analyze the human kallikrein gene locus to identify any possible kallikrein pseudogenes. Although the functional significance of these molecules is not yet clear, characterization of such pseudogenes will help in our understanding of the evolutionary history of kallikreins and the functional significance of kallikreins among species. In this paper, we characterized the first human kallikrein-processed pseudogene and identified its chromosomal localization and its structural features. We also identified another serine protease fragment with a conserved histidine residue.

Materials and methods

Expressed sequence tag (EST) searching

The predicted structure of the putative new pseudogene was subjected to homology search using the BLASTN algorithm [21] on the National Center for Biotechnology Information web server (http://www ncbi.nlm.nih.gov/BLAST/) against the human EST database. A clone with ≥98% identity was obtained from the IMAGE consortium through Research Genetics Inc, Huntsville, AL. This clone was propagated, purified, and sequenced from both directions with an automated sequencer using insert-flanking vector primers.

Reverse transcriptase polymerase chain reaction (RT-PCR)

The RNA was treated with DNase before reverse transcription. Total RNA (2 μg) was reverse transcribed into first-strand cDNA using the Superscript™ pre-amplification system (Invitrogen, Carlsbad, CA). The final volume was 20 μl. Gene-specific primers (A-F2: 5′ TCA CTA CTG CTC ACT GCA TC 3′ and A-R3: 5′ CAT ATG TAG GTA CTG TAG GG 3′) were used for PCR-based amplification of a human tissue panel as described below. PCR was carried out in a reaction mixture containing 1 μl of cDNA,10 mM Tris–HCl (pH 8.3), 50 mM KCl, 2 mM $MgCl_2$, 200 μM dNTPs (deoxynucleoside triphosphates), 100 ng of primers, and 2.5 units of HotStar Taq polymerase (Qiagen, Valencia, CA) on an Eppendorf thermal cycler. The cycling conditions were 95°C for 15 min to activate the HotStar Taq polymerase, followed by 40 cycles of 94°C denaturation for 30 s, 56°C annealing for 30 s, and 72°C extension for 30 s, and a final extension at 72°C for 10 min. Equal amounts of PCR products were electrophoresed on 1.5% agarose gels and visualized by ethidium bromide staining.

Cloning and sequencing of the PCR products

Due to the high degree of homology between the genes in this genomic region, primers were designed to be specific for the pseudogene, annealing away from conserved regions. To further verify the identity of the PCR products, they were cloned into the pCR 2.1-TOPO vector (Invitrogen) according to the manufacturer's instructions. The inserts were sequenced from both directions using vector-specific primers with an automated DNA sequencer.

Tissue expression

Total RNA isolated from 36 different human tissues was purchased from Clontech, Palo Alto, CA. We prepared cDNA as described above and used it for PCR amplification. Tissue cDNAs were amplified at various dilutions using gene-specific primers. The RNA was treated with DNase before reverse transcription.

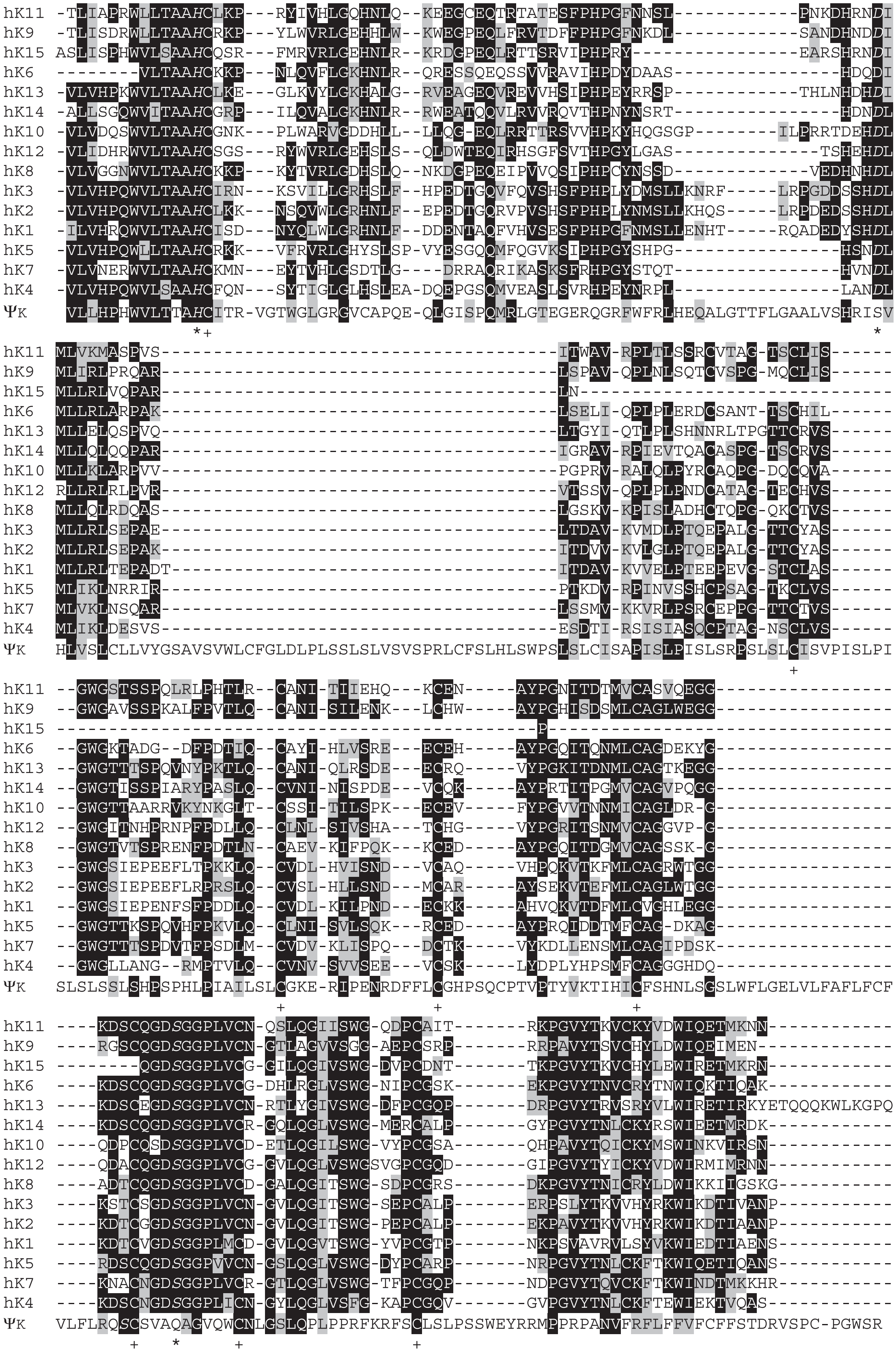
Structure analysis

Genomic sequences generated by the Human Genome Project (HGP) were obtained from the NCBI web site (www.ncbi.nlm.nih.gov). Several computer programs were used to predict the presence of putative new genes in the genomic area of interest, as described before [22]. Protein translation was performed using the "Translate" software available from the ExPasy web server (www.ExPasy.org). Multiple alignment was performed using the "Clustal X" software package. Conserved domain search was performed using the "Conserved Domain" (CD) and "ProDom" programs. Motif search was performed using the "ScanProsite" and "MotifScan" programs available from the ExPasy web server.

Results

Identification and genomic organization of the kallikrein pseudogene (ψKLK1)

We have previously characterized the human kallikrein gene locus. The locus spans a region of 261,558 bp on



```
hK11 -TLIAPRWLLTAAHCLKP---RYIVHLGQHNLQ--KEEGCEQTRTATESFPHPGFNNSL------------PNKDHRNDI
hK9  -TLISDRWLLTAAHCRKP---YLWVRLGEHHLW--KWEGPEQLFRVTDFFPHPGFNKDL------------SANDHNDDI
hK15 ASLISPHWVLSAAHCQSR---FMRVRLGEHNLR--KRDGPEQLRTTSRVIPHPRY----------------EARSHRNDI
hK6  --------VLTAAHCKKP---NLQVFLGKHNLR--QRESSQEQSSVVRAVIHPDYDAAS----------------HDQDI
hK13 -VLVHPKWVLTAAHCLKE---GLKVYLGKHALG--RVEAGEQVREVVHSIPHPEYRRSP------------THLNHDHDI
hK14 -ALLSGQWVITAAHCGRP---ILQVALGKHNLR--RWEATQQVLRVVRQVTHPNYNSRT----------------HDNDL
hK10 -VLVDQSWVLTAAHCGNK---PLWARVGDDHLL--LLQG-EQLRRTTRSVVHPKYHQGSGP--------ILPRRTDEHDL
hK12 -VLIDHRWVLTAAHCSGS---RYWVRLGEHSLS--QLDWTEQIRHSGFSVTHPGYLGAS--------------TSHEHDL
hK8  -VLVGGNWVLTAAHCKKP---KYTVRLGDHSLQ--NKDGPEQEIPVVQSIPHPCYNSSD-------------VEDHNHDL
hK3  -VLVHPQWVLTAAHCIRN---KSVILLGRHSLF--HPEDTGQVFQVSHSFPHPLYDMSLLKNRF-----LRPGDDSSHDL
hK2  -VLVHPQWVLTAAHCLKK---NSQVWLGRHNLF--EPEDTGQRVPVSHSFPHPLYNMSLLKHQS-----LRPDEDSSHDL
hK1  -ILVHRQWVLTAAHCISD---NYQLWLGRHNLF--DDENTAQFVHVSESFPHPGFNMSLLENHT-----RQADEDYSHDL
hK5  -VLVHPQWLLTAAHCRKK---VFRVRLGHYSLSP-VYESGQQMFQGVKSIPHPGYSHPG----------------HSNDL
hK7  -VLVNERWVLTAAHCKMN---EYTVHLGSDTLG----DRRAQRIKASKSFRHPGYSTQT----------------HVNDL
hK4  -VLVHPQWVLSAAHCFQN---SYTIGLGLHSLEA-DQEPGSQMVEASLSVRHPEYNRPL----------------LANDL
ΨK    VLLHPHWVLTTAHCITR-VGTWGLGRGVCAPQE-QLGISPQMRLGTEGERQGRFWFRLHEQALGTTFLGAALVSHRISV
                  *+                                                                *

hK11 MLVKMASPVS--------------------------------------ITWAV-RPLTLSSRCVTAG-TSCLIS------
hK9  MLIRLPRQAR--------------------------------------LSPAV-QPLNLSQTCVSPG-MQCLIS------
hK15 MLLRLVQPAR--------------------------------------LN------------------------------
hK6  MLLRLARPAK--------------------------------------LSELI-QPLPLERDCSANT-TSCHIL------
hK13 MLLELQSPVQ--------------------------------------LTGYI-QTLPLSHNNRLTPGTTCRVS------
hK14 MLLQLQQPAR--------------------------------------IGRAV-RPIEVTQACASPG-TSCRVS------
hK10 MLLKLARPVV--------------------------------------PGPRV-RALQLPYRCAQPG-DQCQVA------
hK12 RLLRLRLPVR--------------------------------------VTSSV-QPLPLPNDCATAG-TECHVS------
hK8  MLLQLRDQAS--------------------------------------LGSKV-KPISLADHCTQPG-QKCTVS------
hK3  MLLRLSEPAE--------------------------------------LTDAV-KVMDLPTQEPALG-TTCYAS------
hK2  MLLRLSEPAK--------------------------------------ITDVV-KVLGLPTQEPALG-TTCYAS------
hK1  MLLRLTEPADT-------------------------------------ITDAV-KVVELPTEEPEVG-STCLAS------
hK5  MLIKLNRRIR--------------------------------------PTKDV-RPINVSSHCPSAG-TKCLVS------
hK7  MLVKLNSQAR--------------------------------------LSSMV-KKVRLPSRCEPPG-TTCTVS------
hK4  MLIKLDESVS--------------------------------------ESDTI-RSISIASQCPTAG-NSCLVS------
ΨK   HLVSLCLLVYGSAVSVWLCFGLDLPLSSLSLVSVSPRLCFSLHLSWPSLSLCISAPISLPISLSRPSLSLCISVPISLPI
                                                                           +

hK11 --GWGSTSSPQLRLPHTLR--CANI-TIIEHQ---KCEN-----AYPGNITDTMVCASVQEGG-----------------
hK9  --GWGAVSSPKALFPVTLQ--CANI-SILENK---LCHW-----AYPGHISDSMLCAGLWEGG-----------------
hK15 ----------------------------------------------P---------------------------------
hK6  --GWGKTADG--DFPDTIQ--CAYI-HLVSRE---ECEH-----AYPGQITQNMLCAGDEKYG-----------------
hK13 --GWGTTTSPQVNYPKTLQ--CANI-QLRSDE---ECRQ-----VYPGKITDNMLCAGTKEGG-----------------
hK14 --GWGTISSPIARYPASLQ--CVNI-NISPDE---VCQK-----AYPRTITPGMVCAGVPQGG-----------------
hK10 --GWGTTAARRVKYNKGLT--CSSI-TILSPK---ECEV-----FYPGVVTNNMICAGLDR-G-----------------
hK12 --GWGITNHPRNPFPDLLQ--CLNL-SIVSHA---TCHG-----VYPGRITSNMVCAGGVP-G-----------------
hK8  --GWGTVTSPRENFPDTLN--CAEV-KIFPQK---KCED-----AYPGQITDGMVCAGSSK-G-----------------
hK3  --GWGSIEPEEFLTPKKLQ--CVDL-HVISND---VCAQ-----VHPQKVTKFMLCAGRWTGG-----------------
hK2  --GWGSIEPEEFLRPRSLQ--CVSL-HLLSND---MCAR-----AYSEKVTEFMLCAGLWTGG-----------------
hK1  --GWGSIEPENFSFPDDLQ--CVDL-KILPND---ECKK-----AHVQKVTDFMLCVGHLEGG-----------------
hK5  --GWGTTKSPQVHFPKVLQ--CLNI-SVLSQK---RCED-----AYPRQIDDTMFCAG-DKAG-----------------
hK7  --GWGTTTSPDVTFPSDLM--CVDV-KLISPQ---DCTK-----VYKDLLENSMLCAGIPDSK-----------------
hK4  --GWGLLANG--RMPTVLQ--CVNV-SVVSEE---VCSK-----LYDPLYHPSMFCAGGHDQ------------------
ΨK   SLSLSSLSHPSPHLPIAILSLCGKE-RIPENRDFFLCGHPSQCPTVPTYVKTIHICFSHNLSGSLWFLGELVLFAFLFCF
                          +              +                  +

hK11 ----KDSCQGDSGGPLVCN-QSLQGIISWG-QDPCAIT-------RKPGVYTKVCKYVDWIQETMKNN------------
hK9  ----RGSCQGDSGGPLVCN-GTLAGVVSGG-AEPCSRP-------RRPAVYTSVCHYLDWIQEIMEN-------------
hK15 --------QGDSGGPLVCG-GILQGIVSWG-DVPCDNT-------TKPGVYTKVCHYLEWIRETMKRN------------
hK6  ----KDSCQGDSGGPLVCG-DHLRGLVSWG-NIPCGSK-------EKPGVYTNVCRYTNWIQKTIQAK------------
hK13 ----KDSCEGDSGGPLVCN-RTLYGIVSWG-DFPCGQP-------DRPGVYTRVSRYVLWIRETIRKYETQQQKWLKGPQ
hK14 ----KDSCQGDSGGPLVCR-GQLQGLVSWG-MERCALP-------GYPGVYTNLCKYRSWIEETMRDK------------
hK10 ----QDPCQSDSGGPLVCD-ETLQGILSWG-VYPCGSA-------QHPAVYTQICKYMSWINKVIRSN------------
hK12 ----QDACQGDSGGPLVCG-GVLQGLVSWGSVGPCGQD-------GIPGVYTYICKYVDWIRMIMRNN------------
hK8  ----ADTCQGDSGGPLVCD-GALQGITSWG-SDPCGRS-------DKPGVYTNICRYLDWIKKIIGSKG-----------
hK3  ----KSTCSGDSGGPLVCN-GVLQGITSWG-SEPCALP-------ERPSLYTKVVHYRKWIKDTIVANP-----------
hK2  ----KDTCGGDSGGPLVCN-GVLQGITSWG-PEPCALP-------EKPAVYTKVVHYRKWIKDTIAANP-----------
hK1  ----KDTCVGDSGGPLMCD-GVLQGVTSWG-YVPCGTP-------NKPSVAVRVLSYVKWIEDTIAENS-----------
hK5  ----RDSCQGDSGGPVVCN-GSLQGLVSWG-DYPCARP-------NRPGVYTNLCKFTKWIQETIQANS-----------
hK7  ----KNACNGDSGGPLVCR-GTLQGLVSWG-TFPCGQP-------NDPGVYTQVCKFTKWINDTMKKHR-----------
hK4  ----KDSCNGDSGGPLICN-GYLQGLVSFG-KAPCGQV-------GVPGVYTNLCKFTEWIEKTVQAS------------
ΨK   VLFLRQSCSVAQAGVQWCNLGSLQPLPPRFKRFSCLSLPSSWEYRRMPPRPANVFRFLFFVFCFFSTDRVSPC-PGWSR
            +   *     +                +
```

Fig. 1. Alignment of the deduced amino acid sequence of the human kallikrein pseudogene-1 (*ΨKLK1*) with members of the kallikrein multigene family. For kallikrein protein accession numbers, please see our recent review [5]. Dashes represent gaps to bring the sequences to better alignment. The residues of the catalytic triad are shown in italics and are marked by stars. Identical residues are highlighted in black and similar residues in grey. The conserved cysteine residues are indicated by (+). ΨK, human kallikrein pseudogene-1.

chromosome 19q13.4 and is formed from 15 kallikrein genes with no intervening non-kallikrein genes [5,6,11]. Several gene prediction programs were used to identify any potential new genes in this region, but none was identified. A hypothetical non-kallikrein protein was predicted between *KLK2* and *KLK4* by the NCBI's annotation project (GenBank accession # XM_115594). This gene was not, however, supported by the EST database. We were also not able to amplify different exons of this predicted gene by PCR from cDNAs of any of 35 different tissues, thus questioning the validity of this prediction. Sequence analysis of potentially translated nucleotides in the region between the *KLK2* and *KLK4* genes revealed, however, a relatively conserved trypsin-like domain of serine proteases. Searching the expressed sequence tag (EST) databases, we identified an EST clone (AA559303) from a prostatic intraepithelial neoplasia library that maps to this conserved domain region. Attempts to translate the nucleotides of the putative gene in all possible frames resulted in the identification of one frame that produces a polypeptide with a considerable degree of homology to kallikreins (Fig. 1) and the trypsin domain of serine proteases. Because this frame is, however, interrupted by three in-frame stop codons, we considered this gene to be a pseudogene. We named it human kallikrein pseudogene 1 (*ψKLK1*). Based on matching ESTs, our PCR analysis, and the homologous region with other kallikreins, we calculated that *ψKLK1* spans an area of approximately 1070 nucleotides of genomic sequence on chromosome 19q13.4 and is located 16986 nucleotides centromeric to *KLK4* and 8142 nucleotides telomeric to *KLK2* (Fig. 2A). It is interesting to note that the mouse kallikrein locus, located on cytogenic region B2 on chromosome 7, syntenic to the human kallikrein locus on 19q13.4, contains a pseudogene (*ψMGK25)* in the position and orientation equivalent to *ψKLK1* in the human locus (Fig. 2B) [17]. The full sequence of *ψKLK1* was submitted to GenBank (GenBank accession # AY302756). Because there are no more homologous sequences on either end and due to the presence of repeat elements flanking this sequence, we assume that the sequence we submitted to GenBank represents the full sequence of the pseudogene. It should be mentioned, however, that as is the case with many other pseudogenes, the exact extension of the sequence might be impossible to verify.

To reveal the structural identity of this pseudogene, we performed an analysis of its hypothetical protein product after removal of the three interrupting stop codons. "Conserved Domain" (CD) and "ProDom" searches together with homology alignment indicated the presence of one of the three conserved amino acids of the catalytic triad of serine proteases. The highest degree of structural similarity was found with the adjacent kallikrein subfamily of serine proteases. The histidine residue of the catalytic triad and the amino acid motif around it is highly conserved (Fig. 1). The serine and aspartate residues are not conserved. In addition, many of the kallikrein conserved cysteine residues were found to be conserved in the structure of the pseudogene. The pseudogene shows highest similarity with the adjacent *KLK4* gene (40–50%). There is also 30–40% homology with other kallikreins. Taken together, these data suggest that *ψKLK1* should be considered a kallikrein pseudogene. One hundred forty nucleotides near the end of the pseudogene (the area not conserved with kallikrein sequences) showed multiple hits with many chromosomal regions, indicating the presence of a possible repeat element.

To verify the structure of this pseudogene and to examine its transcriptional activity, we screened the human EST database and we were able to identify an EST clone from a prostatic intraepithelial neoplasia library with >98% homology with the predicted sequence of the gene (AA559303). Four other EST clones with partial matching with the 3′ end of the pseudogene were also identified from prostate,

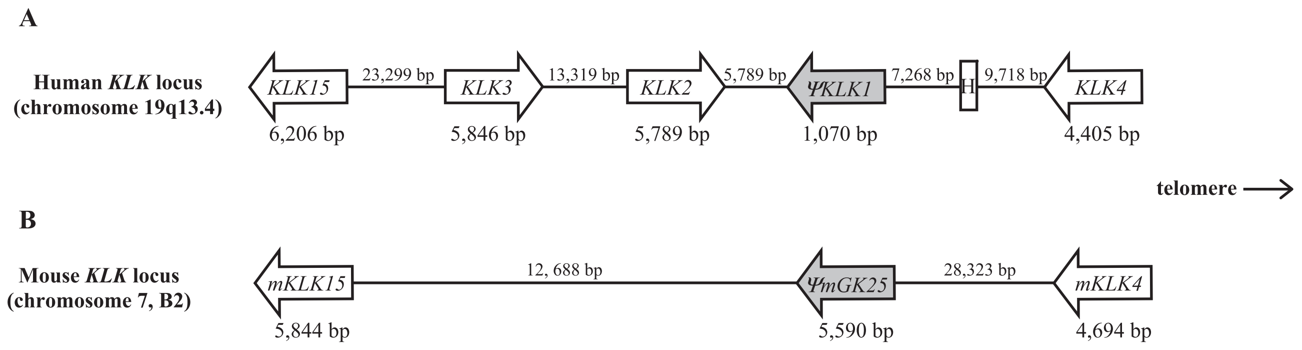


Fig. 2. Chromosomal localization and direction of transcription of (A) *ψKLK1* in the human *KLK* locus on chromosome 19q13.4 and (B) *ψmGK25* in the mouse *KLK* locus on chromosome 7, region B2. Genes are represented by white horizontal arrows, which denote the direction of the coding sequence. Pseudogenes are drawn as shaded horizontal arrows. Intergenic and gene lengths distances are shown in base pairs (bp). The genomic sequence encoding the serine protease fragment with a conserved histidine domain is indicated by "H". Figure is not drawn to scale. (Note: The gene lengths for *mKLK4*, *ψmGK25*, and *mKLK15* were derived from the "gene" coordinates in GenBank entries AF198031, AY152430, and AY152434, respectively. Intergenic distances were determined using the mouse genome map available from NCBI. Because the mouse genome is currently incomplete, distances shown may change in the future.)

adrenal, and uterine cancer libraries. These EST clones, however, show only a partial match with the genomic sequence and show other partial matches with other chromosomes, thus questioning their reliability. We also screened a panel of 36 tissues by RT-PCR using gene-specific primers. The mRNA of this gene was found to be highly expressed in the pituitary gland, testis, prostate, adrenal gland, breast, esophagus, and many tissues of the female genital system including the fallopian tube, ovary, cervix, uterus, and vagina, and to a lower extent in other tissues (Table 1 and Fig. 3). These results are, in general, consistent with the EST findings. Screening of this panel of tissues did not identify any other alternatively spliced form of this gene that might be encoding a non-truncated protein. BLAST search against the GenBank database did not reveal the presence of any orthologues for this pseudogene.

We compared the 5′ and 3′ flanking ends of the pseudogene against each other and against all direct repeat sequences reported by Vanin [2] for all known processed pseudogenes. Interestingly, the two regions were highly similar (Fig. 4). The boxed areas in Fig. 4 show the highly similar direct repeats in these regions.

Table 1
Tissue expression of *ΨKLK1* by RT-PCR

Tissue	Expression level[a]
Adipose	++
Adrenal	+
Bone marrow	−
Brain	−
Cerebellum	+
Cervix	++
Colon	−
Esophagus	++
Fallopian tube	+++
Fetal brain	−
Fetal liver	++
Heart	−
Hippocampus	++
Kidney	−
Liver	−
Lung	−
Mammary gland	++
Ovary	++
Pancreas	−
Placenta	−
Prostate	++
Pituitary	+++
Salivary	−
Skeletal muscle	−
Skin	++
Spinal cord	−
Spleen	+
Small intestine	−
Stomach	−
Testis	++
Thymus	−
Thyroid	+
Tonsil	+
Trachea	−
Uterus	++
Vagina	++

[a] +++, high; ++, moderate; +, low; −, negative.

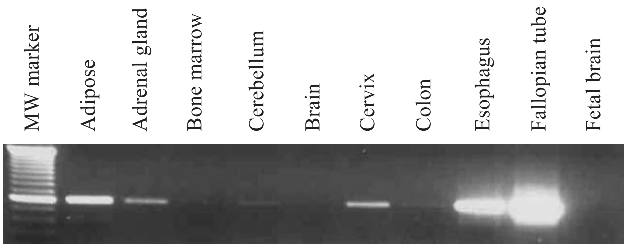


Fig. 3. Representative tissue expression pattern of *ψKLK1* as determined by RT-PCR. For additional data, see Table 1.

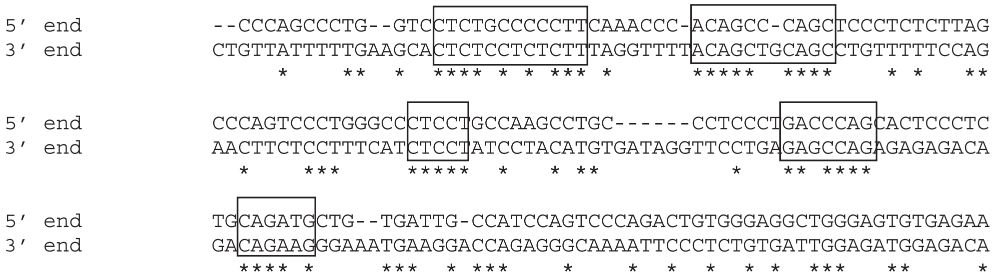
Sequence analysis of the kallikrein locus indicated the presence of another serine protease fragment with a conserved histidine residue. We identified the “LSAAHC” motif, which has a statistically significant homology with the SwissProt signature motif of the histidine residue of serine proteases ([LIVM]-[ST]-A-[STAG]-H-C [H is the active site residue] and the underlined amino acids are present in the identified motif) (SwissProt ID: PDOC00124). The location of this fragment is shown in Fig. 2. No EST matches were found for this fragment.

Discussion

In this paper, we report the cloning of the first human kallikrein pseudogene *ΨKLK1*. It was classified as a kallikrein pseudogene based on its structural similarity with other kallikrein proteins and its defective structure that will make it biologically inactive as a serine protease (all three possible reading frames encode predicted truncated protein products that are interrupted by stop codons, in addition to the missing aspartate and serine residues of the catalytic triad).

Pseudogenes fall into two major categories: those which retain their intervening sequence and those lacking intervening intronic sequences termed “processed pseudogenes” (a more abundant category). Our newly identified pseudogene meets the characteristic criteria for processed pseudogenes, including the lack of intronic sequences in the genomic structure, cessation of homology with the functional gene at the start and end of transcription, and the presence of flanking direct repeats [2]. Extra poly A sequences that are present in many processed pseudogenes were not found in *ΨKLK1*.

Many processed pseudogenes were reported in different chromosomes away from their functional counterparts. Few, however, were found near the functional paralog [1]. As the term pseudogene is a “negative” definition, it should be noted that the possibility still exists that there are functional splice variants of these genes that exist only in certain tissues or certain developmental stages or pathological situations. Also, as is the case with other pseudogenes, the possibility still exists that they can encode for a truncated, yet functional protein [2,23,24].



```
5′ end    --CCCAGCCCTG--GTCCTCTGCCCCCTTCAAACCC-ACAGCC-CAGCTCCCTCTCTTAG
3′ end    CTGTTATTTTTGAAGCACTCTCCTCTCTTTAGGTTTTACAGCTGCAGCCTGTTTTTCCAG
               *    **  *  **** * * *** *      *****  ****    * *   **

5′ end    CCCAGTCCCTGGGCCCTCCTGCCAAGCCTGC------CCTCCCTGACCCAGCACTCCCTC
3′ end    AACTTCTCCTTTCATCTCCTATCCTACATGTGATAGGTTCCTGAGAGCCAGAGAGAGACA
            *    ***     *****  *   * **          *   ** ****

5′ end    TGCAGATGCTG--TGATTG-CCATCCAGTCCCAGACTGTGGGAGGCTGGGAGTGTGAGAA
3′ end    GACAGAAGGGAAATGAAGGACCAGAGGGCAAAATTCCCTCTGTGATTGGAGATGGAGACA
            **** *     ***  * ***    *    *  *  *  * *  ***   **     *
```

Fig. 4. Alignment of the nucleotide sequence of the 5′ and 3′ regions flanking the *ψKLK1* gene. Dashes represent gaps to bring the sequences to better alignment. Identical residues are indicated by asterisks. Potential direct repeat areas are boxed.

It should be noted that not all processed pseudogenes reported are exact DNA copies of their respective RNAs. Two human immunoglobulin pseudogenes (human immunoglobulin ε and immunoglobulin λΨ1) and mouse corticoprotein β lipoprotein precursor pseudogenes have been shown to correspond to only part of their respective mRNAs [2].

The possible transcriptional activity of the *ΨKLK1* gene, shown by EST and PCR analysis, is not unprecedented. Consideration of how pseudogenes are formed suggested that most are unlikely to be transcribed. Transcripts of pseudogenes, however, have been previously reported for the mouse Ψ α_3-globin pseudogene, two Siglec genes [25], and others [26–28]. As some of the ESTs identified for the gene show also partial matches with other genomic sequences in other chromosomes and the genomic contamination of the RNA used for the PCR cannot be absolutely excluded, the transcriptional activity of the pseudogene should be interpreted with caution. The functional relevance of pseudogene transcripts remains unclear [3].

The study of pseudogenes should not be entirely regarded as a cul-de-sac. Although it is unlikely that reactivating (reverse) mutations will occur so as to restore their function, pseudogenes may nevertheless influence the evolution of other functionally significant sequences by, for example, mediating recombination events or acting as sequence donors in gene conversion [4].

The most acceptable model for the evolution of pseudogenes proposes the insertion of an mRNA intermediate through DNA breaks [4]. The presence of flanking direct repeats supports the theory of retrotransposition of a processed RNA intermediate lacking intervening sequences, which have probably been acquired during the process of transposition.

In conclusion, we cloned the first kallikrein pseudogene in the human kallikrein gene locus on chromosome 19q13.4. This will be useful in further understanding the evolution and function of these genes in humans.

Acknowledgment

This work was supported by a grant to E. P. Diamandis from the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Genomic Projects Initiative.

References

[1] Martinez-Arias R, Calafell F, Mateu E, et al. Sequence variability of a human pseudogene. Genome Res 2001;11:1071–85.

[2] Vanin EF. Processed pseudogenes: characteristics and evolution. Annu Rev Genet 1985;19:253–72.

[3] Mighell AJ, Smith NR, Robinson PA, et al. Vertebrate pseudogenes. FEBS Lett 2000;468:109–14.

[4] Cooper DN. Pseudogenes and their formation in human gene evolution. Oxford: BIOS Scientific Publishers Ltd, 1999. p. 265–93.

[5] Yousef GM, Diamandis EP. Human kallikreins: common structural features, sequence analysis and evolution. Curr Genomics 2003;4:147–65.

[6] Yousef GM, Diamandis EP. The new human tissue kallikrein gene family: structure, function, and association to disease. Endocr Rev 2001;22:184–204.

[7] Yousef GM, Diamandis EP. An overview of the kallikrein gene families in humans and other species: emerging candidate tumour markers. Clin Biochem 2003;36:443–52.

[8] Yousef GM, Kishi T, Diamandis EP. Role of kallikrein enzymes in the central nervous system. Clin Chim Acta 2003;329:1–8.

[9] Komatsu N, Takata M, Otsuki N, et al. Expression and localization of tissue kallikrein mRNAs in human epidermis and appendages. J Invest Dermatol 2003;121:542–9.

[10] Hansson L, Stromqvist M, Backman A, et al. Cloning, expression, and characterization of stratum corneum chymotryptic enzyme. A skin-specific human serine proteinase. J Biol Chem 1994;269: 19420–6.

[11] Diamandis EP, Yousef GM. Human tissue kallikreins: a family of new cancer biomarkers. Clin Chem 2002;48:1198–205.

[12] Yousef GM, Diamandis EP. Expanded human tissue kallikrein family—A novel panel of cancer biomarkers. Tumour Biol 2002;23:185–92.

[13] Yousef GM, Diamandis EP. Kallikreins, steroid hormones and ovarian cancer: is there a link? Minerva Endocrinol 2002;27:157–66.

[14] Yousef GM, Chang A, Scorilas A, et al. Genomic organization of the human kallikrein gene family on chromosome 19q13.3–q13.4. Biochem Biophys Res Commun 2000;276:125–33.

[15] Gan L, Lee I, Smith R, et al. Sequencing and expression analysis of the serine protease gene cluster located in chromosome 19q13 region. Gene 2000;257:119–30.

[16] Clements J, Hooper J, Dong Y, et al. Harvey, the expanded human kallikrein (KLK) gene family: genomic organisation, tissue-specific expression and potential functions. Biol Chem 2001;382:5–14.

[17] Olsson AY, Lundwall L. Organization and evolution of the glandular kallikrein locus in Mus musculus. Biochem Biophys Res Commun 2002;299:305–11.

[18] Ashley PL, MacDonald RJ. Tissue-specific expression of kallikrein-related genes in the rat. Biochemistry 1985;24:4520–7.

[19] Clements J. The molecular biology of the kallikreins and their roles in inflammation. In: Farmer S, editor. The kinin system. New York: Academic Press, 1997. p. 71–97.

[20] Stephenson SA, Verity K, Ashworth LK, et al. Localization of a new prostate-specific antigen-related serine protease gene, KLK4, is evidence for an expanded human kallikrein gene family cluster on chromosome 19q13.3–13.4. J Biol Chem 1999;274:23210–4.

[21] Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.

[22] Yousef GM, Diamandis EP. The new kallikrein-like gene, KLK-L2. Molecular characterization, mapping, tissue expression, and hormonal regulation. J Biol Chem 1999;274:37511–6.

[23] Karin M, Richards RI. Human metallothionien genes—Primary structure of the metallothionein-II gene and a related processed gene. Nature 1982;299:797–802.

[24] Scarpulla RC. Processed pseudogenes for rat cytochrome *c* are preferentially derived from one of three alternate mRNAs. Mol Cell Biol 1984;4:2279–88.

[25] Yousef GM, Ordon MH, Foussias G, et al. Genomic organization of the siglec gene locus on chromosome 19q13.4 and cloning of two new siglec pseudogenes. Gene 2002;286:259–70.

[26] McCarrey JR, Kumari M, Aivaliotis MJ, et al. Analysis of the cDNA and encoded protein of the human testis-specific PGK-2 gene. Dev Genet 1996;19:321–32.

[27] Fujii GH, Morimoto AM, Berson AE, et al. Transcriptional analysis of the PTEN/MMAC1 pseudogene, psiPTEN. Oncogene 1999;18:1765–9.

[28] Olsen MA, Schechter LE. Cloning, mRNA localization and evolutionary conservation of a human 5-HT7 receptor pseudogene. Gene 1999;227:63–9.