Identification of Novel Human Kallikrein-Like Genes on Chromosome 19q13.3 - q13.4

GEORGE M. YOUSEF[1,2], LIU-YING LUO[1,2] and ELEFTHERIOS P. DIAMANDIS[1,2]

[1]*Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Toronto, Ontario, M5G 1X5;*
[2]*Department of Laboratory Medicine and Pathobiology, University of Toronto, Ontario, M5G 1X5, Canada*

Abstract. *The human kallikrein gene family is localized on chromosome 19q13.3-q13.4 and currently includes three members: KLK1 or pancreatic/renal kallikrein, KLK2 or human glandular kallikrein and KLK3 or prostate-specific antigen (PSA). The latter two genes are almost prostate-specific and they are used for diagnosis and monitoring of prostate cancer and more recently, in breast cancer applications. In this paper, we analyzed a 300Kb genomic DNA region around chromosome 19q13.3 - q13.4 in an effort to map known kallikrein or kallikrein-like genes and identify new kallikrein-like genes. Using the known kallikrein or kallikrein-like genes PSA, KLK2, zyme and normal epithelial cell-specific 1 gene (NES1) as landmarks, we have identified another six novel genes of which, five have protein homologies and gene structure similarities with other kallikreins or kallikrein-like genes. We conclude, contrary to the current belief, that the human kallikrein gene locus contains a large number of kallikrein-like genes (at least thirteen). In this paper, we present a detailed description of the human kallikrein gene locus, encompassing the already known and newly identified genes. These new genes, like the already known kallikreins, may have utility for diagnosis, monitoring and therapeutics of various cancers including those of the breast, prostate and testis.*

Kallikreins and kallikrein-like proteins are a subgroup of the serine protease enzyme family and exhibit a high degree of substrate specificity [1]. The biological role of these kallikreins is the selective cleavage of specific polypeptide precursors (substrates) to release peptides with potent biological activity [2]. In mouse and rat, kallikreins are encoded by large multigene families. In the mouse genome, at least 24 genes have been identified [3]. Expression of 11 of these genes has been confirmed; the rest are presumed to be pseudogenes [4]. A similar family of 15-20 kallikreins has been found in the rat genome [5] where at least 4 of these are known to be expressed [6].

Three human kallikrein genes have been described, *i.e.* prostatic specific antigen (PSA or KLK3) [7], human glandular kallikrein (KLK2) [8] and tissue (pancreatic-renal) kallikrein KLK1) [9]. The PSA gene spans 5.8 Kb of sequence which has been published [7]; the KLK2 gene has a size of 5.2 Kb and its complete structure has also been elucidated [8]. The KLK1 gene is approximately 4.5 Kb long and the exon sequences and the exon/intron junctions of this gene have been determined [9].

The mouse kallikrein genes are clustered in groups of up to 11 genes on chromosome 7 and the distance between the genes in the various clusters can be as small as 3-7 Kb [3]. All three established human kallikrein genes have been assigned to chromosome 19q13.2 - 19q13.4 and the distance between PSA and KLK2 have been estimated to be 12 Kb [9].

A major difference between mouse and human kallikreins is that two of the human kallikreins (KLK2 and KLK3) are expressed almost exclusively in the prostate while in mouse, none of the kallikreins is localized in this organ. Other candidate new members of the human kallikrein gene family include protease M [10] (also named zyme [11] or neurosin [12]) and the normal epithelial cell-specific gene 1 (NES1) [13]. Both genes have been assigned

Non-Standard Abbreviations: PSA, prostate-specific antigen; hK2, human glandular kallikrein (KLK2); NES1, normal epithelial cell-specific 1 gene; Kb, kilobase; KLK, kallikrein; KLK-L, kallikrein-like; cM, centi-Morgan; EST, expressed sequence tag; RT-PCR, reverse transcription polymerase chain reaction; PAC, P1-derived artificial chromosome; BAC, bacterial artificial chromosome; TLSP, trypsin-like serine protease; HSCCE, human stratum corneum chymotryptic enzyme; FW, forward strand.

Correspondence to: Dr. E.P. Diamandis, Mount Sinai Hospital, Department of Pathology and Laboratory Medicine, 600 University Avenue, Toronto, Ontario M5G 1X5, Canada. Fax: 416-586-8628, E-Mail: ediamandis@mtsinai.on.ca

Key Words: Novel kallikrein-like genes, serine proteases, prostate-specific antigen, prostate cancer, breast cancer, gene identification, gene mapping, human kallikrein gene family.

0250-7005/99 $2.00+.40

Table I. *Exon or gene prediction programs used in this study*[1].

No.	Program name	Source	Website or e-mail address
1	GeneBuilder (gene prediction)	Institute of Advanced Biomedical Technologies	http://125.itba.mi.cnr.it/ ~ webgene/genebuilder.html
2	GeneBuilder(exon prediction)	Institute of Advanced Biomedical Technologies	http://125.itba.mi.cnr.it/ ~ webgene/genebuilder.html
3	ORF gene	Institute of Advanced Biomedical Technologies	http://125.itba.mi.cnr.it/ ~ webgene/wwworfgene2.html
4	GENEID-3	BioMolecular Engineering Research Center, Boston University	http://apolo.imim.es/geneid.html (geneid@darwin.bu.edu)
5	Grail 2	Oak Ridge National Laboratory	http://compbio.ornl.gov
6	FGENEH	Baylor College of Medicine,Houston, Texas	http://mcrb.bcm.tmc.edu

[1]. In the final analysis of the sequences we used programs 1, 2, 4 and 5 only.

to chromosome 19q13.3 [10,14] and show structural homology with other serine proteases as well as the kallikrein gene family [10-14]. The value of PSA for prostate cancer diagnostics is very well established. Human glandular kallikrein, protease M and NES1 have potential as new markers of breast and prostate cancer.

In our efforts to precisely define the relative genomic location of PSA, KLK2, zyme and NES1 genes, we studied an area spanning approximately 300 Kb of contiguous sequence on human chromosome 19 (19q13.3 -q13.4). We were able to identify the relative location of the known kallikrein genes and, in addition, we describe the discovery of other putative kallikrein-like genes which exhibit both location proximity and structural similarity with the known members of the human kallikrein gene family.

Materials and Methods

Identificatioi. of positive PAC and BAC genomic clones from a human genomic DNA library. The cDNA sequence of PSA, KLK1, KLK2, NES1 and zyme genes is already known. We have developed polymerase chain reaction (PCR)-based amplification protocols which allowed us to generate PCR products specific for each one of these genes. Using these PCR products as probes, labeled with ^{32}P, we screened a human genomic DNA PAC library and a human genomic DNA BAC library for the purpose of identifying positive clones of approximately 100-150 Kb long. The general strategies for these experiments have been published elsewhere [14]. Positive clones were further confirmed by Southern blot analysis as described [14].

DNA sequences on chromosome 19. Large sequencing information on chromosome 19 is available at the website of the Lawrence Livermore National Laboratory http: //www-bio.llnl.gov /genome/genome.html. We have obtained approximately 300 Kb of genomic sequences from that website, encompassing a region on chromosome 19q13.3-13.4, where the known kallikrein genes are localized. This 300 Kb of sequence is represented by 8 contigs of variable lengths. By using a number of different computer programs, we were able to construct an almost contiguous sequence of the region as shown diagramatically in Figure 1. Some of the contigs were reversed, as shown in Figure 1, in order to reconstruct the area on both strands of DNA.

By using the published sequences of PSA, KLK2, NES1 and zyme and the computer software BLAST 2, we were able, using alignment strategies, to identify the relative positions of these genes on the contiguous map (Figure 1). These known genes served as landmarks for further studies. An EcoR1 restriction map of the area is also available at the website of the Lawrence Livermore National Laboratory. Using this restriction map and the computer program WebCutter (http://www.firstmarket.com/cutter/cut2.html), we performed a restriction study analysis of the available sequence to further confirm the assignment and relative positions of these contigs along chromosome 19. The obtained configuration and the relative location of the known genes are presented in Figure 1.

Gene prediction analysis. For exon prediction analysis of the whole genomic area, we have used a number of different computer programs, listed in Table I. We have originally tested these programs using the known genomic sequences of the PSA, zyme and NES1 genes. The more reliable computer programs, GeneBuilder (gene prediction), GeneBuilder (exon prediction), Grail 2 and GENEID-3 were selected for further use.

Protein homology searching. Putative exons of the new genes were first translated to the corresponding aminoacid sequences. BLAST homology searching for the proteins encoded by the exons of the putative new genes was performed using the BLASTP program and the Genbank databases.

Results

Relative position of PSA, KLK2, Zyme and NES1 on Chromosome 19. Screening of the human BAC library identified two clones which were positive for the zyme gene

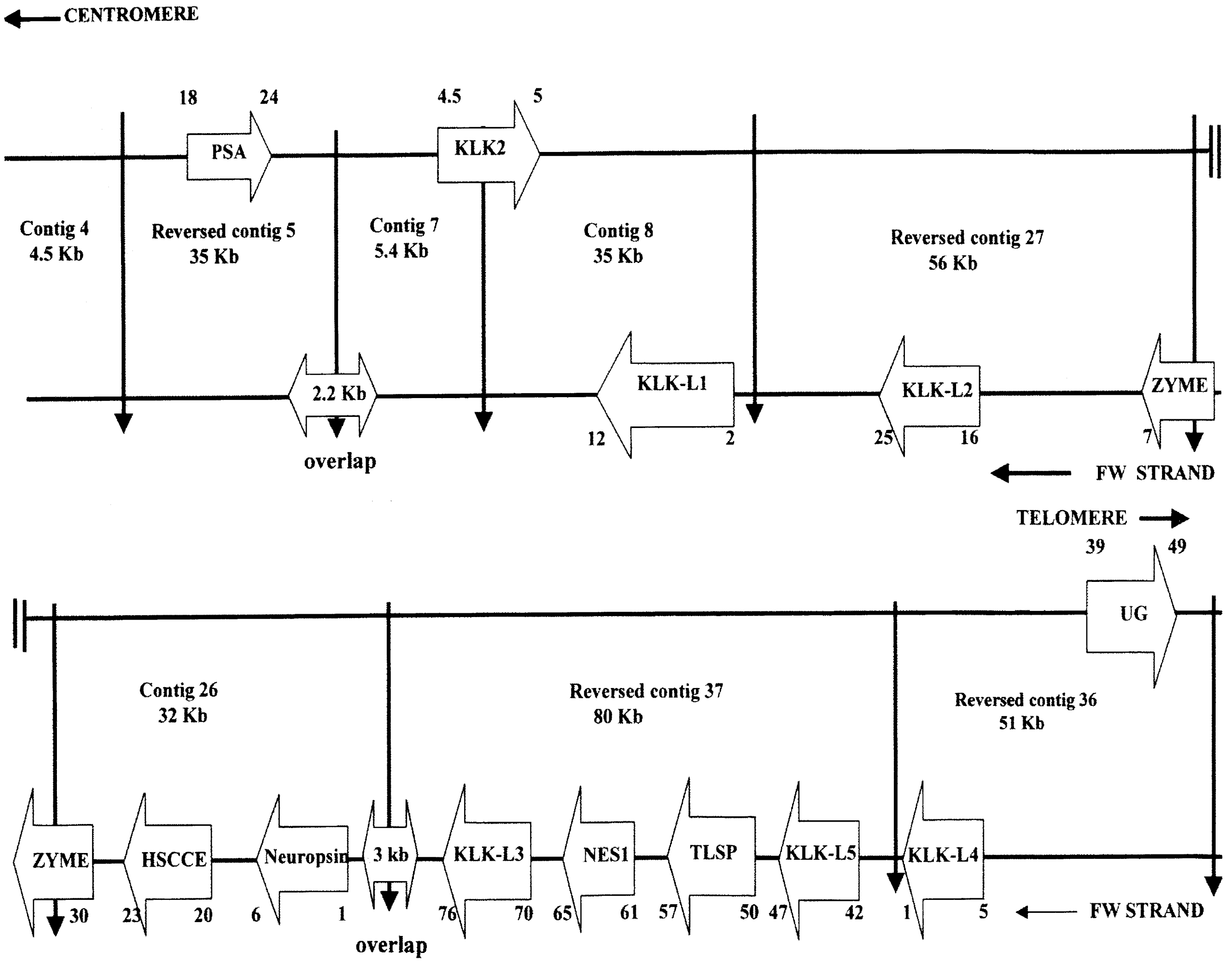

Figure 1. *An approximate 300 Kb of contiguous genomic sequence around chromosome 19q13.3 - q13.4 represented by 8 contigs, each one shown with its length in Kb. The contig numbers refer to those reported in the Lawrence Livermore National Laboratory website. Note the localization of the seven known genes (PSA, KLK2, Zyme, NES1, HSCCE, neuropsin and TLSP) (see abbreviations for full names of these genes). All genes are represented with arrows denoting the direction of transcription. The gene with no homology to human kallikreins is termed UG (unknown gene). Numbers just below or just above the arrows indicate approximate Kb lengths in each contig. The length of each of these genes may change in the future since not all exons were identified for each new gene, as shown in Tables II-VII.*

(clones BAC 288H1 and BAC 76F7). These BACs were further analyzed by PCR and primers specific for PSA, NES1, KLK1 and KLK2. These analyses indicated that both BACs were positive for zyme, PSA and KLK2 and negative for KLK1 and NES1 genes.

Screening of the human PAC genomic library identified a PAC clone which was positive for NES1 (PAC 43B1). Further PCR analysis indicated that this PAC clone was positive for NES1 and KLK1 genes and negative for PSA, KLK2 and zyme. Combination of this information with the EcoR1 restriction map of the region allowed us to establish the relative positions of these four genes. PSA is the most centromeric, followed by KLK2, zyme and NES1. Further alignment of the known sequences of these genes with the 300 Kb contig enabled us to precisely localize all four genes and determine the direction of transcription, as shown by the arrows in Figure 1. The KLK1 gene sequence was not identified on any of these contigs and appears to be further telomeric to NES1 (since it co-localized on the same PAC as NES1). We did not attempt to characterize the genomic position of the KLK1 gene.

Identification of new genes. We have used a set of arbitrary rules to consider presence of a new gene in the genomic area of interest, as follows:

1. Clusters of at least 3 exons should be found.

Table II. *Predicted exons of the putative gene KLK-L1. The translated protein sequences of each exon (open reading frame) are shown.*

Exon No.[1]	Putative coding region[2]		No. of bases	Translated protein sequence	EST match[3]	Intron phase[4]	Stop codon[5]	Catalytic triad[6]	Exon prediction program[7]
	From(bp)	To (bp)							
2	2263	2425	163	SLVSGSCSQIINGEDCSPHSQP WQAALVMENELFCSGVLVH PQWVLSAA**H**CFQ	+	II	–	H	A,B,D
3	2847	3109	263	NSYTIGLGLHSLEADQEPGSQ MVEASLSVRHPEYNRPLLAN**D** LMLIKLDESVSESDTIRSISIASQ CPTAGNSCLVSGWGLLAN	+	I	–	D	A,B,C,D
4	3180	3317	137	GRMPTVLQCVNVSVVSEEVCS KLYDPLYHPSMFCAGGGQDQ KDSCN	+	0	–	–	A,B,C,D
5	4588	4737	150	GD**S**GGPLICNGYLQGLVSFGKA PCGQVGVPGVYTNLCKFTEWIE KTVQAS	+	–	+	S	A,B,C

1. Conventional numbering of exons in comparison to the five coding exons of PSA,as described in Ref.14.
2. Nucleotide numbers refer to the related contig (see text and Figure 1).
3. (+) = >95% homology with published human EST sequences.
4. Intron phase: 0=the intron occurs between codons; I=the intron occurs after the first nucleotide of the codon; II=the intron occurs after the second nucleotide of the codon.
5. (+) denotes the exon containing the stop codon.
6. H=histidine, D=aspartic acid, S=serine.The aminoacids of the catalytic triad are bold and underlined.
7. A = GeneBuilder (gene analysis), B = GeneBuilder (exon analysis), C = Grail 2, D = GENEID-3

2. Only exons with high prediction score ("good" or "excellent" quality, as indicated by the searching programs) were considered for the construction of the putative new genes.
3. We considered the exons predicted as reliable only if they were identified by at least two different exon prediction programs.

By using this strategy, we identified nine putative new genes of which three were found by subsequent homology analysis to be known genes not previously mapped, i.e. the human stratum corneum chymotryptic enzyme (HSCCE), the human neuropsin and the human trypsin-like serine protease (TLSP). Their relative location is shown in Figure 1. In addition, we have identified one other new gene (gene UG) which showed no homology, at the protein level, with the kallikrein proteins. This gene has homology with the OB binding proteins 1 and 2 and with the surface antigen CD33. The five remaining genes all have significant homologies with known human or animal kallikrein proteins and/or other known serine proteases. We named these new genes as KLK-L1, KLK-L2, KLK-L3, KLK-L4 and KLK-L5 to underline their close relationship to the already known kallikreins (KLK-L = kallikrein-like).

In Tables II to VII, we present the preliminary exon structure and partial protein sequence for each one of the six newly identified genes. In Table 8, we present proteins which are homologous to the proteins encoded by the new genes. The genomic sequences and predicted exon sequences of the six newly identified genes have now been deposited in Genbank (Accession numbers: AF135023 for KLK-L1, AF135028 for KLK-L2, AF135026 for KLK-L3, AF135024 for KLK-L4, AF135025 for KLK-L5 and AF 135027 for the UG gene).

Discussion

Prediction of protein-coding genes in newly sequenced DNA becomes very important after the establishment of

Table III. *Predicted exons of the putative gene KLK-L2. The translated protein sequences of each exon (open reading frame) are shown*.*

Exon No.[1]	Putative coding sequence[2]		No. of bases	Translated protein sequence	EST match[3]	Intron phase[4]	Stop codon[5]	Catalytic triad[6]	Exon prediction program[7]
	From(bp)	To (bp)							
1	15,361	15,433	73	MATARPPWMWVLCALITAL LLGVT	+	I	–	–	–
2	17,904	18,165	262	EHVLANNDVSCDHPSNTVPSG SNQDLGAGAGEDARSDDSSSR IINGSDCDMHTQPWQAALLLR PNQLYCGAVLVHPQWLLTAA **H**CRKK	+	II	–	H	A,B,C,D
3	18,903	19,159	257	VFRVRLGHYSLSPVYESGQQMF QGVKSIPHPGYSHPGHSN**D**LML IKLNRRIRPTKDVRPINVSSHCPS AGTKCLVSGWGTTKSPQ	+	I	–	D	C,D
4	19,245	19,378	134	VHFPKVLQCLNISVLSQKRCEDA YPRQIDDTMFCAGDKAGRDSCQ	+	0	–	–	B,C
5	24,232	24,384	153	GD**S**GGPVVCNGSLQGLVSWGDY PCARPNRPGVYTNLCKFTKWIQE TIQANS	+	–	+	S	A,B,C

* All footnotes same as Table II.

Table IV. *Predicted exons of the putative gene KLK-L3. The translated protein sequences of each exon (open reading frame) are shown*.*

Exon No.[1]	Putative coding region[2]		No. of bases	Translated protein sequence	EST match[3]	Intron phase[4]	Stop codon[5]	Catalytic triad[6]	Exon prediction program[7]
	From(bp)	To (bp)							
1	70,473	70,584	112	MEEEGDGMAYHKEALDA GCTFQDP	–	I	–	–	A,B,C,D
2	70,764	70,962	199	ACSSLTPLSLIPTPGHGWAD TRAIGAEECRPNSQPWQAG LFHLTRLFCGATLISDRWLL TAA**H**CRK	+	II	–	H	A,B,C,D
3	73,395	73,687	293	PLTSEACPSRYLWVRLGEHH LWKWEGPEQLFRVTDFFPHP GFNKDLSANDHND**D**IMLIRL PRQARLSPAVQPLNLSQTCVS PGMQCLISGWGAVSSPK	+	I	–	D	A,B,C,D
4	76,305	76,441	137	ALFPVTLQCANISILENKLCH WAYPGHISDSMLCAGLWEG GRGSCQ	+	0	–	–	A,B,C,D
5	76,884	77633	749	GD**S**GGPLVCNGTLAGVVSGG AEPCSRPRRPAVYTSVCHYLD WIQEIMEN	–	–	+	S	A,B

* All footnotes same as Table II.

Table V. *Predicted exons of the putative gene KLK-L4. The translated protein sequences of each exon (open reading frame) are shown**.

Exon No.[1]	Putative coding region[2]		No. of bases	Translated protein sequence	EST match[3]	Intron phase[4]	Stop codon[5]	Catalytic triad[6]	Exon prediction program[7]
	From(bp)	To (bp)							
2	24,945	25,120	176	ESSKVLNTNGTSGFLPGGYT CFPHSQPWQAALLVQGRLLC GGVLVHPKWVLTAA**H**CLKE	+	II	–	H	C
3	25,460	25,728	269	GLKVYLGKHALGRVEAGEQ VREVVHSIPHPEYRRSPTHLN HDH**D**IMLLELQSPVQLTGYIQ TLPLSHNNRLTPGTTCRVSGW GTTTSPQ	+	I	–	D	A,B,C,D
4	26,879	27,015	137	VNYPKTLQCANIQLRSDEECR QVYPGKITDNMLCAGTKEGG KDSCE	+	0	–	–	A,B,C,D
5	28,778	28,963	189	GD**S**GGPLVCNRTLYGIVSWGD FPCGQPDRPGVYTRVSRYVLW IRETIRKYETQQQKWLKGPQ	+	–	+	S	A,B,C

* All footnotes same as Table II.

Table VI. *Predicted exons of the putative gene KLK-L5. The translated protein sequences of each exon (open reading frame) are shown**

Exon No.[1]	Putative coding region[2]		No. of bases	Translated protein sequence	EST match[3]	Intron phase[4]	Stop codon[5]	Catalytic triad[6]	Exon prediction program[7]
	From(bp)	To (bp)							
2	1588	1747	160	LSQAATPKIFNGTECGRNSQ PWQVGLFEGTSLRCGGVLID HRWVL TAA**H**CSG	–	II	–	H	A,B,C
3	3592	3851	260	SRYWVRLGEHSLSQLDWTEQ IRHSGFSVTHPGYLGASTSHEH **D**LRLLRLRLPVRVTSSVQPLPLP NDCATAGTECHVSGWGITNHPR	+	I	–	D	A,B,C,D
4	4806	4939	134	NPFPDLLQCLNLSIVSHATCHGV YPGRITSNMVCAGGVPGQDACQ	+	0	–	–	A,B,C,D

* All footnotes same as Table II.

large genome sequencing projects. This problem is complicated due to the exon-intron structure of the eukaryotic genes which interrupts the coding sequence in many unequal parts. In order to predict the protein-coding exons and overall gene structure, a number of computer programs were developed. All these programs are based on the combination of potential functional signals with the global statistical properties of known protein-coding regions [15]. However, the most powerful approach for gene structure prediction is to combine information about potential functional signals (splice sites, translation start or stop signal etc.) together with the statistical properties of

Table VII. *Predicted exons of the unknown gene UG . The translated protein sequences of each exon (open reading frame) are shown*.*

Exon No.	Putative coding region[1]		No. of bases	Translated protein sequence	EST match[2]	Intron phase[3]	Stop codon[4]	Catalytic program[5]
	From(bp)	To (bp)						
1	44,129	44,641	513	PPLSLEPAVPERRTLRNRRSLAALAPL TPDMLLLLLPLLWGRERAEGQTSKLL TMQSSVTVQEGLCVHVPCSFSYPSHG WIYPGPVVHGYWFREGANTDQDAPV ATNNPARAVWEETRDRFHLLGDPHTK NCTLSIRDARRSDAGRYFFRMEKGSIK WNYKHH RLSVNVT	+	I	–	B,C
2	44,843	45,121	279	ALTHRPNILIPGTLESGCPQNLTCSVPW ACEQGTPPMISWIGTSVSPLDPSTTRSSV LTLIPQPQDHGTSLTCQVTFPGASVTTN KTVHLNVS	+	I	–	A,B,C,D
3	45,327	45,374	48	YPPQNLTMTVFQGDGT	–	I	–	A,B,D
4	46,318	46,542	225	EGQSLRLVCAVDAVDSNPPARLSLSWR GLTLCPSQPSNPGVLELPWVHLRDAAE FTCRAQNPLGSQQVYLNVSLQ	+	I	–	A,B,C
5	47,195	47,283	186	SKATSGVTQGVVGGAGATALVFLSFC VIFV	+	0	–	A,B,C,D
6	49,136	49,554	186	GPLTEPWAEDSPPDQPPPASARSSVGE GELQYASLSFQMVKPWDS RGQEATD TEYSEIKIHR	+	–	+	A,B,C

* All footnotes same as Table II.

coding sequences (coding potential) along with information about homologies between the predicted protein and already known protein families [16].

In mouse and rat, kallikreins are encoded by large multigene families and these genes tend to cluster in groups with a distance as small as 3.3 - 7.0 Kb [3]. A strong conservation of gene order between human chromosome 19q13.1 - q13.4 and 17 loci in a 20-cM proximal part of mouse chromosome 7, including the kallikrein locus, has been documented [17].

In humans, only a few kallikrein genes were identified. In fact, only KLK1, KLK2 and KLK3 (PSA) are considered to represent the human kallikrein gene family [9,18]. In this paper, we provide strong evidence that a large number of kallikrein-like genes are clustered within a 300Kb region around chromosome 19q13.2 - q13.4. The three established human kallikreins (KLK1, KLK2, KLK3), zyme and NES1, as well as the stratum corneum chymotryptic enzyme, neuropsin and TLSP (trypsin-like serine protease) and another five new genes, KLK-L1 to KLK-L5, may constitute a large gene family. This will bring the total number of kallikrein or kallikrein-like genes in humans, in this region of chromosome 19, to thirteen.

The human stratum corneum chymotryptic enzyme [19], neuropsin [20] and trypsin-like serine protease (TLSP) [21] are three previously characterized genes which have many structural similarities with the kallikreins and other members of the serine protease family. However, they have not been mapped in the past. Our precise mapping of all three genes in the region of the kallikrein gene family further suggests that these genes, along with the ones that were newly identified by us, and the already known ones, constitute a family that likely originated by duplication of an ancestral gene. The relative localization of all these genes is depicted in Figure 1. We consider Figure 1 to describe the human kallikrein gene family, consisting of thirteen genes.

Kallikrein genes are a subfamily of serine proteases, traditionally characterized by their ability to liberate lysyl-bradykinin (kallidin) from kininogen [22]. More recently, however, a new, structural concept has emerged to describe kallikreins. From accumulated sequence data, it is now clear that the mouse has many genes with high homology to kallikrein coding sequences [23-24]. Richards and co-workers have contributed to the concept of a "kallikrein multigene family" to refer to these genes [25-26]. This definition is not based much on specific enzymatic function of the gene product, but more on its sequence homology and their close linkage on mouse chromosome 7. In humans, only KLK1 meets the functional definition of a kallikrein. KLK2 has trypsin-like enzymatic activity and KLK3 (PSA) has very weak chymotrypsin-like enzymatic activity. These activities of KLK2 and KLK3 are not known to liberate biologically active peptides from precursors. Based on the newer definition, members of the kallikrein family include, not only the gene for the kallikrein enzyme, but also genes encoding other homologous proteases, including the enzyme that processes the precursors of the nerve growth factor and epidermal growth factor [8]. Therefore, it is important to note the clear distinction between the enzyme kallikrein and a kallikrein or a kallikrein-like gene.

It is important to mention that the prediction of new genes by computer programs is still not a straightforward process. Many shortcomings are known to exist in such programs. Most of these programs are unable to detect non-coding exons and non-coding portions of exons. Some programs are unable to detect exons without the presence of a genomic context (when the regions adjacent to an exon are not present). Also, the power for detection of small exons (less than 100 bp) is low in some programs. About 5% of real splice sites are usually lost by some programs but over-prediction is usually small [27]. However, the detection power of some programs (e.g. Grail 2) is about 91% when tested with known genomic sequences. An indication of the quality of prediction is provided with these programs. In our study, we considered only exons which were predicted with "good" or "excellent" quality and only exons which were predicted by at least two different programs. Moreover, we considered the presence of a putative gene only when at least three exons clustered coordinately in that region. Additional evidence that these new genes are indeed homologous to the known kallikreins and other serine proteases comes from comparison of the intron phases. As we have published previously [14], trypsinogen, PSA and NES1 have 5 coding exons of which the first has intron phase I (the intron occurs after the first nucleotide of the codon), the second has intron phase II (the intron occurs after the second nucleotide of the

Table VIII. *Homology between the predicted amino acid sequences of the newly identified putative genes and protein sequences deposited in Genbank.*

No.	Gene identity	Homologous known protein	Identity% (number of amino acids)
1	KLK-L1	• Human stratum corneum chymotryptic enzyme	44 (101/227)
		• Rat kallikrein	40 (96/237)
		• Mouse glandular kallikrein K22	39 (94/236)
		• Human glandular kallikrein	38 (93/241)
		• Human prostatic specific antigen	37 (91/241)
		• Human protease M	37 (87/229)
2	KLK-L2	• Human neuropsin	48 (106/219)
		• Human stratum corneum chymotryptic enzyme	47 (103/216)
		• Human protease M	45 (99/219)
		• Human trypsinogen I	45 (100/221)
		• Rat trypsinogen	44 (98/220)
3	KLK-L3	• Human neuropsin	44 (109/244)
		• Rat trypsinogen 4	39 (95/241)
		• Human protease M	38 (98/253)
		• Human glandular kallikrein	37 (94/248)
		• Human prostatic specific antigen	36 (89/242)
4	KLK-L4	• Human protease M	52 (118/225)
		• Human neuropsin	51 (116/225)
		• Mouse neuropsin	51 (116/226)
		• Human glandular kallikrein	48 (113/234)
		• Human prostatic specific antigen	47 (108/227)
5	KLK-L5	• Human neuropsin	44 (81/184)
		• Rat trypsinogen I	42 (76/178)
		• Rat trypsinogen II	42 (75/178)
		• Human protease M	41 (73/178)
6	UG	• Human myeloid cell surface antigen CD33	61 (144/233)
		• Human OB binding protein-2	50 (166/328)
		• Human OB binding protein-1	43 (189/431)
		• Human myelin associated glycoprotein	27 (86/311)

codon), the third has intron phase I and the fourth has intron phase 0 (the intron occurs between codons). The fifth exon contains the stop codon. The intron phases of the predicted new kallikrein-like genes follow these rules and are shown in the respective tables. Further support comes from our identification in the new genes, of the conserved amino acids of the catalytic domain of the serine proteases, as presented in Tables II - VI.

In order to test the accuracy of the gene prediction programs, we tested known genomic areas containing the PSA, zyme and KLK2 genes. Two of these programs (Grail 2 and GeneBuilder) were able to detect about 95% of the tested known genes (data not shown). Matches with expressed sequence tag sequences (EST) can also be employed for gene structure prediction in the GeneBuilder program and this can significantly improve the power of the program, especially at high stringency (e.g. >95% homology). In the respective Tables for each putative new gene we provide evidence for matching ESTs from the Genbank human EST database. The presence of EST matches is additional strong evidence that these newly identified genes are expressed.

The question remains if these new genes are functional. In mouse, ten of the kallikrein genes appear to be pseudogenes [9]. One of our new genes (UG) does not show homology with the kallikrein genes. However, it has some protein homology with myelin associated glycoprotein (Table VIII). There may still be an association between UG and the kallikrein genes since some mouse kallikreins are related to nerve growth factor, as discussed earlier [8] and zyme as well as neuropsin and TLSP were found to be highly expressed in brain tissue and is claimed that zyme may be related to Alzheimer's disease [11].

We are now screening and sequencing EST clones and studing the tissue expression of these new genes by RT-PCR. Our goal is to fully characterize their mRNA sequence, study their expression and regulation and examine if they are involved, or can be used, like other human kallikreins (e.g. PSA, KLK2, zyme and NES1), in breast, prostate, testicular or other cancer diagnostic, prognostic or therapeutic applications. The expansion of the kallikrein locus in humans to thirteen genes will allow us to better understand the role of this family in various cancers. There is already evidence that some of these genes encode for tumor suppressors [10, 28, 29].

Acknowledgements

We thank Dr. Stephen Scherer for his valuable discussions and advice.

References

1 Evans BAE, Yun ZX, Close JA, Tregear GW, Kitamura N, Nakanishi S, *et al:* Structure and chromosomal localization of the human renal kallikrein gene. Biochemistry *27:* 3124-3129, 1988.

2 Clements JA: The glandular kallikrein family of enzymes: Tissue-specific expression and hormonal regulation. Endocr Rev *10:* 393-419, 1989.

3 Evans BA, Drinkwater CC, Richards RI: Mouse glandular kallikrein genes: structure and partial sequence analysis of the kallikrein gene locus. J Biol Chem *262:* 8027-8034, 1987.

4 Drinkwater CC, Evans BA, Richards RI: Kallikreins, kinins and growth factor biosynthesis. Trends Biochem Sci *13:* 169-172, 1988b.

5 Ashley PL, MacDonald RJ: Tissue-specific expression of kallikrein-related genes in the rat. Biochemistry *24:* 4520-5427, 1985.

6 Gerald WL, Chao J, Chao L: Sex dimorphism and hormonal regulation of rat tissue kallikrein mRNA. Biochim Biophys Acta *867:* 16-23, 1986.

7 Riegman PHJ, Vlietstra RJ, van der Korput JAGM, Romijn JC, Trapman J: Characterization of the prostate-specific antigen gene: a novel human kallikrein-like gene. Biochem Biophys Res Commun *159:* 95-102, 1989.

8 Schedlich LJ, Bennetts BH, Morris BJ: Primary structure of a human glandular kallikrein gene. DNA *6:* 429-437, 1987.

9 Riegman PH, Vlietstra RJ, Suurmeijer L, Cleutjens CBJM, Trapman J: Characterization of the human kallikrein locus. Genomics *14:* 6-11, 1992.

10 Anisowicz A, Sotiropoulou G, Stenman G, Mok SC, Sager R: A novel protease homolog differentially expressed in breast and ovarian cancer. Mol Med *2:* 624-636, 1996.

11 Little SP, Dixon EP, Norris F, Buckley W, Becker GW, Johnson M, et al: Zyme, a novel and potentially amyloidogenic enzyme cDNA isolated from Alzheimer's disease brain. J Biol Chem *272:* 25135-25142, 1997.

12 Yamashiro K, Tsuruoka N, Kodama S, Tsujimoto M, Yamamura Y, Tanaka T, *et al:* Molecular cloning of a novel trypsin-like serine protease (neurosin) preferentially expressed in brain. Biochim Biophys Acta *1350:* 11-14, 1997.

13 Liu XL, Wazer DE, Watanabe K, Band V: Identification of a novel serine protease-like gene, the expression of which is down-regulated during breast cancer progression. Cancer Res *56:* 3371-3379, 1996.

14 Luo L, Herbrick J-A, Scherer SW, Beatty B, Squire J, Diamandis EP: Structural characterization and mapping of the normal epithelial cell-specific 1 gene. Biochem Biophys Res Commun *247:* 580-586, 1998.

15 Milanesi L, Kolchanov N, Rogozin I, Kel A, Titov I: Sequence functional inference. *In:* "Guide to human genome computing", ed. M.J. Bishop, Academic Press, Cambridge, 249-312, 1994.

16 Burset M, Guigo R: Evaluation of gene structure prediction programs. Genomics *34:* 353-367, 1996.

17 Nadeau J, Grant P, Kosowsky M: Mouse and human homology map. Mouse Genome *89:* 31-36, 1991.

18 Rittenhouse HG, finlay JA, Mikolajczyk, Partin AW: Human Kallikrein 2 (hK2) and prostate-specific antigen (PSA): two closely related, but distinct, kallikreins in the prostate. Crit Rev Clin Lab Sci *35:* 275-268, 1998.

19 Hansson L, Stromqvist M, Backman A, Wallbrandt P, Carlstein A, Egelrud T: Cloning, expression and characterization of stratum corneum chymotryptic enzyme. A skin-specific human serine proteinase. J Biol Chem *269:* 19420-19426, 1994.

20 Yoshida S, Taniguchi M, Hirata A, Shiosaka S: Sequence analysis

and expression of human neuropsin cDNA and gene. Gene *213:* 9-16, 1998.

21 Yoshida S, Taniguchi M, Suemoto T, Oka T, He X, Shiosaka S: cDNA cloning and expression of a novel serine protease, TLSP. Biochem Biophys Acta *1399:* 225-228, 1998.

22 Schachter M: Kallikreins (kininogenases) - a group of serine proteases with bioregulatory actions. Pharmacol Rev *31:* 1-17, 1980.

23 Morris BJ, Catanzaro DF, Richards RI, Mason AJ, Shine J: Kallikrein and renin: molecular biology and biosynthesis. Clin Sci *61:* 351s-353s, 1981.

24 Richards RI, Catanzaro DF, Mason AJ, Morris BJ, Baxter JD, Shine J: Mouse glandular kallikrein genes. Nucleotide sequence of cloned cDNA coding for a member of the kallikrein arginyl esteropeptidase group of serine proteases. J Biol Chem *257:* 2758-2761, 1982.

25 Van Leeuwen BH, Evans BA, Tregear GW, Richards RI: Mouse glandular kallikrein genes. Identification, structure and expression of the renal kallikrein gene. J Biol Chem *261:* 5529-5535, 1986.

26 Evans BA, Richards RI: Genes for the a and g subunits of mouse nerve growth factor. EMBO J *4:* 133-138, 1985.

27 Rogozin IB, Milanesi L, Kolchanov NA: Gene structure prediction using information on homologous protein sequence. Comput Applic Biosci *12:* 161-170, 1996.

28 Diamandis EP, Yu H: New biological functions of prostate specific antigen? J Clin Endocrinol Metab *80:* 1515-1517, 1995.

29 Goyal J, Smith KM, Cowan JM, Wazer DE, Lee SW, Band V: The role of NES1 serine protease as a novel tumor suppressor. Cancer Res *58:* 4782-4786, 1998.

Received May 14, 1999
Accepted May 27, 1999