LATEST PAPERS SEARCH for PAPERS Printer-friendly PDF Comment(s)

>Abstract >Introduction >Results >Discussion >References >Materials & methods >Supplemental data >Contact authors

Cancer Immunity, Vol. 4, p. 2 (11 February 2004) Submitted: 5 November 2003. Resubmitted: 15 January 2004. Accepted: 27 January 2004.
Communicated by: LJ Old

Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues

Yves Stauffer[1], Gregory Theiler[1], Peter Sperisen[2], Yuri Lebedev[3], and C. Victor Jongeneel[1]

[1]Ludwig Institute for Cancer Research, Office of Information Technology, Epalinges, Switzerland
[2]Swiss Institute of Bioinformatics, Epalinges, Switzerland
[3]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Russia

Keywords: human, neoplasms, HERVs, gene expression profiling, ESTs, massively parallel signature sequencing

Abstract

Human endogenous retroviruses (HERVs) are remnants of ancient retroviral infections that became fixed in the germ line DNA millions of years ago. The fact that humoral and cellular immune responses against HERV-encoded proteins have been identified in cancer patients suggests that these antigens might be used in cancer immunotherapy or diagnosis. We analyzed the digital expression patterns of the HERV-K (HML-2), -W, -H and -E families in normal and cancerous tissues. Thirty-one proviral members of the HERV-K family and one representative each for the other HERV families were used as probes to search human EST data. Matching of HERV proviruses to ESTs was HERV family-specific and the expression profiles of the HERV families distinct. The HERV-K family was expressed in normal tissues such as muscle, skin and brain, as well as in germ cell tumors and other cancerous tissues. HERV-H was the only family expressed in cancers of the intestine, bone marrow, bladder and cervix, and was more highly expressed than the other families in cancers of the stomach, colon and prostate. In contrast, HERV-W was predominantly expressed in normal placenta. Expression patterns were confirmed by MPSS (massively parallel signature sequencing) data where available. For the HERV-K family, we mapped most ESTs to their corresponding proviruses and assessed the coding capacities of the matched proviruses. This study shows that HERV families are more widely expressed than originally thought and that some members of the HERV-K and -H families could encode targets for cancer immunotherapy.

Introduction

Human endogenous retroviruses originate from ancient retroviral infections that became fixed in the germ line DNA between <1 and 40 million years ago and represent approximately 1% of the human genome (1, 2). So far, about twenty HERV families, named according to the transfer RNA used to prime reverse transcription, have been identified (3). Most HERV families are present in multiple copies in the genome and all those examined to date are defective due to the accumulation of mutations affecting their coding potential (4). However, some families still contain open reading frames (ORFs) for one or more of the retroviral genes. HERVs have simple genomes containing *gag*, *pol* and *env* genes encoding retroviral polyproteins, flanked by two long terminal repeats (LTRs). The 5' LTR contains enhancer and promoter sequences that provide signals recognized by the cellular machinery for transcription initiation and the 3' LTR provides a polyadenylation signal. Typically, two polyadenylated transcripts are produced: a full-length transcript, used for the synthesis of new viral genomes and as mRNA for the *gag* and *pol* genes, and a spliced transcript, used as mRNA for the *env* gene. A HERV provirus could therefore be considered as a single retroviral gene containing different translational units. Translation of the different ORFs present on the full-length transcript is regulated and involves some nonsense and frameshift suppression events to produce about 20-fold more Gag than Pol protein products. These polyproteins are then cleaved by the viral protease (5).

The HERV-K (HML-2) family is present as 30-50 proviral copies in the human genome (6) and is the most conserved HERV family, since some members of this family have intact open reading frames for the *gag*, *pol* or *env* retroviral genes (7). Moreover, members of the HERV-W and -H families have been shown to encode intact Env proteins (8, 9). HERVs are transcriptionally silent in most normal human tissues (10). However, HERVs have been found to be expressed in some normal tissues, such as placenta, and under pathological conditions (8, 11, 12). For instance, the HERV-K family has been reported to be expressed in teratocarcinoma and breast cancer cell lines (13, 14), HERV-E in prostate carcinoma (15), HERV-H in leukemia cell lines (16) and HERV-W in normal placenta and brain tissues from multiple sclerosis or schizophrenic patients (8, 17). There is no evidence, however, that HERVs are directly implicated in carcinogenesis.

On the other hand, antibody responses against HERV-K proteins have been observed in patients with germ cell tumors (18, 19). In addition, antibodies reactive against cDNA clones encoding HERV-K Gag or Env were identified in the sera of testicular, melanoma and prostate cancer patients using the SEREX (serological analysis of recombinant cDNA expression libraries) methodology, demonstrating that a humoral response was mounted against these proteins (SEREX sequence IDs 1630, 1631, 92, 289 and 2312). Schiavetti *et al*. recently identified a short HERV-K ORF encoding an antigen that elicited a CTL response in melanoma patients (20). The fact that HERV-encoded proteins have been found to be able to elicit humoral and/or T-cell mediated responses in cancer patients suggests that these proteins could be a source of antigens for use in cancer immunotherapy or diagnosis.

In order to identify additional HERV antigen candidates in different malignancies, the expression patterns of each HERV family need to be analyzed and the coding capacity of differentially expressed transcripts assessed. In this study, we analyzed digital expression patterns of four HERV families, HERV-K, -H, -W and -E, for which proviruses containing at least one full-length viral open reading frame have been described. When available, we compared expression profiles obtained by EST analysis with those based on MPSS of a set of 31 normal tissues and 3 cancer cell lines. MPSS is a technology that allows the generation of millions of signature tags proximal to the 3' end of transcripts, sufficient to cover cellular transcripts up to 10-fold (21, 22). Moreover, since most HERV-K proviruses have recently been identified, we could assign most ESTs to their corresponding proviruses and assess their expression levels and coding capacities.

Results

Mapping HERV proviral sequences to the genome

We mapped HERV proviruses onto the human genome based on previously published analyses (23, 24, 25, 26) and using the lalign (27) and bl2seq (28) pairwise alignment tools with known family members. Only proviruses that were not extensively degenerate or truncated were selected for the analysis of their expression patterns. Table 1 shows the list of analyzed proviruses and their mapping to human genomic sequences and Ensembl chromosomes (based on the NCBI 33 assembly of the human genome).

EST-based expression profiles of HERV families

In order to analyze the expression patterns of the four less degenerate HERV families, two query sets containing proviral sequences representative of each HERV family were created. Since only a few proviral sequences have been described for the HERV-H, -W and -E families, one representative proviral member for each of these families was included in the query sets. The first query set contained these sequences as well as thirty-one HERV-K proviral sequences, thus covering most of the HERV-K family. For comparison purposes, the second query set contained HERV-K108 as the representative of the HERV-K family. This HERV-K provirus was selected because it contains large retroviral ORFs. LTR sequences were removed from proviral sequences prior to inclusion in the query sets in order to avoid matching solo LTRs lacking adjacent retroviral sequences.

These query sets were then used to search the human EST data using Megablast (29). Blast results were sorted according to matched ESTs and only HERV proviruses that matched ESTs with the best alignment score were kept. ESTs matching Alu sequences inserted in the proviruses HERV-K 22q11, 19p13_11B, 6p21 and 9q34_3 were excluded from the analysis. Moreover, in order to confirm that ESTs stemmed from the matched HERV-K provirus, ESTs were searched against the human genome (NCBI build 33) using BLAT (BLAST-like alignment tool) and the chromosome coordinates compared to the ones obtained for the HERV-K proviruses present in the query set. A majority of ESTs (95/143) matched the corresponding HERV-K provirus unambiguously, 6 ESTs matched more than one HERV-K provirus in the query set with identical BLAT scores, while 42 ESTs matched yet uncharacterized HERV-K provirus loci with a better BLAT score (see Supplementary Table 4). The total number of matched ESTs from non-normalized libraries is listed in Table 1 for each provirus and HERV family, respectively.

The HERV-K C3_NT005863 provirus, corresponding to the transcriptionally active HERV-K(II) described by Sugimoto *et al*. (30) and the HERV-K 22q11 provirus, a provirus encoding a Gag protein identified by Y. Obata using the SEREX methodology in a prostate cancer patient, are the most expressed members of the HERV-K family in our analysis. As expected from the sequence divergence between the four HERV families, EST matches were always HERV family-specific (i.e. ESTs were matched only by members of the same HERV family). Since there was no competition for the matching of the HERV-W, -H and -E proviruses to their family-specific ESTs, we assumed that the expression of the member selected represented the expression pattern of the entire family.

Information about the tissue of origin and its status (cancerous or normal) was then retrieved for each matched EST based on the classification of the corresponding cDNA library using a controlled hierarchical ontology (eVOC) (31). Based on this information, we sorted the ESTs into 32 different tissue categories (see Supplementary Table 1). The list of normal and cancerous tissues in which HERV families are expressed is shown in Table 2. Only ESTs from non-normalized cDNA libraries were taken into account. For each HERV family, tissues are sorted from the highest to the lowest relative expression, based on the ratio of the number of matched ESTs to the total number of ESTs sequenced for the tissue.

Table 1. Mapping of HERV proviral sequences to the human genome.

Provirus	EST[a]	Chr.	Accession[b]	Start	Stop	S[c]	Length	Chr. Start[d]	Chr. Stop	S[c]	Prov. Ac.[e]	Comment	Alt. Name[f]	Ref.[g]
HERV-K														
C3_NT005863.11	32	3	AC084198	94740	103862	+	9123	102691966	102701086	+	AB047240		HERV-K(II)	30
22q11[h]	25	22	NT_011520.9	3270496	3281184	+	10686	22204481	22215170	+				24, 25
K102	11	1	AL353807	157140	166319	-	9178	152374332	152383510	-	AF164610			23
K101	5	22	NT_011519.10	2078337	2087511	+	9179	17300741	17309916	+	AF164609			23
K108	3	7	AC072054	29447	38914	-	9467	4332396	4341866	-	AF164614			23
8q24	3	8	AF235103	13722	21316	+	7594	145851347	145858940	-		truncated provirus		24
19q13.1	3	19	AC012309	19871	29389	-	9518	42273599	42283115	-				24
C3_NT022411.11	2	3	AC0922903	143424	152483	-	9112	126890913	126900025	+	AB047209		HERV-K(I)	30
8p23	2	8	AC068020	83065	92586	+	9522	7925735	7935257	+				24
10p14	2	10	AL392086	32893	42355	-	9463	7016381	7025842	-				24
12q24	2	12	AC026786	106398	112339	+	5940	133292096	133298037	-		truncated gag-pol (del 3240bp)		24
K18.2	1	1	AL121985	78059	87290	+	9234	157438449	157447681	+	Y18890			39
K10	1	5	AC016577	20931	30110	-	9180	156020261	156029439	-	M14123			6
9q34.11	1	9	AL441992	150822	158043	+	7221	125066185	125073406	+		truncated LTR		24
11q22.1	1	11	AP000776	101081	110546	+	9466	101599871	101609337	+				24
12q14	1	12	AC025420	37159	46625	+	9466	58437980	58447445	-				24
1q23	0	1	AC015623	152324	157979	+	5656	163276291	163270635	-	AL611962	truncated provirus		24
1p36	0	1	AL365443	85732	95255	-	9522	12695498	12705020	-				24
3p25	0	3	AC018809	124537	131427	+	6890	9881193	9888082	-		truncated provirus		24
3q27	0	3	AC069420	61265	70444	-	9180	186682364	186691542	-				24
C3_AC078785	0	3	AC078785	5275	14433	-	9158	114024354	114033511	-			HERVKC3	38
K104	0	5	AC025757	13109	114122	-	9445	30486349	30495793	-	AF164612			23
6p21[h]	0	6	AL035587	47775	57735	+	9961	42862823	42872782	-				24
6p22	0	6	AL671879	109025	119389	+	10363	28712907	28723272	+				24
K109	0	6	AL590785	33991	43412	-	9422	78376819	78386239		AF164615			23
K115	0	8	AC134684	55950	65412	-	9463	N/A			AY037929	insertion present in 15% human		37
9q34.3[h]	0	9	AL355987	58799	68261	-	9463	133115833	133125294	-				24
HERV-K(C11a)	0	10	AC015686	157484	166945	+	9462	7016381	7025842	-		same sequence as 10p14		38
11q23	0	11	AP000831	6156	15315	-	9160	118625638	118634796	-				24
19p13.11B[h]	0	19	AC078899	61897	72009	+	10113	20232609	20242722	+				24
K113	0	19	N/A				9472	N/A			AY037928	insertion present in 30% human		37
K103	0	N/A	N/A				9181	N/A			AF164611			23
HERV-K tot	143													
HERV-W	98	7	AC000064	28068	38289	-	10222	91695512	91705734	+				8
HERV-H	152	2	AC009495	789	9575	-	8787	166527996	166536785	-	AJ289709		HERV-H/env62	26
HERV-E	35	19	AC010329	96143	104955	-	8813	20705675	20714487	-	M10976		HERV-E (4-1)	40

[a] Number of ESTs from non-normalized, non-subtracted libraries matching the HERV-K provirus or the HERV family.

[b] Accession number for the BAC or contig sequence that contains the HERV proviral sequence.

[c] Indicates if the HERV provirus is in the sense (+) or complementary strand of the DNA sequence.

[d] Start and stop coordinates of the HERV provirus on the chromosome (EMBL Build 33).

[e] Accession number for the proviral sequence.

[f] Alternative name for the HERV provirus as published in the corresponding reference.

[g] Reference(s) reporting the HERV provirus.

[h] Presence of an inserted Alu sequence in the HERV provirus.

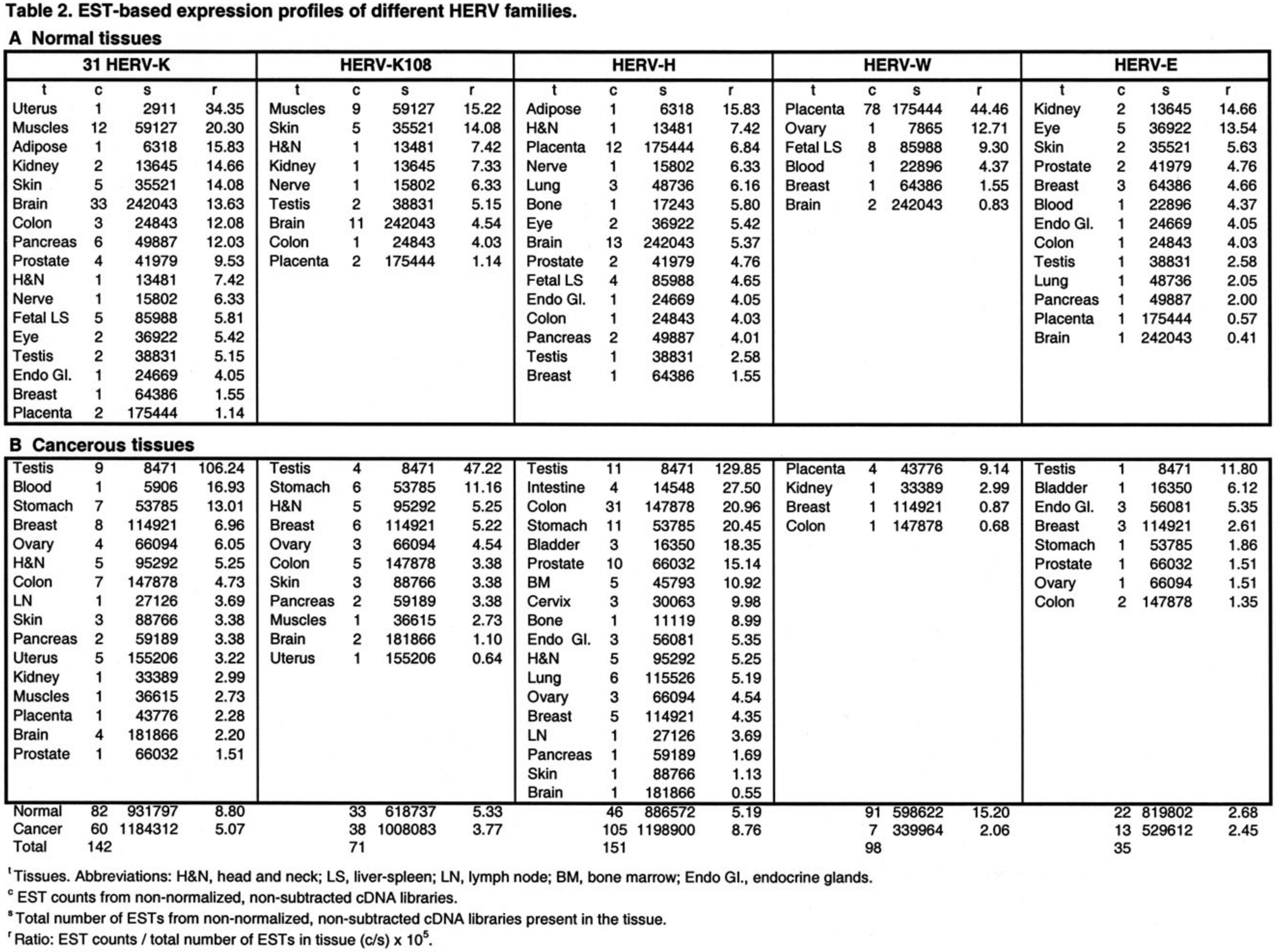


Table 2. EST-based expression profiles of different HERV families.

A Normal tissues

31 HERV-K				HERV-K108				HERV-H				HERV-W				HERV-E			
t	c	s	r	t	c	s	r	t	c	s	r	t	c	s	r	t	c	s	r
Uterus	1	2911	34.35	Muscles	9	59127	15.22	Adipose	1	6318	15.83	Placenta	78	175444	44.46	Kidney	2	13645	14.66
Muscles	12	59127	20.30	Skin	5	35521	14.08	H&N	1	13481	7.42	Ovary	1	7865	12.71	Eye	5	36922	13.54
Adipose	1	6318	15.83	H&N	1	13481	7.42	Placenta	12	175444	6.84	Fetal LS	8	85988	9.30	Skin	2	35521	5.63
Kidney	2	13645	14.66	Kidney	1	13645	7.33	Nerve	1	15802	6.33	Blood	1	22896	4.37	Prostate	2	41979	4.76
Skin	5	35521	14.08	Nerve	1	15802	6.33	Lung	3	48736	6.16	Breast	1	64386	1.55	Breast	3	64386	4.66
Brain	33	242043	13.63	Testis	2	38831	5.15	Bone	1	17243	5.80	Brain	2	242043	0.83	Blood	1	22896	4.37
Colon	3	24843	12.08	Brain	11	242043	4.54	Eye	2	36922	5.42					Endo Gl.	1	24669	4.05
Pancreas	6	49887	12.03	Colon	1	24843	4.03	Brain	13	242043	5.37					Colon	1	24843	4.03
Prostate	4	41979	9.53	Placenta	2	175444	1.14	Prostate	2	41979	4.76					Testis	1	38831	2.58
H&N	1	13481	7.42					Fetal LS	4	85988	4.65					Lung	1	48736	2.05
Nerve	1	15802	6.33					Endo Gl.	1	24669	4.05					Pancreas	1	49887	2.00
Fetal LS	5	85988	5.81					Colon	1	24843	4.03					Placenta	1	175444	0.57
Eye	2	36922	5.42					Pancreas	2	49887	4.01					Brain	1	242043	0.41
Testis	2	38831	5.15					Testis	1	38831	2.58								
Endo Gl.	1	24669	4.05					Breast	1	64386	1.55								
Breast	1	64386	1.55																
Placenta	2	175444	1.14																

B Cancerous tissues

31 HERV-K				HERV-K108				HERV-H				HERV-W				HERV-E			
Testis	9	8471	106.24	Testis	4	8471	47.22	Testis	11	8471	129.85	Placenta	4	43776	9.14	Testis	1	8471	11.80
Blood	1	5906	16.93	Stomach	6	53785	11.16	Intestine	4	14548	27.50	Kidney	1	33389	2.99	Bladder	1	16350	6.12
Stomach	7	53785	13.01	H&N	5	95292	5.25	Colon	31	147878	20.96	Breast	1	114921	0.87	Endo Gl.	3	56081	5.35
Breast	8	114921	6.96	Breast	6	114921	5.22	Stomach	11	53785	20.45	Colon	1	147878	0.68	Breast	3	114921	2.61
Ovary	4	66094	6.05	Ovary	3	66094	4.54	Bladder	3	16350	18.35					Stomach	1	53785	1.86
H&N	5	95292	5.25	Colon	5	147878	3.38	Prostate	10	66032	15.14					Prostate	1	66032	1.51
Colon	7	147878	4.73	Skin	3	88766	3.38	BM	5	45793	10.92					Ovary	1	66094	1.51
LN	1	27126	3.69	Pancreas	2	59189	3.38	Cervix	3	30063	9.98					Colon	2	147878	1.35
Skin	3	88766	3.38	Muscles	1	36615	2.73	Bone	1	11119	8.99								
Pancreas	2	59189	3.38	Brain	2	181866	1.10	Endo Gl.	3	56081	5.35								
Uterus	5	155206	3.22	Uterus	1	155206	0.64	H&N	5	95292	5.25								
Kidney	1	33389	2.99					Lung	6	115526	5.19								
Muscles	1	36615	2.73					Ovary	3	66094	4.54								
Placenta	1	43776	2.28					Breast	5	114921	4.35								
Brain	4	181866	2.20					LN	1	27126	3.69								
Prostate	1	66032	1.51					Pancreas	1	59189	1.69								
								Skin	1	88766	1.13								
								Brain	1	181866	0.55								
Normal	82	931797	8.80		33	618737	5.33		46	886572	5.19		91	598622	15.20		22	819802	2.68
Cancer	60	1184312	5.07		38	1008083	3.77		105	1198900	8.76		7	339964	2.06		13	529612	2.45
Total	142				71				151				98				35		

[t] Tissues. Abbreviations: H&N, head and neck; LS, liver-spleen; LN, lymph node; BM, bone marrow; Endo Gl., endocrine glands.

[c] EST counts from non-normalized, non-subtracted cDNA libraries.

[s] Total number of ESTs from non-normalized, non-subtracted cDNA libraries present in the tissue.

[r] Ratio: EST counts / total number of ESTs in tissue (c/s) x 10^5.

These results suggest that HERV families have distinct expression profiles and are expressed in both normal and cancerous tissues. The overall level of expression is similar for the HERV-K, -H and -W families, but is lower for the HERV-E family. Although their level of expression is relatively low in most tissues, as suggested by the low number of ESTs identified, certain HERV families are more highly expressed in some tissues. For example, the HERV-W family is highly expressed in normal placenta (78 ESTs), consistent with the fusiogenic function recently described for its envelope protein in human placenta (32, 33). The HERV-H and -K families are highly expressed in testicular cancers but are also expressed in different normal and cancerous tissues. The HERV-H family is particularly well represented in cancerous tissues, such as colon, stomach and prostate.

HERV-K expression in testicular cancer (see Table 2B) was confirmed by the fact that about 70% (41/58) of matched ESTs from subtracted libraries and 47% (7/15) from normalized libraries were generated from germ cell tumors. Moreover, a majority of these ESTs matched the HERV-K 22q11 (22 ESTs, of which 18 were from subtracted cDNA libraries) or HERV-K101 (16 ESTs, of which 14 were from subtracted cDNA libraries) proviruses (see Supplementary Table 2). Preferential HERV-W expression in normal placenta was also confirmed since this tissue represented 96% (45/47) and 60% (3/5) of matched ESTs from normalized and subtracted libraries, respectively. Fewer ESTs from these normalized or subtracted cDNA libraries were identified for the HERV-H (13 ESTs) and -E (22 ESTs) families. Nevertheless, expression patterns in these cDNA libraries were similar to the ones observed in non-normalized libraries.

Assessment of the homogeneity of HERV-K provirus expression patterns and HERV family-specific matching to ESTs

In order to determine the level of similarity between the expression profiles of the 31 HERV-K proviruses and to confirm that matching to ESTs was family-specific, all HERV provirus sequences were aligned to all ESTs matched by the HERV-K family. Normalized alignment scores and binary distances were computed as described in Materials and Methods and proviruses were clustered based on these binary distances. As shown in Figure 1, expression patterns were relatively similar, though not identical, for the different proviral members of the HERV-K family. Most of the variability in the HERV-K provirus expression patterns was indeed attributable to truncated HERV-K proviruses that failed to match some ESTs (HERV-K proviruses on the left side of Figure 1). In addition, this analysis provides experimental evidence that matching to EST sequences is actually HERV family-specific, since proviruses from the HERV-W, -H and -E families failed to match HERV-K family-specific ESTs.

Comparison of the expression levels of the different HERV families in normal and cancerous tissues

As shown in Table 2, the number of ESTs present in normal and cancerous tissues for the HERV families analyzed were obtained for two query sets containing one representative for each of the HERV-H, -W and -E families and either thirty-one members of the HERV-K family or the HERV-K108 provirus (representative of the HERV-K family). In order to compare the expression levels of the different HERV families in each tissue, these observed frequencies were used to build two contingency tables. For each HERV family, expected frequencies and chi square values were then calculated for each tissue as described in Materials and Methods. The results for the two query sets are shown in Table 3 (A and B, respectively). Normal tissues were sorted according to their contribution to the total chi square value, while cancerous tissues were listed in same order as normal tissues for comparison purposes. Moreover, tissues for which HERV expression was observed in only one tissue state (normal/cancer) were grouped in the separate "Others Norm." (adipose, nerve, fetal liver-spleen, eye and bone) or "Others Canc." (stomach, lymph node, intestine, bone marrow, bladder and cervix) categories, respectively.

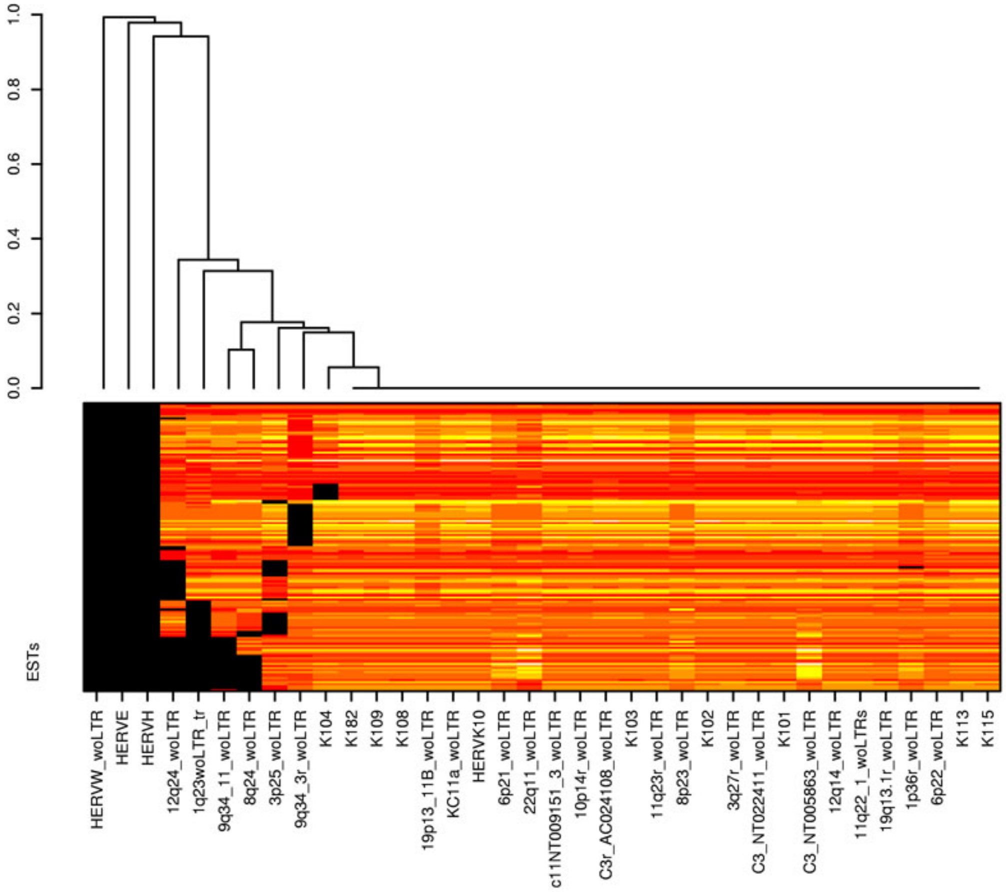


Figure 1. Clustering HERV proviruses according to their expression patterns. All EST sequences matched by the HERV-K family were aligned to all HERV proviruses present in the query set (31 HERV-K proviruses and 1 representative member for each of the HERV-H, W and E family). Alignment scores were normalized to obtain bit scores. A color-code ranging from red (low) to white (high) was used to represent bit scores (lower panel). Bit scores lower than the threshold of 36 are represented in black. HERV proviruses were clustered, using average-linkage agglomerative hierarchical clustering, based on the binary distances between pairs of proviral sequences (upper panel).

Differential expression of a HERV family in a tissue was inferred by comparing the observed and expected frequencies of the different HERV families. A HERV family was considered differentially expressed in a tissue if its observed frequency was at least two-fold higher than its expected frequency, with observed frequencies for the other HERV families being equal or lower than expected. Other parameters were taken into account, such as the contribution of the tissue to the total chi square value and, for the HERV-K family, the correlation between the two query sets.

Table 3. Comparison of expression levels of different HERV families in normal and cancerous tissues.

A

	ESTs[a]	Tissues[b]	31 K[c]		H		W		E		Tot[d]	Chisq[e]	%[f]
			obs	exp	obs	exp	obs	exp	obs	exp			
Normal	175444	Placenta	2	31	12	33	**78**	21	1	8	93	196	41.1
	242043	Brain	**33**	16	13	17	2	11	1	4	49	28	5.9
	59127	Muscles	**12**	4	0	4	0	3	0	1	12	24	5.0
	64386	Breast	1	2	1	2	1	1	**3**	0	6	14	2.9
	13645	Kidney	2	1	0	1	0	1	**2**	0	4	11	2.3
	35521	Skin	**5**	2	0	2	0	2	2	1	7	11	2.2
	20803	Blood	0	1	0	1	1	0	1	0	2	6	1.3
	49887	Pancreas	**6**	3	2	3	0	2	1	1	9	6	1.2
	41979	Prostate	4	3	2	3	0	2	2	1	8	5	1.2
	48736	Lung	0	1	**3**	1	0	1	1	0	4	5	1.1
	24843	Colon	3	2	1	2	0	1	1	0	5	3	0.7
	7865	Ovary	0	0	0	0	1	0	0	0	1	3	0.7
	24669	Endocrine Gl.	1	1	1	1	0	1	1	0	3	3	0.6
	38831	Testis	2	1	1	1	0	1	1	0	4	3	0.6
	2911	Uterus	1	0	0	0	0	0	0	0	1	2	0.4
	13481	H&N	1	1	1	1	0	0	0	0	2	1	0.2
	162273	Others Norm.	9	10	9	11	8	7	5	3	31	3	0.6
Cancerous	43776	Placenta	1	2	0	2	**4**	1	0	0	5	10	2.0
	181866	Brain	**4**	2	1	2	0	1	0	0	5	5	1.1
	36615	Muscles	1	0	0	0	0	0	0	0	1	2	0.4
	114921	Breast	8	6	5	6	1	4	**3**	1	17	5	1.1
	33389	Kidney	1	1	0	1	1	0	0	0	2	2	0.4
	88766	Skin	**3**	1	1	1	0	1	0	0	4	3	0.7
	5906	Blood	1	0	0	0	0	0	0	0	1	2	0.4
	59189	Pancreas	2	1	1	1	0	1	0	0	3	2	0.4
	66032	Prostate	1	4	**10**	4	0	3	1	1	12	13	2.7
	115526	Lung	0	2	**6**	2	0	1	0	0	6	11	2.3
	147878	Colon	7	14	**31**	15	1	9	2	3	41	30	6.3
	66094	Ovary	4	3	3	3	0	2	1	1	8	3	0.6
	56081	Endocrine Gl.	0	2	3	2	0	1	**3**	0	6	16	3.5
	8471	Testis	9	7	11	7	0	5	1	2	21	7	1.6
	155206	Uterus	**5**	2	0	2	0	1	0	0	5	10	2.1
	95292	H&N	5	3	5	4	0	2	0	1	10	5	1.0
	198784	Others Canc.	8	13	**28**	13	0	9	2	3	38	27	5.6
		Total	142		151		98		35		426	477	100

B

	ESTs[a]	Tissues[b]	K108[c]		H		W		E		Tot	Chisq	%
			obs	exp	obs	exp	obs	exp	obs	exp			
Normal	175444	Placenta	2	19	12	40	**78**	26	1	9	93	148	35.0
	242043	Brain	**11**	5	13	11	2	7	1	3	27	11	2.6
	59127	Muscles	**9**	2	0	4	0	2	0	1	9	36	8.5
	64386	Breast	0	1	1	2	1	1	**3**	0	5	14	3.4
	13645	Kidney	1	1	0	1	0	1	**2**	0	3	12	2.9
	35521	Skin	**5**	1	0	3	0	2	2	1	7	17	3.9
	20803	Blood	0	0	0	1	1	1	1	0	2	5	1.2
	49887	Pancreas	0	1	2	1	0	1	1	0	3	4	0.8
	41979	Prostate	0	1	2	2	0	1	**2**	0	4	8	2.0
	48736	Lung	0	1	3	2	0	1	1	0	4	4	0.9
	24843	Colon	1	1	1	1	0	1	1	0	3	3	0.7
	7865	Ovary	0	0	0	0	1	0	0	0	1	3	0.6
	24669	Endocrine Gl.	0	0	1	1	0	1	1	0	2	4	1.0
	38831	Testis	2	1	1	2	0	1	1	0	4	4	1.0
	2911	Uterus	0	0	0	0	0	0	0	0			
	13481	H&N	1	0	1	1	0	1	0	0	2	2	0.4
	162273	Others Norm.	1	5	9	10	8	6	**5**	2	23	7	1.6
Cancerous	43776	Placenta	0	1	0	2	**4**	1	0	0	4	10	2.5
	181866	Brain	2	1	1	1	0	1	0	0	3	4	1.1
	36615	Muscles	1	0	0	0	0	0	0	0	1	4	0.9
	114921	Breast	**6**	3	5	6	1	4	**3**	1	15	7	1.7
	33389	Kidney	0	0	0	0	1	0	0	0	1	3	0.6
	88766	Skin	**3**	1	1	2	0	1	0	0	4	8	1.9
	5906	Blood	0	0	0	0	0	0	0	0			
	59189	Pancreas	2	1	1	1	0	1	0	0	3	4	1.1
	66032	Prostate	0	2	**10**	5	0	3	1	1	11	11	2.7
	115526	Lung	0	1	**6**	3	0	2	0	1	6	8	1.9
	147878	Colon	5	8	**31**	17	1	11	2	4	39	23	5.5
	66094	Ovary	**3**	1	3	3	0	2	1	1	7	4	0.9
	56081	Endocrine Gl.	0	1	3	3	0	2	**3**	1	6	13	3.0
	8471	Testis	4	3	11	7	0	4	1	2	16	7	1.8
	155206	Uterus	1	0	0	0	0	0	0	0	1	4	0.9
	95292	H&N	**5**	2	5	4	0	3	0	1	10	8	2.0
	198784	Others Canc.	6	7	**28**	15	0	10	2	4	36	21	5.0
		Total	71		151		98		35		355	423	100

[a] Number of ESTs from non-normalized, non-subtracted cDNA libraries sequenced for the tissue.

[b] Tissues in which HERV families are expressed. Others Norm. (adipose, nerve, fetal liver-spleen, eye, bone) and Others Canc. (stomach, lymph node, small intestine, bone marrow, bladder, cervix), sets of tissues in which no HERV expression was observed in the corresponding cancerous or normal tissues, respectively.

[c] Observed and expected frequencies (red) in the tissue for each HERV family, based on query sets composed of either 31 members of the HERV-K family (A) or the HERV-K108 provirus alone (B) and one member each of the HERV-H, -W and -E families. Expected frequency = (sum obs ESTs for HERV family x sum obs ESTs in tissue) / total ESTs. Observed frequencies that are at least two-fold higher than the expected frequencies are shown in bold. The yellow background highlights significant differential expression of a proviral HERV family compared to the other HERV families.

[d] Sum of the number of HERV ESTs observed in the tissue.

[e] Chi square value for the tissue. Chisq = $\Sigma((obs-exp)^2/exp)$.

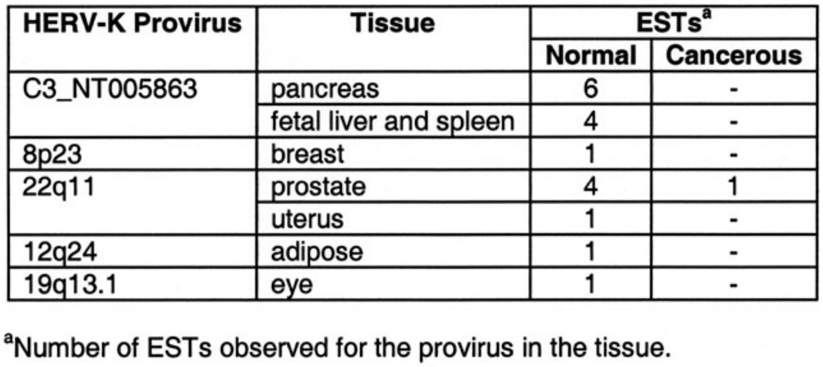
[f] Contribution (%) of the chi square value for the tissue to the total chi square value.

As shown in Table 3A, the HERV-K family is the only HERV family expressed in normal muscle and is more expressed in normal and cancerous brain and skin tissues than the other HERV families. Expression of the HERV-K family in these tissues is confirmed by the results obtained with the HERV-K108 provirus as the only representative of the HERV-K family (Table 3B). The HERV-K family is also expressed in other cancerous tissues, such as testis, uterus and head and neck, as well as in normal pancreas. However, the observed frequency in testicular cancers is just above the expected frequency (9/7) and therefore does not meet the two-fold criteria set above. It is interesting to note that most of the HERV-K ESTs identified in the "Others Canc." category (7/8) were isolated from stomach cancer cDNA libraries. Half of these ESTs stem from either the HERV-K101 (3 ESTs) or the K108 (1 EST) provirus. Taking the size of the cDNA libraries into account and comparing the frequencies observed in normal and cancerous tissues for the HERV-K family, we observed higher HERV-K expression levels in the following normal tissues: brain, muscle, skin and pancreas. On the other hand, the family is more expressed in testis, breast, and head and neck cancerous tissues than in the respective normal tissues. Therefore, the HERV-K family is expressed in both normal and cancerous tissues. In contrast, the HERV-H family is preferentially expressed in cancerous tissues such as prostate and colon. It is also over-represented in testicular cancers (11/7), though not reaching the two-fold threshold between observed and expected frequencies. On the other hand, the over-representation of observed HERV-H ESTs in the "Others Canc." category (28/13) indicates that the HERV-H family is specifically expressed in a set of cancerous tissues comprising bone marrow (5 ESTs), small intestine (4), bladder (3) and cervix (3) and is more highly expressed in stomach cancers (11) than the HERV-K family (7). The HERV-W family is predominantly expressed in normal placenta, but is also expressed in cancerous placenta. Finally, the HERV-E family is the least expressed of the HERV families analyzed. This family seems to be expressed in normal and cancerous breast tissues, normal kidney and cancerous endocrine glands but the number of ESTs identified is relatively low (<5). However, the 5 ESTs observed in the "Others Norm." category were isolated from normal eye tissue.

HERV-K proviruses with atypical tissue expression

Approximately two-fold more ESTs were matched when thirty-one members of the HERV-K family were used instead of HERV-K108 alone in the query set (Table 2), indicating that some HERV-K proviruses matched additional ESTs. Consequently, when all HERV-K proviruses were present in the query set, additional normal and cancerous tissues showing HERV-K expression were identified. Table 4 shows HERV-K proviruses that are expressed in tissues where other members of the family are not usually expressed. For instance, C3_NT005863 is the only HERV-K provirus expressed in normal pancreas and is the most expressed provirus in fetal liver and spleen. The HERV-K 22q11 provirus accounts for all matched ESTs originating from normal uterus and normal or malignant prostate.

Table 4. HERV-K proviruses expressed in tissues in which other members of the family are not usually expressed.



HERV-K Provirus	Tissue	ESTs[a]	
		Normal	Cancerous
C3_NT005863	pancreas	6	-
	fetal liver and spleen	4	-
8p23	breast	1	-
22q11	prostate	4	1
	uterus	1	-
12q24	adipose	1	-
19q13.1	eye	1	-

[a] Number of ESTs observed for the provirus in the tissue.

HERV expression analysis by MPSS

Massively parallel signature sequencing is a technology that allows the generation of millions of signature tags proximal to the 3' end of transcripts (21). Unlike other methods of sequence-based transcriptome analysis, such as ESTs and serial analysis of gene expression (SAGE), the MPSS technology can obtain, in a single experiment, up to a 10-fold clone coverage of the transcripts present in a human cell. MPSS data have been generated for 31 normal tissues and 3 cancer cell lines. Therefore, we were interested in comparing the HERV expression data based on EST counts with that based on MPSS data. We first predicted potential MPSS tags in HERV proviral sequences by extracting 13 nt sequences adjacent to the *Dpn*II site proximal to the polyadenylation site present in the 3' LTR of HERV proviruses. After removing tags that also matched cellular genes, we checked if those sequences were present in the MPSS data and counted the number of times each tag was identified in each tissue. Three MPSS tags meeting these criteria were identified. The first one is shared by 17 HERV-K proviruses including provirus C3_NT005863, the second is specific for HERV-K 22q11, while the third is specific for the HERV-W provirus. The number of instances each MPSS tag was detected in the different tissues, reflecting the expression levels of the corresponding transcripts, is shown in Table 5. Because a modified protocol has been used to generate the MPSS data for the three cancer cell lines (Table 5B), the normal breast cell line and a normal placenta tissue sample, the number of MPSS tags in these tissues cannot be compared with those present in the normal tissues (Table 5A).

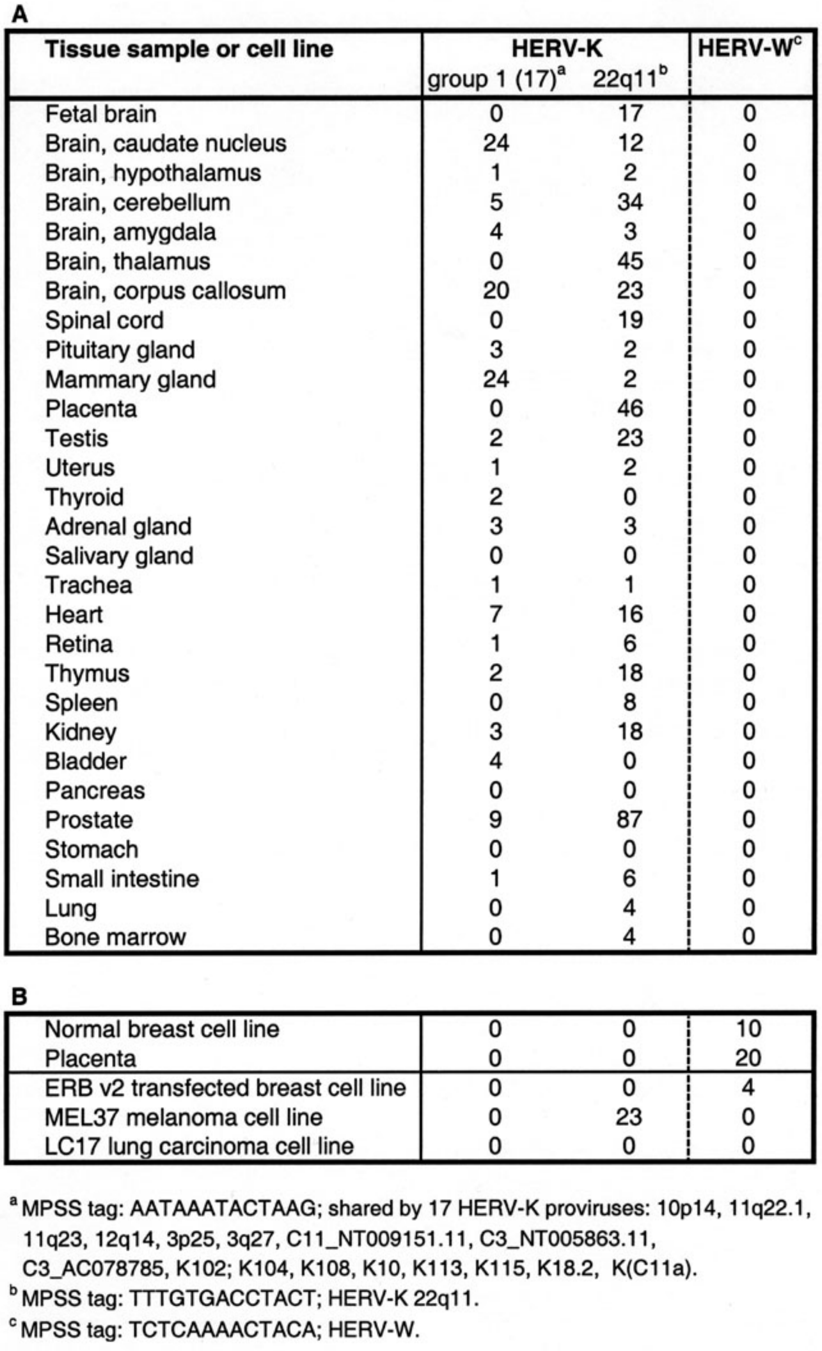
Expression of the HERV-K 22q11 provirus in normal brain and prostate tissues (see Supplementary Table 2) is confirmed by the MPSS data. Moreover, the MPSS analysis reveals expression of this provirus in normal tissues (Table 5A) that were not identified by the EST expression analysis, such as placenta, testis, kidney and other tissues, as well as in a melanoma cell line (Table 5B). In contrast, the expression pattern of the MPSS tag shared by 17 HERV-K proviruses is similar but not identical to the ones obtained based on the EST analysis. For instance, there was a correlation between MPSS and EST data for expression of these HERV-K proviruses in normal brain and breast tissues, but no expression could be detected in normal pancreas using MPSS while expression of the C3_NT005863 provirus in this tissue had been observed in the ESTs analysis (Supplementary Table 2). HERV-W expression in normal placenta and breast tissues is, however, confirmed by the MPSS analysis (Table 5B).

Analysis of ESTs matching retroviral open reading frames

Because of the accumulation of mutations in HERV sequences, most retroviral ORFs are fragmented or shortened compared to their original ancestor. In order to determine whether ESTs preferentially matched conserved retroviral ORFs, we analyzed the proportion of ESTs matching open reading frame sequences longer than 450 nucleotides for each HERV family studied. We found that a majority of ESTs from non-normalized cDNA libraries matched such HERV ORFs. The proportion of ESTs matching ORFs was 66% (63/95) for HERV-K, 67% (66/98) for HERV-W, 75% (27/36) for HERV-E and 41% (74/157) for HERV-H (Supplementary Table 3). Therefore, with the exception of HERV-H, a majority of the ESTs matched ORFs potentially encoding at least 150 aa.

Among the ESTs matching HERV ORFs, all HERV-W specific ESTs (66/66) and a majority of HERV-H ESTs (58%, 43/74) matched a full-length envelope ORF encoded by the respective provirus. Moreover, 50% (7/14) of ESTs matching HERV-K C3_NT005863 and 37.5% (6/16) matching the 22q11 proviral ORFs matched regions encoding potentially functional full-length Gag proteins. The remaining HERV-K C3_NT005863 and 22q11 ESTs matched partial *pol* ORFs. For the other HERV-K proviruses in the query set, most ESTs matched either *pol* or *env* partial ORFs.

Table 5. HERV expression analysis by massively parallel signature sequencing (MPSS).



A

Tissue sample or cell line	HERV-K		HERV-W[c]
	group 1 (17)[a]	22q11[b]	
Fetal brain	0	17	0
Brain, caudate nucleus	24	12	0
Brain, hypothalamus	1	2	0
Brain, cerebellum	5	34	0
Brain, amygdala	4	3	0
Brain, thalamus	0	45	0
Brain, corpus callosum	20	23	0
Spinal cord	0	19	0
Pituitary gland	3	2	0
Mammary gland	24	2	0
Placenta	0	46	0
Testis	2	23	0
Uterus	1	2	0
Thyroid	2	0	0
Adrenal gland	3	3	0
Salivary gland	0	0	0
Trachea	1	1	0
Heart	7	16	0
Retina	1	6	0
Thymus	2	18	0
Spleen	0	8	0
Kidney	3	18	0
Bladder	4	0	0
Pancreas	0	0	0
Prostate	9	87	0
Stomach	0	0	0
Small intestine	1	6	0
Lung	0	4	0
Bone marrow	0	4	0

B

Normal breast cell line	0	0	10
Placenta	0	0	20
ERB v2 transfected breast cell line	0	0	4
MEL37 melanoma cell line	0	23	0
LC17 lung carcinoma cell line	0	0	0

[a] MPSS tag: AATAAATACTAAG; shared by 17 HERV-K proviruses: 10p14, 11q22.1, 11q23, 12q14, 3p25, 3q27, C11_NT009151.11, C3_NT005863.11, C3_AC078785, K102; K104, K108, K10, K113, K115, K18.2, K(C11a).

[b] MPSS tag: TTTGTGACCTACT; HERV-K 22q11.

[c] MPSS tag: TCTCAAAACTACA; HERV-W.

Discussion

The fact that HERV expression has been reported in multiple cancer tissues and that HERV-K endogenous retroviral proteins have been identified using the SEREX methodology suggests that these proteins could be useful antigens for diagnostic purposes or cancer immunotherapy. We therefore studied digital expression patterns of the four least degenerate HERV families using members of each family to search human ESTs. As expected from the sequence differences between the HERV families at the DNA level, we observed that ESTs derived from different HERV families could easily be distinguished: ESTs matching members of the HERV-K family did not match proviral sequences from the HERV-W, H or E families. Moreover, although individual HERV-K members often did not match all family-specific ESTs, their expression patterns were similar and clustered together (Figure 1). Since our query set contained most of the proviral members of the HERV-K family, we also attempted to match each EST to its respective provirus. In order to be able to compare expression levels between or within HERV families, we compared EST counts from non-normalized, non-subtracted cDNA libraries.

Total EST counts for the HERV families analyzed in this study were relatively low, suggesting that HERVs are expressed at low levels. This is consistent with the fact that transcription of most HERV proviruses is usually shut down, possibly by methylation (10, 34). However, this observation might also be due to limitations in the depth of cDNA sequencing. Firstly, the number of cDNA libraries and EST sequences generated vary greatly from tissue to tissue. For example, about 80 times more EST sequences have been generated from normal brain than from normal uterus. Secondly, poorly expressed genes may be under-represented in relatively small cDNA libraries. The digital expression patterns reported here are consistent with the patterns already reported for the respective HERV families. For example, the HERV-K and -H families are known to be expressed in testicular and germ cell tumors. However, using digital expression profiling, expression in additional tissues was found. The HERV-K family is more highly expressed in normal tissues than in tumors. It is the only HERV family expressed in normal muscle and is overexpressed in normal brain, skin and pancreas, as well as in cancers of the brain, head and neck, and uterus, compared to the other HERV families. Moreover, HERV-K101 and K108 are expressed in stomach cancers but not in normal stomach tissues. Our analysis suggests that the HERV-K C3_NT005863 (HERV-KII) is expressed in normal pancreas and confirms the expression of the HERV-K 22q11 provirus in normal and malignant prostate tissues, contrasting with the rest of the HERV-K family. The level of expression of HERV-K 22q11 is however higher in normal prostate tissues than in prostate cancers. On the other hand, the HERV-H family is preferentially expressed in cancer tissues (105 ESTs from cancerous tissues versus 46 ESTs from normal tissues) and has a higher level of expression in stomach, colon, prostate and testis tumors than other families. The observed versus expected frequency for HERV-H in testicular cancers, however, was less than two-fold (11/7). In addition, this family seems to be specifically expressed in a set of cancer tissues such as small intestine, bone marrow, bladder and cervix. The HERV-W family is mostly expressed in placenta (8) where its envelope protein (syncytin) has been reported to play a role in human placenta morphogenesis by mediating cytotrophoblast fusion (33). Finally, the HERV-E family had a relatively low digital expression.

Tissue expression patterns of HERV-K 22q11 (expressed in brain and prostate tissues) and HERV-W (predominantly expressed in placenta) were confirmed by the MPSS data (22) in a set of 31 different normal tissue samples and 3 cancer cell lines. The higher sensitivity of the MPSS analysis relative to other digital expression methods allowed us to identify additional tissues in which the HERV-K 22q11 provirus is expressed, such as normal kidney, heart and thymus. On the other hand, the MPSS analysis was limited by the fact that MPSS data were available to us for only a few tissue samples, and thus expression detected by ESTs could not always be confirmed by MPSS.

We observed that a majority of ESTs matched relatively long and sometimes full-length ORFs of the HERV-K, -W and -E families. The active transcription of HERV ORFs and the detection of antibody responses against the HERV-K Gag or Env proteins in about 60% of germ cell tumor patients (18, 35) suggest that these proteins could be potential targets in cancer immunotherapy. However, since peptides encoded by these long HERV ORFs might not be produced *in vivo*, it is important to assess their tissue expression levels experimentally.

Because only one provirus each was used to represent the HERV-W, -H or -E families, we cannot be sure whether ESTs matching the corresponding family were transcribed from the provirus used in the query set or another member of the family. However, since the HERV-H family was found to be expressed in some cancerous tissues (small intestine, bone marrow, bladder and cervix) but not in the corresponding normal tissues, members of this family should be further investigated for their expression patterns and immunological properties.

For the HERV-K family, we were able to match a majority of ESTs to their respective proviruses. Our digital expression analysis showed that C3_NT005863 and 22q11 are the most highly expressed HERV-K members. The fact that numerous ESTs matched full-length *gag* ORFs encoded by these two proviruses and that an antibody response against the HERV-K 22q11 Gag protein has been observed in prostate cancer patients (36), suggests that these proteins are good antigen candidates. Moreover, since HERV-K 22q11 is a member of the HERV-K(Old) subfamily (25), it still has a 96 bp insertion in the *gag* ORF, deleted in more recent HERV-K proviruses. It would be interesting to study the immunogenicity of the peptide encoded by this insertion since this subfamily is present at a lower copy number in the human genome than the rest of the HERV-K family. An antibody response against such an antigen might be more specific for the tissues in which HERV-K 22q11 is expressed, prostate in particular. However, the fact that the HERV-K 22q11 provirus is more highly expressed in normal prostate than in tumors and is also expressed in normal brain might raise some safety issues for an immunotherapy approach. Additionally, we identified two HERV-K stomach cancer target candidates as a few ESTs from this tissue matched *pol* or full-length *env* ORFs encoded by the K101 (3 ESTs) or K108 (1 EST) proviruses, respectively. Therefore, these proteins and HERV-K proviruses should be further investigated *in vivo* to determine their level of expression in cancer and normal tissues as well as their immunological properties.

In summary, EST and MPSS data indicate that HERV-derived RNAs are more widely expressed than originally thought. The HERV-K101 and K108, expressed in stomach cancers, as well as the HERV-H family that seems to be specifically expressed in some cancers might be good candidates for cancer immunotherapy or diagnosis. Their precise tissue distribution will have to be verified using more sensitive and quantitative methods, and the presence of HERV-specific proteins in these tissues investigated histologically. It is clear that the HERV families are expressed differentially across tissues. Whether this reflects different biological roles, if any, remains an open question.

Abbreviations

BLAT, BLAST-like alignment tool; HERVs, human endogenous retroviruses; LTR, long terminal repeat; MPSS, massively parallel signature sequencing; ORF, open reading frame

Acknowledgements

The authors would like to acknowledge the support provided to this project by Dr. Lloyd Old. We thank Prof. Ernest Feytmans and Dr. Pierre Farmer for advice on statistical analysis and Dr. Andrew Simpson for critical reading of the manuscript. This work was supported by the European Cancer Immunome Program (QLGT-CT1999-01211), the Cancer Research Institute (New York) and the Ludwig Institute for Cancer Research.

References

1. Lower R, Lower J, Kurth R. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proc Natl Acad Sci USA* 1996; **93**: 5177-84. (PMID: 8643549)

2. Smit AF. The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev* 1996; **6**: 743-8. (PMID: 8994846)

3. Tristem M. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *J Virol* 2000; **74**: 3715-30. (PMID: 10729147)

4. Griffiths DJ. Endogenous retroviruses in the human genome sequence. *Genome Biol* 2001; **2**: REVIEWS1017. (PMID: 11423012)

5. Coffin JM. Retroviridae: the viruses and their replication. In: Howley PM, editor. Fundamental Virology. 3rd ed. Philadelphia (PA): Lippincott-Raven; 1996. p. 798-806.

6. Ono M. Molecular cloning and long terminal repeat sequences of human endogenous retrovirus genes related to types A and B retrovirus genes. *J Virol* 1986; **58**: 937-44. (PMID: 3009897)

7. Tonjes RR, Lower R, Boller K, Denner J, Hasenmaier B, Kirsch H, Konig H, Korbmacher C, Limbach C, Lugert R, Phelps RC, Scherer J, Thelen K, Lower J, Kurth R. HERV-K: the biologically most active human endogenous retrovirus family. *J Acquir Immune Defic Syndr Hum Retrovirol* 1996; **13 Suppl 1**: S261-7. (PMID: 8797733)

8. Blond JL, Beseme F, Duret L, Bouton O, Bedin F, Perron H, Mandrand B, Mallet F. Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family. *J Virol* 1999; **73**: 1175-85. (PMID: 9882319)

9. Lindeskog M, Mager DL, Blomberg J. Isolation of a human endogenous retroviral HERV-H element with an open env reading frame. *Virology* 1999; **258**: 441-50. (PMID: 10366582)

10. Florl AR, Lower R, Schmitz-Drager BJ, Schulz WA. DNA methylation and expression of LINE-1 and HERV-K provirus sequences in urothelial and renal cell carcinomas. *Br J Cancer* 1999; **80**: 1312-21. (PMID: 10424731)

11. Armbruester V, Sauter M, Krautkraemer E, Meese E, Kleiman A, Best B, Roemer K, Mueller-Lantzsch N. A novel gene from the human endogenous retrovirus K expressed in transformed cells. *Clin Cancer Res* 2002; **8**: 1800-7. (PMID: 12060620)

12. Johnston JB, Silva C, Holden J, Warren KG, Clark AW, Power C. Monocyte activation and differentiation augment human endogenous retrovirus expression: implications for inflammatory brain diseases. *Ann Neurol* 2001; **50**: 434-42. (PMID: 11601494)

13. Boller K, Konig H, Sauter M, Mueller-Lantzsch N, Lower R, Lower J, Kurth R. Evidence that HERV-K is the endogenous retrovirus sequence that codes for the human teratocarcinoma-derived retrovirus HTDV. *Virology* 1993; **196**: 349-53. (PMID: 8356806)

14. Wang-Johanning F, Frost AR, Johanning GL, Khazaeli MB, LoBuglio AF, Shaw DR, Strong TV. Expression of human endogenous retrovirus k envelope transcripts in human breast cancer. *Clin Cancer Res* 2001; **7**: 1553-60. (PMID: 11410490)

15. Wang-Johanning F, Frost AR, Jian B, Azerou R, Lu DW, Chen DT, Johanning GL. Detecting the expression of human endogenous retrovirus E envelope transcripts in human prostate adenocarcinoma. *Cancer* 2003; **98**: 187-97. (PMID: 12833471)

16. Lindeskog M, Blomberg J. Spliced human endogenous retroviral HERV-H env transcripts in T-cell leukaemia cell lines and normal leukocytes: alternative splicing pattern of HERV-H transcripts. *J Gen Virol* 1997; **78 (Pt 10)**: 2575-85. (PMID: 9349478)

17. Karlsson H, Bachmann S, Schroder J, McArthur J, Torrey EF, Yolken RH. Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. *Proc Natl Acad Sci USA* 2001; **98**: 4634-9. (PMID: 11296294)

18. Goedert JJ, Sauter ME, Jacobson LP, Vessella RL, Hilgartner MW, Leitman SF, Fraser MC, Mueller-Lantzsch NG. High prevalence of antibodies against HERV-K10 in patients with testicular cancer but not with AIDS. *Cancer Epidemiol Biomarkers Prev* 1999; **8**: 293-6. (PMID: 10207631)

19. Sauter M, Schommer S, Kremmer E, Remberger K, Dolken G, Lemm I, Buck M, Best B, Neumann-Haefelin D, Mueller-Lantzsch N. Human endogenous retrovirus K10: expression of Gag protein and detection of antibodies in patients with seminomas. *J Virol* 1995; **69**: 414-21. (PMID: 7983737)

20. Schiavetti F, Thonnard J, Colau D, Boon T, Coulie PG. A human endogenous retroviral sequence encoding an antigen recognized on melanoma by cytolytic T lymphocytes. *Cancer Res* 2002; **62**: 5510-6. (PMID: 12359761)

21. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K. Gene

expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 2000; **18**: 630-4. (PMID: 10835600)

22. Jongeneel CV, Iseli C, Stevenson BJ, Riggins GJ, Lal A, Mackay A, Harris RA, O'Hare MJ, Neville AM, Simpson AJ, Strausberg RL. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc Natl Acad Sci USA* 2003; **100**: 4702-5. (PMID: 12671075)

23. Barbulescu M, Turner G, Seaman MI, Deinard AS, Kidd KK, Lenz J. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Curr Biol* 1999; **9**: 861-8. (PMID: 10469592)

24. Hughes JF, Coffin JM. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. *Nat Genet* 2001; **29**: 487-9. (PMID: 11704760)

25. Reus K, Mayer J, Sauter M, Zischler H, Muller-Lantzsch N, Meese E. HERV-K(OLD): ancestor sequences of the human endogenous retrovirus family HERV-K(HML-2). *J Virol* 2001; **75**: 8917-26. (PMID: 11533155)

26. de Parseval N, Casella J, Gressin L, Heidmann T. Characterization of the three HERV-H proviruses with an open envelope reading frame encompassing the immunosuppressive domain and evolutionary history in primates. *Virology* 2001; **279**: 558-69. (PMID: 11162811)

27. Huang X, Miller W. A time-efficient, linear-space local similarity algorithm. *Adv Appl Math* 1991; **12**: 373-81.

28. Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* 1999; **174**: 247-50. (PMID: 10339815)

29. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* 2000; **7**: 203-14. (PMID: 10890397)

30. Sugimoto J, Matsuura N, Kinjo Y, Takasu N, Oda T, Jinno Y. Transcriptionally active HERV-K genes: identification, isolation, and chromosomal mapping. *Genomics* 2001; **72**: 137-44. (PMID: 11401426)

31. Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel CV, McCarthy MI, Hide T, Hide W. eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res* 2003; **13**: 1222-30. (PMID: 12799354)

32. Blond JL, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes S, Mandrand B, Mallet F, Cosset FL. An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. *J Virol* 2000; **74**: 3321-9. (PMID: 10708449)

33. Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, Keith JC Jr, McCoy JM. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* 2000; **403**: 785-9. (PMID: 10693809)

34. Gotzinger N, Sauter M, Roemer K, Mueller-Lantzsch N. Regulation of human endogenous retrovirus-K Gag expression in teratocarcinoma cell lines and human tumours. *J Gen Virol* 1996; **77 (Pt 12)**: 2983-90. (PMID: 9000088)

35. Boller K, Janssen O, Schuldes H, Tonjes RR, Kurth R. Characterization of the antibody response specific for the human endogenous retrovirus HTDV/HERV-K. *J Virol* 1997; **71**: 4581-8. (PMID: 9151852)

36. Obata Y. Personal communication.

37. Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J. Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr Biol* 2001; **11**: 1531-5. (PMID: 11591322)

38. Costas J. Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length proviral genomes. *J Mol Evol* 2001; **53**: 237-43. (PMID: 11523010)

39. Tonjes RR, Czauderna F, Kurth R. Genome-wide screening, cloning, chromosomal assignment, and expression of full-length human endogenous retrovirus type K. *J Virol* 1999; **73**: 9187-95. (PMID: 10516026)

40. Repaske R, Steele PE, O'Neill RR, Rabson AB, Martin MA. Nucleotide sequence of a full-length human endogenous retroviral segment. *J Virol* 1985; **54**: 764-72. (PMID: 3999194)

41. Ono M, Yasunaga T, Miyata T, Ushikubo H. Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. *J Virol* 1986; **60**: 589-98. (PMID: 3021993)

42. Stauffer Y, Marguerat S, Meylan F, Ucla C, Sutkowski N, Huber B, Pelet T, Conrad B. Interferon-alpha-induced endogenous superantigen. a model linking environment and autoimmunity. *Immunity* 2001; **15**: 591-601. (PMID: 11672541)

43. Human BLAT Search. URL: http://zeon.well.ox.ac.uk/cgi-bin/hgBlat?command=start&org=human

44. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981; **147**: 195-7. (PMID: 7265238)

45. Altschul SF. The Statistics of Sequence Similarity Scores. URL: http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

46. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA* 1990; **87**: 2264-8. (PMID: 2315319)

47. NCBI ORF Finder. URL: http://www.ncbi.nih.gov/gorf/gorf.html

Materials and methods

Mapping of HERV proviral sequences

Information about previously described HERV-K, -W, -H and -E proviruses was retrieved from the literature (8, 23, 24, 26, 37, 38, 39, 40) and the proviruses mapped to human genomic sequences and Ensembl chromosomes (NCBI 33 assembly). Some proviral sequences have been submitted individually to GenBank/EMBL, but in most cases only large bacterial artificial chromosome (BAC) or contig sequences containing the proviral HERV sequence were available. In order to determine the start and stop coordinates of the provirus on the genomic sequence and analyze the proviral structure, pairwise alignment tools such as lalign (27) or bl2seq (28) were used to align known LTRs, *gag*, *pol* or *env* sequences for a known family member. Coordinates, as well as complete proviral sequences, were then retrieved from the GenBank sequence repository or Ensembl web site. For the HERV-K family, the sequences of the K101 and K108 proviruses described by Barbulescu *et al*. (23) were used as probes in this analysis.

Alternatively, HERV-K proviral sequences were obtained using our own strategy to identify new HERV-K proviruses in the human genome. A query set (probes) composed of 6 HERV-K proviruses (K101, K102, K104, K108 and K113) described by Barbulescu *et al*. (23, 37), HERV-K10 (41) and two HERV-K18 alleles (42), was used to search the NCBI NT genomic contig database using Megablast (29). An alignment of at least 1000 base pairs was used as seed and the alignment was extended taking all other overlapping alignments into account. Information about the matched NT contig and the sequence was then extracted. This new strategy allowed us to also map slightly divergent HERV-K sequences. About 80 HERV-K insertion sites were found in the human genome using this approach, most of them containing only partial, truncated proviral sequences. Among the 22 predicted full-length HERV-K insertions, three potentially new proviral sequences, named after the chromosome and NCBI NT contig on which they were identified, were selected for further analysis (C3_NT005863.11, C3_NT022411.11 and C11_NT009151.11). The first two proviruses, C3_NT005863 and C3_NT022411, correspond to the HERV-K(II) and HERV-K(I) proviral sequences described by Sugimoto *et al*. (30). The third proviral sequence, C11_NT009151, is identical to the HERV-K provirus 11q23 identified by Hughes *et al*. (24).

Digital expression analyses

Query sets containing proviral sequences without LTRs (see supplemental data) as probes were constituted to search human ESTs using Megablast (29). Filters were inactivated (-FF) and mismatches penalized (-q -9). Only alignments with an E value lower than 10^{-40} were selected. The first query set was composed of 31 HERV-K proviruses (see Table 1) and one representative each of the other HERV-W (GenBank Accession No. AC000064 30950-37500), -H (GenBank Accession No. AJ289709) and -E (GenBank Accession No. M10976) families. For comparison purposes, the same analysis was carried out with the HERV-K108 provirus as a representative of the HERV-K family.

Hits identified by Megablast were first clustered by EST, sorted by alignment scores, and the library name and identifier were then retrieved for each EST. The controlled hierarchical vocabulary (eVOC) for development stages, anatomical sites and pathology types was used to classify cDNA libraries and describe the tissues from which they were generated (31). After selecting only the best matching provirus for each EST, tissue information was retrieved based on the library identifier. The first output of this analysis was a tab-delimited file containing a list of ESTs, best matching HERV proviruses, alignment information, cDNA library names and tissue descriptions. The data in the file was then rearranged by HERV families, sorted by HERV provirus and ESTs from standard cDNA libraries were separated from those derived from normalized or subtracted ones (see Supplementary Table 3). Only EST counts from non-normalized, non-subtracted cDNA libraries were used to compare the expression levels of the different HERV families in normal and cancer tissues.

The relative expression of one HERV family in a tissue is characterized by the ratio of the number of ESTs matching members of the HERV family to the total number of ESTs sequenced in the respective tissue. HERV expression patterns were obtained by sorting tissues according to the relative expression of the HERV family in each tissue. The total number of ESTs present in a tissue was compiled by counting all ESTs from non-normalized, non-subtracted cDNA libraries.

Validation of HERV-K provirus loci matched by ESTs

To validate that the EST stemmed from the matched HERV-K provirus in the query set, HERV-K ESTs were searched against the human genome (NCBI build 33) using BLAT (43) and loci coordinates compared to characterized HERV-K provirus loci present in the query set (see Supplementary Table 4). In addition, a tag was added in Supplementary Table 3 for each HERV-K EST: Y, validated EST mapping to the HERV-K provirus; N, the EST matches a different locus from an uncharacterized HERV-K provirus with a higher BLAT score.

Clustering of HERV proviral expression

A comparison of all EST sequences matched by the HERV-K family versus all proviral sequences present in the query set was performed using the Smith-Waterman (SW) algorithm (44) with a standard DNA similarity matrix and gap opening/extension penalties set to -16/-4. The SW scores S were normalized to obtain bit scores (45) using the formula S_{bit}=[lambda$S$-(ln K)]/(ln 2), where the parameter values for the scoring system used above are lambda=0.15812780 and K=0.05411805. These values were estimated through simulation. The binary distance between any pair of proviral sequences was calculated according to the formula d=1-[n_{11}/(n_{11}+n_{10}+n_{01})], where n_{11} is the number of EST sequences with a bit score equal to or greater than 36 - this threshold corresponds to a P-value of 0.01 (46) for both proviral sequences; n_{10} and n_{01} are the number of EST sequences with a bit score equal to or greater than 36 matching only one of the two proviral sequences. This distance measure ranges from 0 to 1. Eventually, proviral sequences were clustered using average-linkage agglomerative hierarchical clustering based on the binary distances.

Comparison of the expression levels in normal and cancerous tissues of the HERV families

The observed frequencies for each HERV family in normal and cancerous tissues were used to build contingency tables and expected cell frequencies, E, calculated using the formula E=(RTxCT)/N, where RT is the row total, CT the column total and N the grand total. Chi square values were then computed for each tissue using the formula chi^2=sum of [$(\text{obs-exp})^2$/exp], where obs is the observed frequency of the HERV family in the tissue and exp its expected frequency. A HERV family was considered to be differentially expressed in a tissue if its observed frequency was at least two-fold higher than its expected frequency in the tissue and the observed frequencies for the other HERV families were equal to or lower than their expected frequencies. The contribution of the tissue to the total chi square value was also taken into account.

The relative expression levels of a HERV family in a normal tissue versus the corresponding cancerous tissue were assessed by comparing the observed frequencies in each tissue and the total number of ESTs sequenced for the respective tissue.

Evaluation of the coding capacity

Proviral sequences without LTRs were analyzed for the presence of open reading frames using the NCBI ORF finder web site (47). The coordinates of ORFs longer than 450 nt were retrieved for each provirus. The number of ESTs matching the ORFs was then compiled by comparing these coordinates to the ones of the matching ESTs on the provirus (Supplementary Table 3). The presence of putative conserved domains, characteristic for Gag, Pol or Env proteins, was also evaluated using the NCBI Conserved Domain search feature available on the NCBI ORF finder web site.

Supplemental data

Table 1. Definition of tissue categories. 17 KB Excel file 040102_suppl_tab1.xls

Table 2. Summary of tissues and number of ESTs matched by HERV proviruses. 93 KB Excel file 040102_suppl_tab2.xls

Table 3. ESTs matched by HERV proviruses (without LTRs) with matching parameters and tissue information. 175 KB Excel file 040102_suppl_tab3.xls

Table 4. ESTs mapping to the human genome using BLAT (HERV-K). 73 KB Excel file 040102_suppl_tab4.xls

HERV proviral sequences without LTRs. 258 KB text file 040102_suppl_seq.txt (FASTA format)

Entire supplemental data set. 170 KB WinZip file 040102_suppl_data.zip

Contact

Address correspondence to:

Victor Jongeneel
Ludwig Institute for Cancer Research, Office of Information Technology
Ch. des Boveresses 155
1066 Epalinges
Switzerland
Fax: + 41 21 692 4065
E-mail: Victor.Jongeneel@licr.org