Publication VII

Final manuscript version of an article published as:
Merja Oja. *In silico* expression profiles of human endogenous retroviruses. In *Proceedings of the Workshop on Pattern Recognition in Bioinformatics (PRIB 2007)*, volume 4774 of *Lecture Notes in Bioinformatics*, pages 253–263, 2007.
[http://www.springerlink.com/content/v257330413700652/]

VII

In Silico Expression Profiles of Human Endogenous Retroviruses

Merja Oja

Helsinki Institute for Information Technology,
Helsinki University of Technology, P.O.Box 5400, 02015 TKK, Finland,
and Department of Computer Science, University of Helsinki
merja.oja@tkk.fi

Abstract. Human endogenous retroviruses (HERVs) are remnants of ancient retrovirus infections and now reside within the human DNA. Recently HERV expression has been detected in both normal and diseased tissues. However, the patterns of expression of individual HERV sequences are mostly unknown. In this work we use a generative mixture model, based on hidden Markov models, for estimating the activities of individual HERV sequences from databases of expressed sequences. We determine the relative activities of sixty HERVs from the HML2 group in five human tissues, i.e. we estimate the *expression profile* of each HERV. This allows us to gain insight into HERV function.

1 Introduction

Human endogenous retroviruses (HERVs) are remains of retrovirus infections that occured millions of years ago. They are viruslike DNA sequences that reside within the human genome. HERV sequences form 8% of the human genomic DNA [3, 4].

Retroviruses can move and copy their DNA to other locations in the genome. These copying events will eventually yield several mutated versions of the original virus. A group of such sequences is called a HERV group and it may contain hundreds of very similar sequences. Most of the HERV sequences are heavily mutated and/or broken due to genomic rearrangements and have partially lost the typical retroviral structure consisting of 4 genes (gag, pro, pol and env) and two long terminal repeat sequences (LTRs), one at each end of the retrovirus sequence.

In this paper we study the HML2 group because it is the youngest and as such has the largest proportion of full length HERVs and the smallest number of mutations. Thus, it has the most potential for containing active HERVs.

HERVs are interesting for two reasons: they can express viral genes in human tissues and their presence in the genome may affect the function of nearby human genes. Retroviral activity might cause disease; retroviral mRNAs have been detected in schizophrenia, autoimmune diseases and cancer [2, 10] although a causal role of HERVs in these conditions is highly uncertain. In addition, a few retroviral genes have adopted functions beneficial to the human host [8].

In this work we study activities of individual HERV sequences in various tissues, i.e. will estimate the *expression profile* of each HERV. The profile contains measurements from several tissues and thus enables us to study the differential expression patterns of individual HERVs. This leads to better understanding of the function of individual HERVs. For example, HERVs that are more/only active in the brain tissue may have functions related to neurodegenerative diseases or to normal brain functions. This profiling approach is widely used in the study of human gene function, see for example [17]. In contrast, the only work that we know of where individual HERVs have been studied in several tissues is [16], where a small set of full-length HERV-K elements (HML2 is a subgroup of HERV-K) were studied using a heuristic method.

We have earlier studied the overall expression of individual HERVs (one expression value for each HERV without distinguishing between different tissues and conditions). In this work we extend the approach to estimation of expression *profiles* over various tissues. Furthermore, we analyze the expression profiles of *individual* HERV sequences. In contrast, most previous studies of HERV expression report activities only for HERV groups (e.g. [13]); the only exceptions we know of are [6] where HERVs are searched from gene mRNAs but activities are not compared across HERVs and [16] mentioned above.

To find evidence of HERV expression, we use a large public database of expressed sequence tags (ESTs). ESTs are short and noisy samples from mRNA sequences. The amount of ESTs originating from a particular HERV is evidence of its activity. However, it is nearly impossible to match an EST sequence to only one HERV sequence: Each EST will match several HERVs very well due to the similarity of the HERV sequences within a HERV group and the noise (sequencing errors) in the ESTs. We have introduced earlier a probabilistic model [11] to handle the uncertainty in EST to HERV matching. In the methods section we describe how this model can be used for estimating HERV expression profiles. The expression profiles for the HERV sequences of the HML2 group are presented in the results section.

2 Methods

In [11] a generative mixture model, based on hidden Markov models, for estimating the activities of individual HERV sequences from ESTs was introduced. Below we briefly describe this model and then move on to describe how it can be used when the aim is to estimate expression profiles instead of overall expression values.

The hidden Markov mixture model is a generative model for the set of EST sequences. It is designed to mimic the actual EST generation from HERVs; each mixture component is a hidden Markov model (HMM) for ESTs from a particular HERV (See Fig. 1). The component HMM resembles the profile HMM [7], with the exception that it is possible to jump from the start state to any of the match states and from any match state either to the end or to a special EEMIT state that is used to emit the low quality end of an EST. The match states, one for each

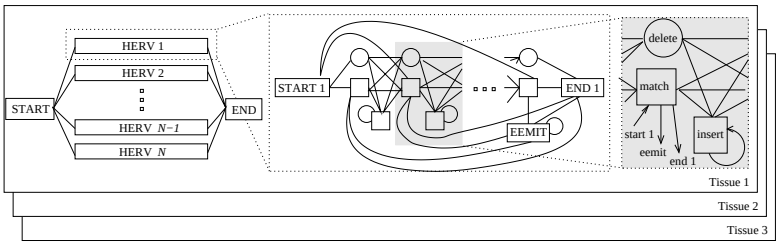

Fig. 1. The structure of the HMM mixture. The model is constrained by sharing parameters. The shaded box is the basic block of the sub-HMM and is repeated length-2 times. It is identical in all sub-HMMs; all other parameters are shared except the emission distribution of the match state which varies between blocks, according to the HERV sequence each sub-HMM corresponds to. EEMIT-state emits the low-quality end part. The plates illustrate that the same model is learned separately for each tissue.

position of the HERV sequence, can either emit the nucleotide in that position of the HERV sequence (with probability p_t) or one of the other nucleotides (with probabilities $(1 - p_t)/3$). The parameter p_t is shared between all match states in the mixture model. The EEMIT states and all the insert states share parameters: they emit nucleotides using the same distribution. The transition parameters are also shared throughout the mixture (see Fig. 1). In summary, the component HMM generates data that roughly matches a subsequence of the source HERV, but with mismatches, insertions, deletions, and a low-quality end part.

The mixture model corresponds to one large HMM where the first transition chooses one of the N HERV-specific sub-HMMs (see Fig. 1). The Baum-Welch algorithm is used to learn the whole mixture. The learned probabilities of the first transition (the mixture weights) are estimates of the HERV activities. We use heuristics to reduce HMM training time to reasonable limits [11].

The hidden Markov mixture model can be extended to estimation of expression profiles. We can simply learn a separate model for each tissue and then combine the results meaningfully. In practice, we need to collect several sets of EST sequences, one set for each tissue. Then we learn the model for each EST set. This results in the relative activity distributions of the HERVs for each tissue.

The relative activity distributions of HERVs from different tissues can be combined in two ways to form the HERV expression profiles. 1) The relative activities of a HERV in different tissues are used directly as the expression profile. In this setting it is assumed that each EST set, irrespective of its size, is a sample of all HERV derived mRNAs in the tissue. 2) The relative activities of a HERV in different tissues are first scaled according to the number of ESTs available from the tissues. This way the expression profile of a HERV is more clearly related to the number ESTs available from the HERV and the activity value of the HERV

can be seen as a *probabilistic EST count*. In this setting it is assumed that the size of the EST set is relevant. In this work we will use this second approach.

3 Data

We study the expression profiles of HERVs of the HML2 group. This HERV group is the youngest one and thus has the largest proportion of relatively intact elements. It contains sixty members, some of which are full-length, i.e. have retained the typical retrovirus structure LTR-gag-pro-pol-env-LTR. A few of these elements even have open reading frames for the env gene, i.e. they could produce retroviral env proteins.

The HML2 group is the most difficult one to study because the sequences within the young HML2 group are more similar to each other than sequences in other groups. It is impossible to match ESTs to individual HML2 HERVs unambiguously. Our statistical approach is able to alleviate this problem to some extent. But, even with our method, the activities of nearly identical HERVs will be correlated.

We studied the expression of HML2 HERVs in five tissue types: brain, lung, breast, placenta and male reproductive tissues (RT). This selection was mainly due to the availability of the ESTs, but some of these tissues are also interesting *per se*: HERV transcripts have been detected in brain related diseases, HERVs active in reproductive tissues could produce new HERV integrations and some HERVs are known to have beneficial functions in placenta. In addition, we know from earlier studies that HERV-K elements are active at least in testis and brain tissues [9].

The HERVs were automatically detected from the human genome by the program RetroTector[1]. Sixty of the HERVs were similar to HML2 reference sequence and were included into the HML2 HERV set.

ESTs matching the HML2 HERVs were searched from the dbEST database [20] with BLAST [1]. The ESTs were divided into tissue-specific sets using eVoc Ontologies [19]. We used a match threshold of E-value 10^{-40} in BLAST.

In addition to HML2 HERVs, some elements from other HERV groups were included in the HERV set. This was done to ensure reliable activity estimates for the HML2 HERVs: If the extra HERVs would not be included, then EST originating from them would be distributed over HML2 HERVs, falsely increasing their activity estimates. In other words, adding the extra HERVs reduces the error due to ESTs that match a non-HML2 HERV better than any of the HML2 HERVs.

The set of extra HERVs was selected based on a heuristic BLAST activity. The BLAST activity of a HERV is the number of EST matching that HERV better than any other HERV. ESTs that match several HERVs equally well are

[1] RetroTector is a program used for detecting retroviral sequences in genomes. It searches for conserved retroviral motifs and then combines the motifs into chains fulfilling distance constraints. It was developed by Jonas Blomberg and Göran Sperber at Uppsala University [15].

Table 1. Data set sizes for different tissues. "HERV-EST pairs" is the number of EST to HERV matches returned by BLAST.

Tissue	HERVs	ESTs	HERV-EST pairs
Brain	94	471	7076
Lung	86	279	4661
Placenta	85	219	2770
Breast	73	164	2987
Male reproductive tissue	89	249	4157

divided to all those HERVs. In our earlier work [11] BLAST activity was shown to correlate with activity estimates from the HMM model. We included all highly BLAST active HERVs (those with more than 2.5 ESTs) and then the most active HERV (still required to have at least one EST match) from each HERV group into the analysis. The set of extra HERVs was different for each tissue. Table 1 list the data sets sizes for all tissues.

4 Results

The method is able to estimate the relative activities of the HERVs. The activity profiles for HML2 HERVs are shown in Fig. 2A. Many of the HERVs exhibit tissue specific expression. There are also some HERVs that are active in all tissues as well as HERVs that are not active in any of them. The activities of most HML2 HERVs were previously unknown. A portion of the full-length HML2 HERVs have been studied before in [16] using a heuristic BLAST approach. Some individual HERVs are analyzed more closely in Section 4.1.

The results show that adding the extra HERVs was necessary to get reliable estimates for the HML2 HERVs. In each case the probability mass allotted to the HML2 HERVs was less than half of the total (ranging from 37% in the placenta to 48% in the lungs). If the extra HERVs would not have been included, then the probability mass now belonging to them would have been distributed over the HML2 HERVs, falsely increasing their activity estimates. Furthermore, some of the non-HML2 HERVs were very active in comparison to the mean activity level of the HML2 HERVs (see Fig. 2B). The high activity of the non-HML2 HERVs indicates that there is a lot of cross-talk between the HERV groups (the ESTs retrieved using the HML2 sequences as queries also match HERVs from the other groups). Some of the cross-talk might be due to portions of the HERVs that resemble other types of retrotransposons (see section 4.1).

We estimated the reliability of the results with a bootstrap-like method. The EST data was resampled with replacement 1000 times, and the activities were reoptimized for each replicate while other parameters were kept fixed (see [11] for more details). Fig. 3 shows the means and standard deviations of these replicates for the HERV activities in the lung tissue. The behavior in the other tissues is very similar. The standard deviations are small compared to the differences in HERV activities and the means are very close to the activities learned from all

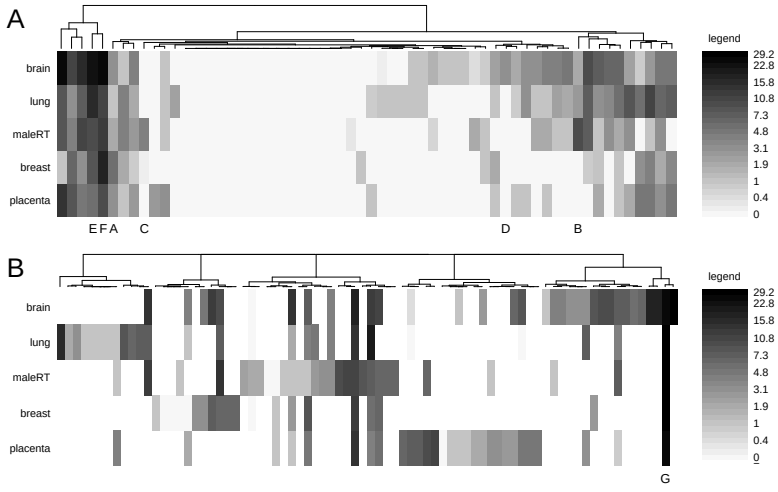

Fig. 2. The activities of the HML2 (panel **A**) and non-HML2 (panel **B**) HERVs. In both panels the rows depict the HERV activity distributions in different tissues and the columns the expression profiles of individual HERV sequences. Letters below the columns are labels for the HERVs analyzed in Section 4.1. The activity values are shown on a logarithmic scale, as can be seen from the legends on the right. The scale is the same in both panels. The numbers next to the legend are the probabilistic EST counts for each gray shade. The highest activity for a HML2 HERV is 24.8 (HERV F in the brain tissue). The columns have been ordered according to a hierarchical clustering based on the (unlogarithmic) Euclidean distances between the HERV expression profiles.

data. The standard deviations of the clearly active HERVs (probabilistic EST count above 5) and almost inactive HERVs (probabilistic EST count below 1) do not overlap. Thus we can trust the active-looking ones to be truly active.

4.1 Closer look on individual active HERVs

Here we take a closer look on some of the individual HERVs. These have been selected as examples of the typical expression patterns of the active HERVs observed in the data. The HERVs analyzed in this subsection are summarized in Table 2. The labels used to denote the HERVs are letters with no special meaning.

HERV A is full-length element with an open reading frame for the env gene. This HML2 HERV is known as HERV-K102. It is somewhat active in all tissues — its highest activity is observed in the breast tissue. The activity is due to ESTs that match the LTRs and the env gene area of the HERV. This HERV is a potential retrovirally active HERV that could produce env protein. HERV A

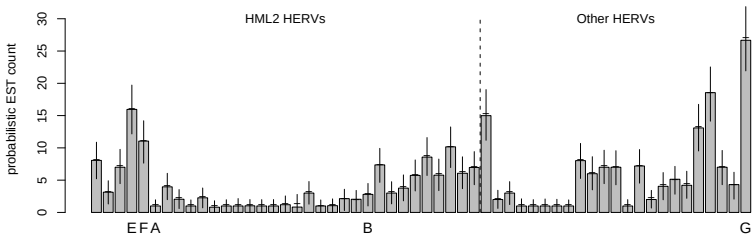

Fig. 3. The activities (probabilistic EST counts) of the HML2 and non-HML2 HERVs in the lung tissue. The crosses are the means and standard deviations from the bootstrap resamples (see text for details) and the bars the activities learned from complete data. The HERVs are in the same order as in Figs. 2A and 2B, but inactive HERVs (with probabilistic EST count below 10^{-7}) have been left out of the visualization to save space. The letters below the columns are the labels for the HERVs analyzed in Section 4.1.

is also mentioned in [16], but no exact details are given. UCSC Genome browser shows a new hypothetical human protein overlapping the LTRs of this HERV. This supports our finding that this HERV is retrovirally active.

HERV B is an almost full-length HERV with no open reading frames and a missing end-LTR. The HERV is active in the brain, lung and male reproductive tissues. Its activity is concentrated on gag and pol genes. This HERV has been studied earlier in [16], where it was found to be expressed in the brain, placenta, testis and prostate tissues. It had low activity in the lung and breast tissues. These results agree with our observations except for placenta and lung, for which our results are just the opposite.

HERV C is active only in the male reproductive tissues. The ESTs match this full-length HERV near the end of pol and at the end-LTR. The ESTs might be coming from the end of a pol gene transcript, however, ESTs from the beginning of the transcript are not observed. UCSC Genome browser shows a short gene sequence, annotated as a retroviral rec gene, between and partly overlapping the sequence segments detected as active by our method. This further supports the observation that this relatively intact HERV locus is active.

HERV D exhibits a clear tissue-specific expression: it is active only in the brain tissues. This non-full-length HERV is active in the gag gene area. However, there is no open reading frame for a gag protein. The observed expression does not resemble that of a retrovirally active HERV [4]. Hence, it seems that this HERV might have been used as a building block for something else than retroviral proteins.

The data set contains some HERVs that are very active in all studied tissues; for example, the HERV sequence E. ESTs match this HERV in the end of the pol gene and parts of env. However, when we look at this genome area at the UCSC Genome Browser, the pol gene area is annotated there as an L1 repeat. Thus, it may be that the (probabilistic) EST count of this HERV is actually measuring

L1 derived ESTs. Similar situation applies to the highly active HERV F, where the expression also seems to be L1 derived. These HERVs are examples of broken down sequences that are harder to detect automatically. For these HERVs the RetroTector program may have misinterpreted some portion of the L1 structure, which as a retrotransposon is similar to that of a retrovirus, as retrovirus-derived DNA.

HERV G measures expression of SVA elements that are composite retrotransposons consisting of an Alu like portion, a tandem repeat portion and a portion originating from the HML2 LTR sequence [12]. The end portion of the SVA repeat is about 95% similar to the HML2 LTR. For this reason, some of the ESTs retrieved using BLAST may actually come from a SVA element. As a consequence, it is necessary to include a SVA like sequence into the HERV set so that possible SVA derived ESTs will not confuse the activity estimates of LTR-containing HML2 HERVs. The SVA-ESTs will match the SVA like "HERV" better and thus have low probability on matches to HERVs. It turns out that one sequence in the HERV collection (marked with G in the figures) obtained by RetroTector is very similar to a SVA element and actually portions of this sequence are annotated as SVA in the UCSC Genome Browser. It was included into the HERV set to serve as the SVA like element. The results show that this "HERV" is very active in all tissues and the activity is in the SVA repeat areas. This indicates SVA activity in all the analyzed tissues.

5 Discussion and Conclusions

We have used a generative model-based method to estimate the expression profiles of individual HERVs rather than those of HERV groups. Such detailed analysis is vital for understanding the functions and control mechanisms of HERVs. Our method allows the exploration of expression patterns within a HERV group and will reveal interesting potentially active HERVs. These can then be studied further and their activity levels in different tissues can be verified with laboratory methods. By contrast, exhaustive search of active HERVs in the laboratory would be too expensive and/or difficult.

The advantage of our method over a simple "find the best matching HERV for each EST" approach (such as the BLAST activity method described in section 3) is the ability to take uncertainties into account. Our model is able to learn the underlying activities from data where the error rate (noise) in the ESTs is larger than differences between two HML2 HERV sequences. In our earlier work [11] we showed with experiments on simulated data that the HMM model outperforms the simple BLAST activity estimation method. The difference was most notable in the case of HML2 HERVs.

The number of ESTs available from each tissue was not as high as we would have hoped: The EST sets were small with only about three ESTs per HERV. As a result, the activity estimates are not as accurate as they would have been with a larger data set. Still, our results were reliable according to bootstrap

Table 2. Details about the HERVs analyzed in Section 4.1. "Label" is the label of the HERV used in the text and figures. "Chr", "strand", "start" and "end" tell the chromosome, the strand, and the sequence start and end positions for the HERV, respectively (in the July 2003 version (hg16) of the human genome). "Subgenes" describes the structure and "group" the group of the HERV. The last column in the upper part of the table gives the name used for the HERV in [16]. The "orf" columns describe how intact the retrovirus protein reading frame is: 0 is intact, i.e. the HERV has a open reading frame for the protein. "Age" is the estimated age of the element measured in percentage of LTR unsimilarity. The two LTRs of a retrovirus are identical on integration and mutate afterwards. The "gene context" column gives the gene nearest to or overlapping with the HERV locus.

label	chr	start	end	strand	subgenes	group	name in [16]
A	1	152822428	152813249	-	LTRgagpropolenvLTR	HML2	K102
B	22	22203232	22213324	+	LTRgagpropolenv	HML2	22q11
C	11	101103511	101112976	+	LTRgagpropolenvLTR	HML2	11q22.1
D	7	140863179	140859365	-	gagpropol	HML2	
E	16	35307416	35314276	+	polenv	HML2	
F	1	75265364	75273509	+	LTRgagpro	HML2	
G	19	21682582	21697392	+	LTRLTR	unknown	

label	gagorf	proorf	polorf	envorf	age	gene context
A	3	0	1	0	0.21	3' LTR the last exon of a hypothetical gene
B	1	5	12	9	-	gene IGLL1 2.5 Kb away downs. (antisense)
C	3	0	1	1	0.41	part annotated as retroviral rec gene
D	8	1	16	-	-	gene SSBP1 1 Kb away downs. (antisense)
E	-	-	15	2	-	nearest gene 20 Kb downstream (sense)
F	2	0	-	-	-	HERV in a long intron of an antisense gene
G	-	-	-	-	10.59	gene ZNF100 100b downstream (antisense)

resampling and as such can give valuable pointers to HERVs that should be studied more closely.

There are few examples of active and potentially protein-coding HERVs. Most of the active HERVs (such as HERVs B and D discussed in section 4.1) display fragmented expression that could be explained by RNA mediated activity or by function as exons, beginings or ends of nearby human genes.

Some of the observed expression may be due to active non-retroviral repeat sequences. In this study we wanted to study fragmented HERVs in addition to the full-length elements. The fragmented HERVs are harder to detect and in the process of ensuring that the more mutated HERVs are not missed some elements that are combinations of retrovirus and retrotransposon sequences may be included into the RetroTector produced HERV set. Actually, some of the most active HERVs were found to contain sequence portions which the RepeatMasker[2]

[2] RepeatMasker is a widely used program for detecting repeats. It relies on a database, the RepBase [5], of consensus sequences for various kinds or repeats. The repeat annotations in the UCSC Genome Browser come from RepeatMasker predictions. A

[14] had annotated as L1, L2 or SVA repeats. The fact that we observe expression similar to L1 or SVA elements is interesting as these elements have been shown to be active recently: The comparison of human chimpanzee genomes revealed thousands of species specific integrations for both L1 and SVA elements [18]. Our results indicate both L1 and SVA elements are still actively expressed in the human genome.

The hidden Markov mixture model can also be applied to other kinds of mRNA data sources or to other types of repetitive elements. For example our method could be used as a post-processing step in a RT-PCR reaction [9] where a broadly targeting primer (all members of a HERV group are amplified) has been used. When the PCR products are sequenced, they can be compared to the members of the targeted HERV group using our hidden Markov mixture model. This way it can be determined which elements within the group of very similar sequences are active. This can be done in one or several tissues.

Acknowledgments

We would like to acknowledge the Microbes and Man (MICMAN) research programme of the Academy of Finland (decision 202502). The author belongs to the Adaptive Informatics Research Centre, which is a Centre of Excellence of the Academy of Finland. We would like to thank Göran Sperber and Jonas Blomberg, Uppsala University, for RetroTector data. We are also grateful to Panu Somervuo, University of Helsinki, for his help with the HMM code. Finally, we would like to thank Samuel Kaski and Jaakko Peltonen, Helsinki University of Technology, and Jonas Blomberg for their valuable comments during the research and writing processes.

References

1. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–10, 1990.
2. J. Blomberg, D. Uschameckis, and P. Jern. Evolutionary aspects of human endogenous retroviral sequences and disease. In E. Sverdlov, editor, *Retroviruses and Primate Evolution*, pages 208–243. Eurekah Bioscience, 2005.
3. E. Lander *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.
4. D. J. Griffiths. Endogenous retroviruses in the human genome sequence. *Genome Biology*, 2(6):reviews1017.1–1017.5, 2001.
5. J. Jurka. RepBase update: a database and an electronic journal of repetitive elements. *Trends in genetics*, 16(9):418–420, 2000.
6. T.-H. Kim, Y.-J. Jeon, W.-Y. Kim, and H.-S. Kim. HESAS: HERVs expression and structure analysis system. *Bioinformatics*, 21(8):1699–1700, 2005.

discussion on the differences between RetroTector and RepeatMasker can be found from [15].

7. A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–31, 1994.
8. A. Muir, A. Lever, and A. Moffett. Expression and functions of human endogenous retroviruses in the placenta: An update. *Placenta*, 25(Suppl. 1):S16–S25, 2004.
9. S. Muradrasoli, A. Forsman, L. Hu, V. Blikstad, and J. Blomberg. Development of real-time PCRs for detection and quantitation of human MMTV-like (HML) sequences. HML expression in human tissues and cell lines. *J Virol Meth*, 136:83–92, 2006.
10. P. N. Nelson, P. R. Carnegie, J. Martin, H. Davari Ejtehadi, P. Hooley, D. Roden, S. Rowland-Jones, P. Warren, J. Astley, and P. G. Murray. Demystified ... human endogenous retroviruses. *Molecular Pathology*, 56:11–18, 2003.
11. M. Oja, J. Peltonen, J. Blomberg, and S. Kaski. Methods for estimating human endogenous retrovirus activities from EST databases. *BMC Bioinformatics*, 8(Suppl 2):S11, 2007.
12. E. M. Ostertag, J. L. Goodier, Y. Zhang, and H. H. J. Kazazian. SVA elements are nonautonomous retrotransposons that cause disease in humans. *The American Journal of Human Genetics*, 73(6):1444–51, 2003.
13. W. Seifarth, O. Frank, U. Zeifelder, B. Spiess, A. D. Greenwood, R. Hehlmann, and C. Leib-Mösch. Comprehensive analysis of human endogenous retrovirus transcriptional activity in human tissues with a retrovirus-specific microarray. *Journal of Virology*, 79(1):341–52, 2005.
14. A. F. A. Smit, R. Hubley, and P. Green. RepeatMasker open-3.0., 1996-2004. http://www.repeatmasker.org.
15. G. Sperber, P. Jern, T. Airola, and J. Blomberg. Automated recognition of retroviral sequences; RetroTector©. *Nucleic Acids Research*, 2007. Accepted with revision.
16. Y. Stauffer, G. Theiler, P. Sperisen, Y. Lebedev, and C. V. Jongeneel. Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues. *Cancer Immunity*, 4(2), 2004.
17. A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *PNAS*, 101(16):6062–6067, 2004.
18. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437:69–87, 2005.
19. eVOC ontologies. http://www.evocontology.org.
20. Expressed sequence tags database (dbEST). http://www.ncbi.nlm.nih.gov/dbEST/.