Journal homepage www.jzs.uos.edu.krd
Journal of Zankoi Sulaimani
Part-A- (Pure and Applied Sciences)

Evolution, and phylogeny of herv-s family in some closely related primates

Hawnaz Othman Najmalddin, Aso Ahmed Taher , and Shad Arif Mohammed

Department of Biology, Faculty of Science and Education Science, College of Science, University of Sulaimani.

email: shad.mohammed@univsul.edu.iq

Article info

Original: 21 Apr. 2015
Revised: 23 May 2015
Accepted: 31 May 2015
Published online:
20 Dec. 2015

Key Words:
HERV-S,
Endogenous,
Phylogeny,
Evolution,
Primate,
Cancer cell line

Abstract

Human endogenous retroviral element S-family (HERV-S) has been present within the human genome as well as various closely related primates. The expression pattern, copy number, chromosomal locations of the proviral element have been extensively studied. Here, we analyse the *pol* fragment of the HERV-S in human tissues, cancer cell lines and in seven other related primates. Phylogenetic analysis of the HERV-S *pol* gene suggests division of the family into two groups among the tested species that have arisen throughout the primate evolution by gene duplication as well as other genetic alterations. Also, the tree analysis of human tissues and cancer cells suggest that expression of the chromosome X *pol* fragment is identical to expression of the *pol* fragment in cancerous cells more than its relation to the *pol* sequence on other human chromosomes. To sum up, our data provide a clarification about the dynamic evolutionary characteristics as well as phylogenetic relationships of HERV-S family *pol* gene in primates.

Introduction

Endogenous retroviruses (ENVRs) are groups of retroviral elements and are considered as genetic leftovers of ancient exogenous viruses that had infected germline and successfully dominated in primate genome. They are now vertically transmitted (Fischer et al., 2014**).** These elements have been reported to be associated with many diseases, including multiple sclerosis, psoriasis, systemic lupus, as well as many types of cancer (Li et al., 2013) and thus they are subjects of many transcription and translation studies.

These elements are thought to be integrated into the ancestral genome of primates around ten to fifty million years ago (Yi et al., 2004), and depending on their integration site in the genome they have shown important roles both beneficial and detrimental. Some might have entered the coding region of a gene and disrupted its function or translation. Many others may exhibit transcription regulation roles when inserted upstream or downstream of a coding sequence. This regulatory function is exerted by the presence of the long terminal repeats (LTRs) acting as enhancers or promoters (Bhardwaj and Coffin, 2014).

Human endogenous retroviruses (HERVs) comprise around 7-8% of the human genome (Jern et al., 2004; Romanish et al., 2010). Some are inactive as a result of major deletions and genetic alterations. However, some others are actively transcribed in normal tissues such as brain, and thymus (Yi et al., 2004), placenta (Kjellman et al., 1999; Kurth and Bannert, 2010) as well as cancer cell lines like HERV-K family (Bhardwaj and Coffin, 2014).

HERV-S family is a type of human endogenous retrovirus elements that has entered the human germline around 40 Mya (Yi, *et al*., 2004). The element was first identified by (Tristem, 2000) using BLAST search and it is 6.7 Kb long. It has a complete LTR-*gag*-*pol*-*env*-LTR structure. Classification of the proviral element is based on the serine tRNA primer binding site. The detailed structure of the element was then annotated by (Yi et al., 2004) using Pfam database.

The aim of the current study is to firstly identify HERV-S family *pol* DNA sequence in chromosomes of closely related primates and discover their phylogenetic relationship based on the *pol* fragment. Secondly, to understand the disease association of HERV-S family in the human genome. Last but not the least, to estimate the evolutionary time of the *pol* fragment among the primate groups.

The objectives of this research include constructing phylogenetic trees and identifying the *pol* fragment within different human chromosomes. Also, to identify evolutionary divergence time as well as calculating pairwise distance within the primate species groups. Finally, the relationship between the *pol* fragment expression on different chromosomes with diseases like cancer will be studied based on phylogenetic tree analysis of *pol* cDNA sequences from different human tissues and cancer cell lines.

Materials and methods

Detection of HERV-S family pol sequences

To detect pol DNA sequences, BLAST (Altschul et al., 1997) method was used which depends on similarity ratios of the sequences. The *pol* region of HERV-S family on human chromosome X was used as a query sequence and six other primates were used as subjects for analysis (Shown in table 1)

Table 1 Genebank accession codes of seven primate species Chromosome X

Species	Scientific name	Accession	Species	Scientific name	Accession
Human	*Homo sapiens*	AC004385	Orangutan	*Pongo abeli*	CM000573.1
Chimpanzee	*Pan tryglotydes*	NC_006491.3	Rhesus monkey	*Macaca mulatta*	CM001273.1
Gorilla	*Gorilla gorilla*	NC_018447.1	Crab-eating monkey	*Macaca fascicularis*	NC_022292.1
New-world monkey	*Callithrix jacchus*	NC_013918.1			
Olive baboon	*Papio anubis*	NC_018172.1			

Alignment methods and amino-acid sequence retrieval

Potential amino acid sequences were generated for the available *pol* sequences by selecting the translated target sequences producing significant hits in the NCBI BLAST database. Then, the DNA sequences were translated in MEGA 6 program. The translated sequences were then aligned using MUSLCE (Edgar, 2004)

In order to prepare the alignment for phylogenetic analysis, a consensus sequence was generated for the aligned sequences using consensus maker function of HIV database. The consensus sequences were then analysed and further annotated manually.

pol fragment analysis

The translated sequence of HERV-S pol from MEGA was analysed to find possible open reading frames (ORFs) that can produce a protein. Using Expasy translate tool, six open reading frames were tested for functionality and finding domains. One of which gave valuable hits for RT-pol which was then analysed using Expasy/prosite database.

Results and Discussion

Structure of HERV-S family *pol* gene

The structure of HERV-S element was previously identified by (Tristem, 2000) by using the available NCBI BLAST search, stating that the general structure of the element is 5' LTR-gag-*pol*-env-LTR 3' is evident in human genome. Later on the detailed structure of the element was further investigated using Pfam HMM database protein domain search, shown in figure 1. (Yi et al., 2004). Here, we examined the structure of HERV-S *Pol* gene sequence using Pfam HMM and domain analysis of Prosite from Expasy database. The ORF of HERV-S *Pol* was retrieved from Expasy translation tool and submitted to Expasy/ Prosite for structure analysis. The results showed high rates of similarities with HERVs RT-*POL pol*yprotein molecules. The *pol* domain spans amnioacid 41-185 that codes for Reverse Transcriptase if expressed. Also, three other metal binding and catalytic domains at amino acid locations 111, 178, and 179 were detected. They all were magnesium binding and catalytic domains coded by Aspartic acid (GAC) summarised in figure 1. The expression pattern of this fragment was earlier investigated by (Yi et al., 2004) in both normal and cancer cell lines. The study showed that the element is highly expressed in cancer cells while it is not expressed in normal human and monkey tissues except for normal brain and thymus tissues. One could relate the importance of expression of this gene in cancer cell lines and use it as target for further expression studies.

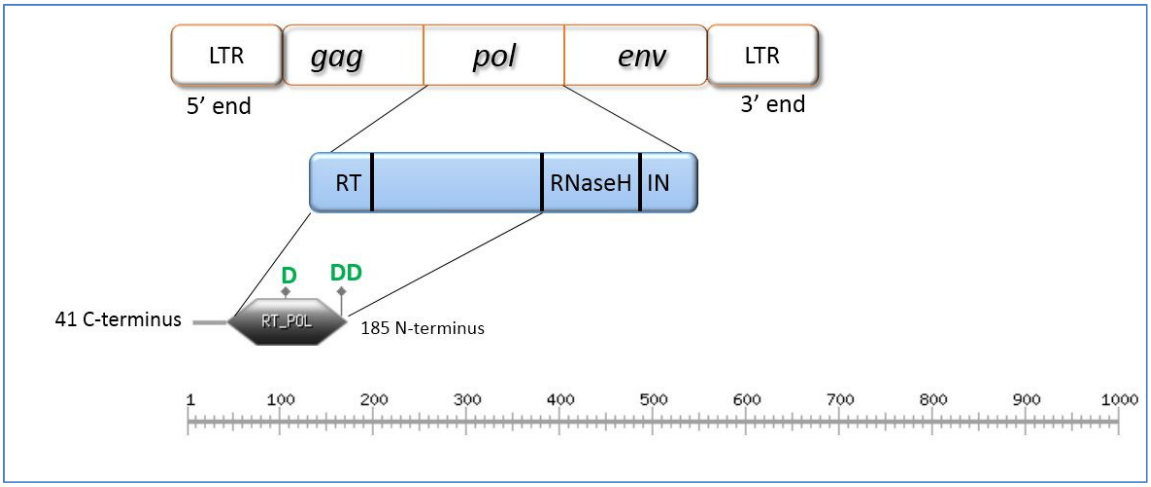

Figure 1 Analysis of HERV-S pol gene fragment. The HERV-S element is composed of two LTRs flanking gag-*pol*-env genes. The *pol* fragment is composed of RT(reverse transcriptase), RNaseH, and IN (Integrase) domains. There are three metal binding domains at positions 111.178, and 179 denoted as D (Aspartic acid). The scale ruler is to show the number of amino acid sequence span of the RT-*POL* domain. Domains are annotated based on pfam and expasy/prosite databases.

Analysis of HERV-S *pol* gene sequence

In this study, the HERV-S *pol* gene sequence was analysed in two different patterns. Firstly, the gene sequence was retrieved from NCBI/nucleotide database for different primate species and aligned for analysis figure 2. Secondly, available HERV-S *pol* gene sequences were retrieved from the same database for different human chromosomes and from different human tissues and cancer cell lines. After that, the gene sequences were aligned for analysis Figure 3.

Figure 2 Alignment of HERV-S* POL *amino acid sequence in seven different primate species. The highly conserved regions are bordered in red. Dashes represent missing sites at particular positions. Analysis was performed using MUSCLE method (Edgar, 2004).

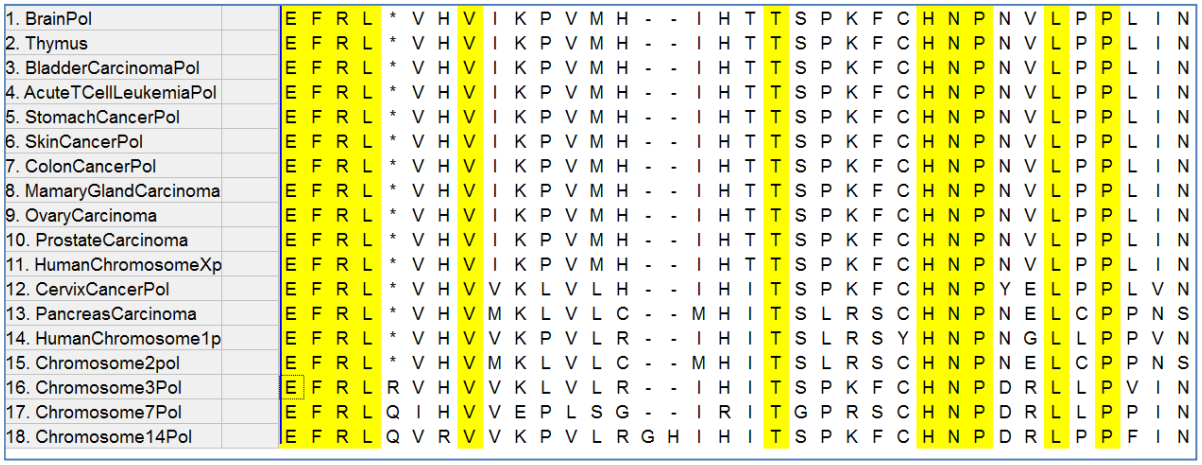

Figure 3 Pol sequence alignment in different human tissuses and cancer cells (C-terminus). The conserved cites are highlighted in yellow. *pol* sequences from fifteen different cell lines as well as pol sequences from four different human chromosomes were retrieved and aligned for analysis.. The alignment was carried out using MUSCLE method under the default settings (Analysis was performed using MUSCLE method (Edgar, 2004).

Phylogenetic analysis of HERV-S *pol* sequence

Phylogenetic analysis of *pol* DNA sequence among primates

In order to better understand the evolutionary relationships within HERV-S family, three methods for phylogenetic tree constructions were used (Neighbor joining (NJ), Maximum-Likelihood (ML), and Divergence time tree based of Maximum likelihood tree). The NJ tree showed the same results as the ML tree. For this purpose, four closely related and four distantly related primate species were subjected to the phylogenetic analysis, and a bootstrap of 100 replications was used to run the trees Shown in (figure 4, and figure 5). The Gene-Bank accession codes of the sequences are shown in (table 1).

As shown in the phylogenetic trees, the HERV-S family *pol* sequences were grouped into two main groups (**A** and **B**) throughout their evolutionary divergence. This might indicate the amplification frequency of the HERV-S *pol* family among the different species of the two groups. The amplification frequency of the *pol* in human genome, however, was estimated to be at least three times since the original integration of HERV-S to human genome (Yi et al., 2004). Our data confirms their results, as shown in (figure 6).

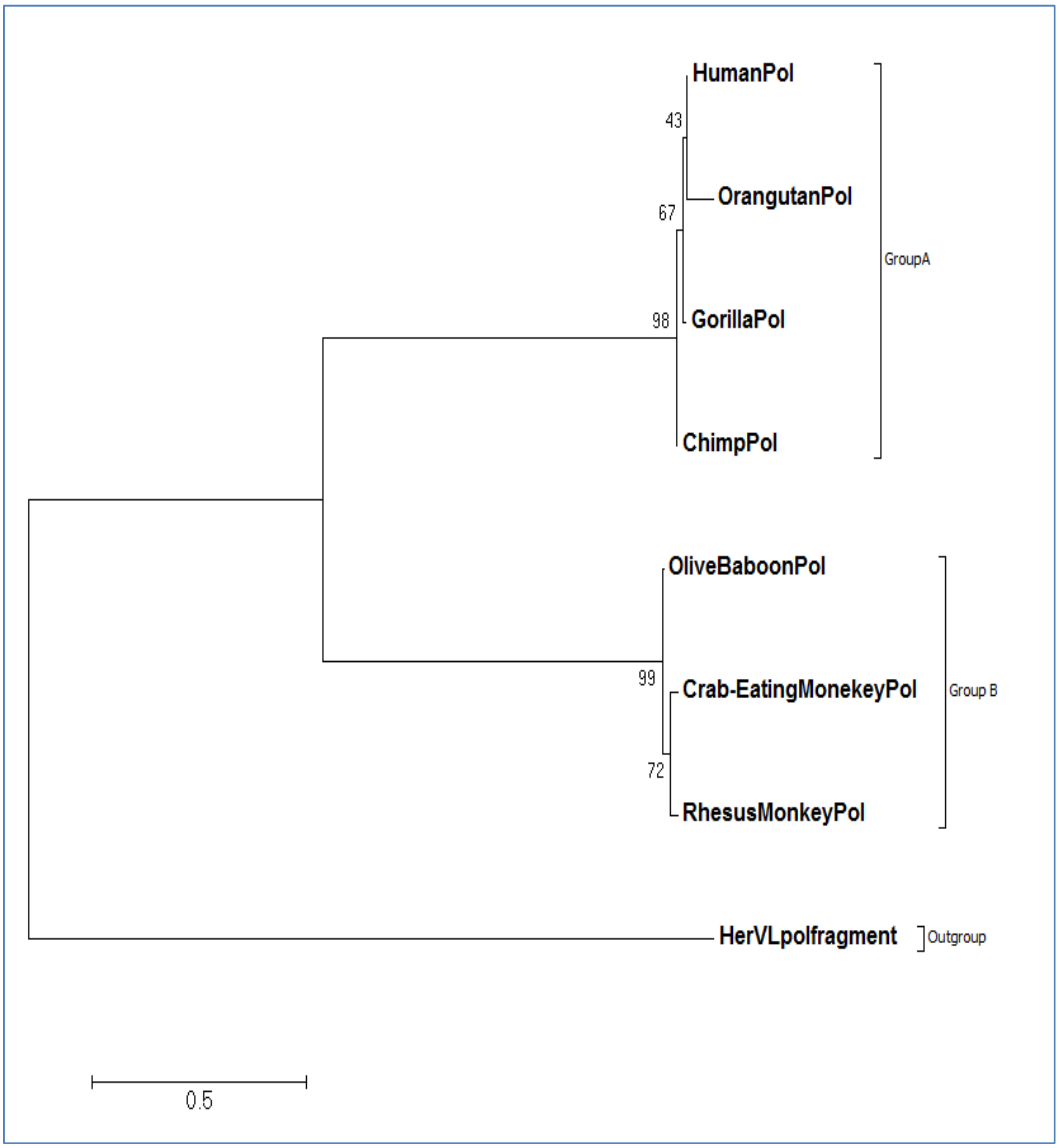

Figure 4 Molecular Phylogenetic analysis by Neighbor-Joining method (Saitou and Nei, 1987).

The optimal tree with the sum of branch length = 4.03295629 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (Zuckerkandl and Pauling, 1965) and are in the units of the number of amino acid substitutions per site. The analysis involved 8 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 132 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura et al., 2013). The species were grouped into two groups A, and B based on their evolutionary relationships, also, HERV-L family was used as an outgroup due to its close relationship with HERV-S family.

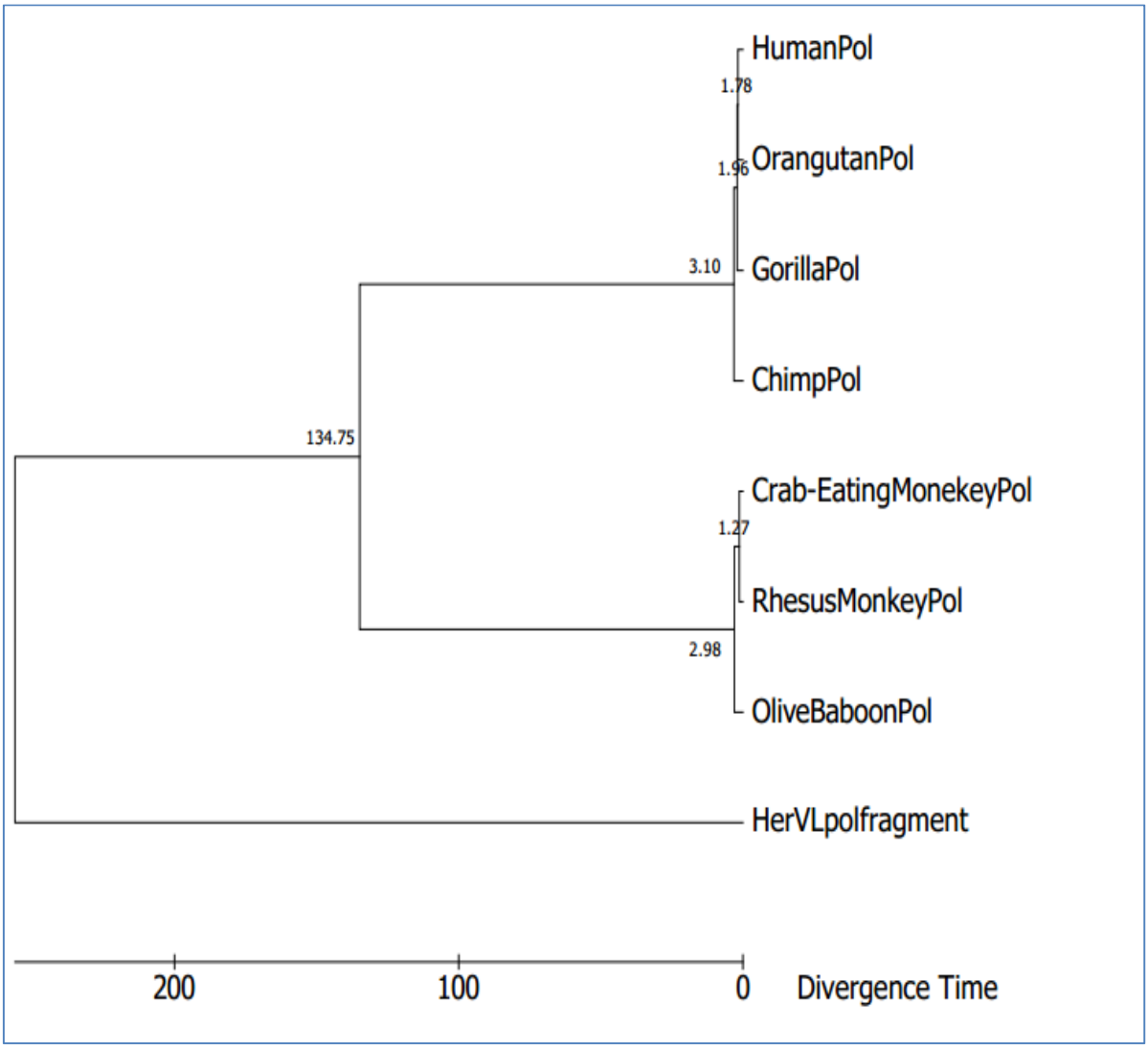

Figure 5 Molecular Phylogenetic analysis by Maximum Likelihood method (timetree).

The timetree shown was generated using the RelTime method (Tamura et al., 2012). Divergence times for all branching points in the topology were calculated using the Maximum Likelihood method based on the JTT matrix-based model (Jones et al., 1992). Bars around each node represent 95% confidence intervals which were computed using the method described in (Tamura et al., 2012). The estimated log likelihood value of the topology shown is -1278.5669. The tree is drawn to scale, with branch lengths measured in the relative number of substitutions per site. The analysis involved 8 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 132 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura et al., 2013).

Phylogenetic analysis of *pol* DNA sequence in human tissues and cancer cell lines

As shown in figure 6, the *pol* fragment were retrieved using NCBI BLAST (Altschul et al., 1997), the sequences were subjected to alignment as shown in figure 3. The alignment was then used to construct phylogenetic trees. The divergence timetree was then created using a maximum likelihood tree as a base tree, the results showed 100% similarities between human Chromosome X *pol* fragment and the *pol* fragments expressed by ten different human cancer cell lines. Which indicates the important roles and disease association of this pol fragment on Chromosome X, especially in these cancerous cells.

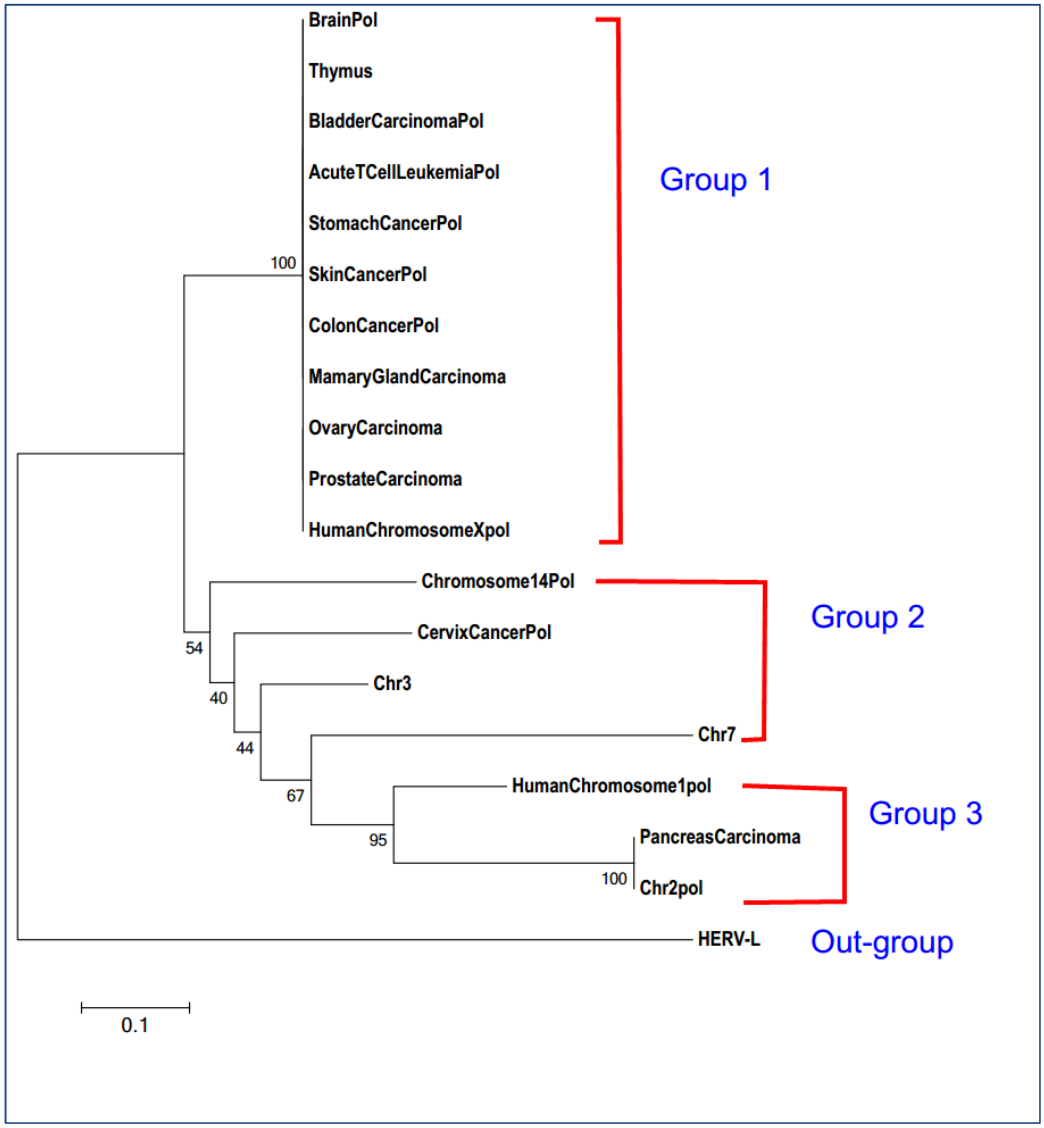

Figure 6Molecular phylogenetic analysis by NeighborJoining method (Saitou and Nei, 1987). The optimal tree with the sum of branch length = 2.21830489 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (Zuckerkandl and Pauling, 1965) and are in the units of the number of amino acid substitutions per site. The analysis involved 19 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 129 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura et al., 2013)

HERV-S pol Evolutionary analysis

Calculation of pairwise distance among different primate species

The phylogenetic tree (time tree) data in figure 5 was used as a base to calculate the pairwise divergence for the two split groups by applying Kimura-2 parameter method (Kimura, 1980). Poisson correction model was used to analyse the 8 nucleotide sequences containing the outgroup sequence. Then the average distance was calculated in three categories (overall, within the two groups and between the two groups). The overall average distance among the eight sequences were 0.6. Also, the average distance among the species within group A was 0.027 and within group B was 0.012. These data were used to calculate the evolutionary time between the two groups using T=d/2r when **T** stands for evolutionary time, **d** stands for divergence and **r** for evolutionary rate. The accepted evolutionary rate (0.3%) that was chosen by (Yi et al., 2004) was also used here to calculate the evolutionary time HERV-S pol family among different primate species. Table 2 summarises the evolutionary time among the species analysed.

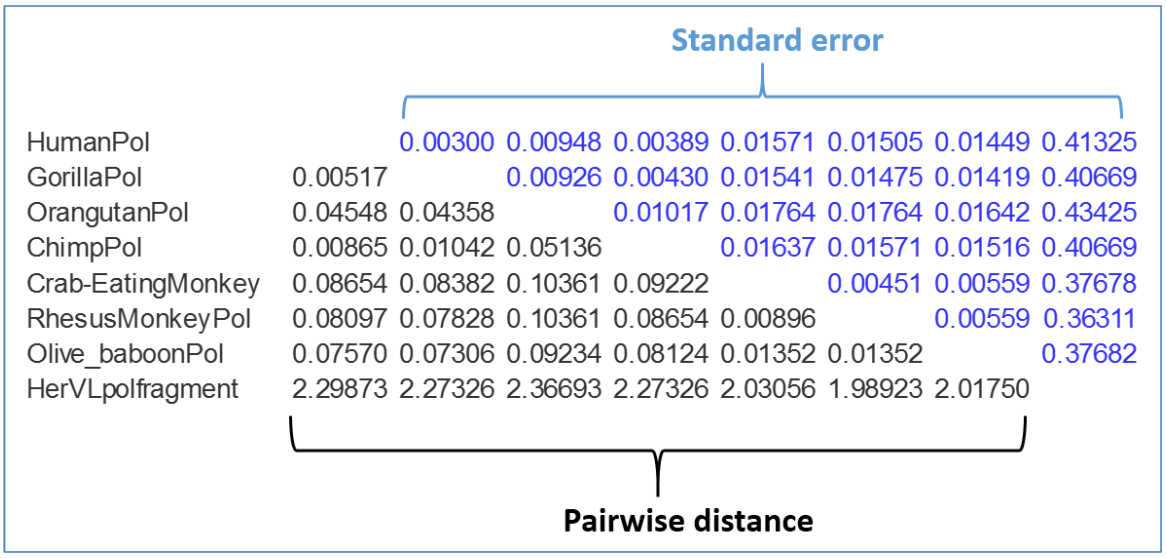

Figure 7 Estimates of Evolutionary Divergence between Sequences.

The number of base substitutions per site from between sequences are shown. Standard error estimate(s) are shown above the diagonal and were obtained by using analytical formulas. Analyses were conducted using the Kimura 2-parameter model (Kimura, 1980). The rate variation among sites was modeled with a gamma distribution (shape parameter = 1). The analysis involved 8 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. Analyses were conducted using the Poisson correction model (Zuckerkandl and Pauling, 1965).All ambiguous positions were removed for each sequence pair. There were a total of 2940 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura et al., 2013).

Table 2 evolutionary time and divergence of HERV-S pol among different primate species.

Groups (HERV-S *pol*)	Percentage of Divergence within groups	Evolutionary time r= 0.3% Myr
A	0.027	4.5
B	0.012	2

According to the evolutionary time rates used here (0.3% per million years) in group A and B, it can be estimated that the members of the groups have proliferated as multiple copies of the sequences around 4.5 Myr ago for the species of the first group, and 2 Myr ago for the species in group B.

The species studied in this research, although genetically distinct, all exhibited HERV-S *pol* fragment when we analysed their Chromosome X. On the other hand, Polymerase Chain Reaction (PCR) analysis of *pol* DNA sequence in Lemur and Galago showed negative results which means the absence of the fragment or the element in these two species (Yi et al, 2004). One would conclude that the ancestor of Lemur and Galago had diverged from the ancestor of the two groups of primates tested in the current study before the entrance on the HERV-S into the genome.

Using Tajima's relative rate test, human HERV-S *pol* (group 1) and Olive baboon *pol* (group 2) (shown in figure 4) were tested. Human Herv-L *pol* was used as an out-group for the test. The results showed that the differences in the *pol* fragment were not significant (X^2 *p*-value= 0.73). Thus, the *pol* DNA sequence on the Chromosome-X of the two phylogenetically distinct groups is from the same origin, and they all share the same common ancestor.

Table 3 Results from the Tajima's test for 3 Sequences

Configuration	**Count**
Identical sites in all three sequences	197
Divergent sites in all three sequences	9
Unique differences in Sequence A	13
Unique differences in Sequence B	9
Unique differences in Sequence C	219

The equality of evolutionary rate between sequences A (*HumanPol*) and B (*Olive baboonPol*), with sequence C (*HerVLpolfragment*) used as an outgroup in Tajima's relative rate test (Tajima, 1993). The χ^2 test statistic was 0.73 (*P* = 0.39377 with 1 degree[s] of freedom) . *P*-value less than 0.05 is often used to reject the null hypothesis of equal rates between lineages. The analysis involved 3 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 447 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura et al., 2013).

Conclusions

In conclusion, HERV-S family *pol* sequence is expressed differently in different human tissues and cancer cell lines. The *pol* fragment expressed on Chromosome X is closely related to the pol fragment expressed in many cancer cell lines which shows the relationship between the two, as well as cancer related disease association of the chromosome X *pol* sequence.

HERV-S *pol* sequence is highly conserved among human, chimpanzee, gorilla and orangutan. However, due to high mutation rates, gene duplication and transposition events, the gene is less conserved in the other three species tested (rhesus monkey, crab eating monkey, and olive baboon).

References

1. ALTSCHUL, S. F., MADDEN, T. L., SCHÄFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research,* 25**,** 3389-3402.
2. BHARDWAJ, N. & COFFIN, J. M. 2014. Endogenous retroviruses and human cancer: is there anything to the rumors? *Cell host & microbe,* 15**,** 255-259.
3. CHAKRABORTY, R. 1977. Estimation of time of divergence from phylogenetic studies. *Canadian Journal of Genetics and Cytology,* 19**,** 217-223.
4. EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research,* 32**,** 1792-1797.
5. FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution***,** 783-791.
6. FISCHER, S., ECHEVERRÍA, N., MORATORIO, G., LANDONI, A. I., DIGHIERO, G., CRISTINA, J., OPPEZZO, P. & MORENO, P. 2014. Human endogenous retrovirus np9 gene is over expressed in chronic lymphocytic leukemia patients. *Leukemia research reports,* 3**,** 70-72.
7. FOLEY, B., LEITNER, T., APETREI, C., HAHN, B., MIZRACHI, I., MULLINS, J., RAMBAUT, A., WOLINSKY, S. & KORBER, B. 2013. HIV sequence compendium 2013. *Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LAUR***,** 13-26007.
8. JERN, P., SPERBER, G. O. & BLOMBERG, J. 2004. Definition and variation of human endogenous retrovirus H. *Virology,* 327**,** 93-110.
9. JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences: CABIOS,* 8**,** 275-282.
10. KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution,* 16**,** 111-120.
11. KJELLMAN, C., SJÖGREN, H.-O., SALFORD, L. G. & WIDEGREN, B. 1999. HERV-F (XA34) is a full-length human endogenous retrovirus expressed in placental and fetal tissues. *Gene,* 239**,** 99-107.
12. KURTH, R. & BANNERT, N. 2010. Beneficial and detrimental effects of human endogenous retroviruses. *International journal of cancer,* 126**,** 306-314.

13. LI, S., LIU, Z., YIN, S., CHEN, Y., YU, H., ZENG, J., ZHANG, Q. & ZHU, F. 2013. Human endogenous retrovirus W family envelope gene activates the small conductance Ca 2+-activated K+ channel in human neuroblastoma cells through CREB. *Neuroscience,* 247**,** 164-174.
14. ROMANISH, M., COHEN, C. & MAGER, D. Potential mechanisms of endogenous retroviral-mediated genomic instability in human cancer. Seminars in cancer biology, 2010. Elsevier, 246-253.
15. SAITOU, N. & NEI, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution,* 4**,** 406-425.
16. SHIMODAIRA, H. & HASEGAWA, M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics,* 17**,** 1246-1247.
17. TAJIMA, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics,* 135**,** 599-607.
18. TAMURA, K., BATTISTUZZI, F. U., BILLING-ROSS, P., MURILLO, O., FILIPSKI, A. & KUMAR, S. 2012. Estimating divergence times in large molecular phylogenies. *Proceedings of the National Academy of Sciences,* 109**,** 19333-19338.
19. TAMURA, K., STECHER, G., PETERSON, D., FILIPSKI, A. & KUMAR, S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution,* 30**,** 2725-2729.
20. TRISTEM, M. 2000. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *Journal of virology,* 74**,** 3715-3730.
21. YI, J.-M., KIM, T.-H., HUH, J.-W., PARK, K. S., JANG, S. B., KIM, H.-M. & KIM, H.-S. 2004. Human endogenous retroviral elements belonging to the HERV-S family from human tissues, cancer cells, and primates: expression, structure, phylogeny and evolution. *Gene,* 342**,** 283-292.
22. ZUCKERKANDL, E. & PAULING, L. 1965. Evolutionary divergence and convergence in proteins. *Evolving genes and proteins,* 97**,** 97-166.