A profile-Hidden Markov Model approach for detecting endoretroviral LTR sequences in the Platypus genome

Hugo Miguel Martins

Yu Qian

Bioinformatics Research Centre, Aarhus University, October 2009

Index

Introduction 3

Background 4

Methods 7

Results 8

Discussion 12

Platypus models 13

Opossum models 18

Conclusions 30

References 32

Appendices

Appendix A – Pictogram of the Monodelphis domestica sequence clusters

Appendix B – Output files for the results of the pHMM runs

Appendix C – Output files for the distance matrices of the pHMM hits

Appendix D – Output files of the test runs for the pHMM models

Introduction

With the advent of high throughput sequencing methods, more and more complete genomes tend to be available for public research. The platypus (*Ornitorhynchus anatinus*) was the last animal to have its complete genome sequenced and published[1] opening new doors to the study of this very strange mammal.

In this work, we aim to build a comparative based profile Hidden Markov Model (pHMM) approach in detecting retrotransposable elements. The project will focus on elements of the LTR class, using available LTR loci from the Platypus itself and pHMMs built on known LTR loci from the closest relative available in the genome databases, the grey-tailed opossum (*Monodelphis domestica*). We will test the feasibility of using pHMMs to detect these genomic structures and hopefully contribute with novel sites.

Background

Endogenous retroviruses are derived sequences from ancient infections of germline cells by exoretroviruses[2]. The provirus resulting from these infections are passed on vertically, becoming a permanent feature of the organism's genome, and subject to mutations and evolutionary processes. Many are partially excised or suffer deleterious mutations that render their once coding genes inoperative.

A typical retrovirus genome contains several key structural features[3]. Coding genes for viral proteins such as the *gag*, *pol* and *env*, and signal sequence such as the PBS – primer binding site or the PSI – packaging site. Another and a quite important structure is the LTR, or long terminal repeat. Retroviruses possess two LTRs, one in each extremity of their genomes. The LTR is the control center for the viral gene expression[4] and has many similarities to a typical eukaryotic promoter, with transcriptional enhancers and some specialized regulatory elements[5]. All requisite signals for gene expression can be found in LTRs: enhancer, promoter, transcription initiation site, transcription terminator and polyadenlyation signal. The enhancer and other transcription regulatory signals are contained in the U3 region of the 5' LTR, after which one can find the TATA box, just about 25 bp before the R sequence. When integrated, a proviral 5' LTR acts as an RNA polymerase II promoter which conducts the transcription process that begins, by definition, at the beginning of the R sequence, proceeding through the U5 along the rest of the provirus. The addition of a poly A tract just after the R sequence in the 3' LTR terminates this process. Interestingly, although both 5' and 3' LTRs have the same sequence arrangement, they perform different functions after insertion.

Being such a common feature in retroviruses, LTRs are optimal targets for detection of integrated retroviral elements. Many recent complex detection pipelines include structural information based on the LTR features, in order to detect *de novo* LTRs in the genome. Some extended models also probe a confined space between two nearby, similar LTR hits for proviral gene remains in order to achieve a full endoretroviral characterization[6].

There are several tools available to detect LTRs, from which we cite a few examples. RepeatMasker[7] annotation can be used to detect interspersed and low complexity LTR based on sequence identity. However, it might miss low copy LTRs and it might fail to detect new LTR as it is mainly based on homology. LTR-STRUC[8] uses structural features of primer binding site, the polypurine tract and the dinucleotides ends of each LTR and LTR insertions sites to identify LTRs. RetroTector[9] focuses on detecting endogenous viruses (ERVs) in genomic material in a repeat-independent way, but it might be weak in detecting single LTR. The initial purpose of our project is to detect LTR using landmarks and fingerprints based on previous homology information.

Algorithms for multiple sequence alignment such as Clustal[10] are sensitive to the number and choice of sequences, and if many sequences with variable amounts of inserts exit, they can destroy the alignment. In case of ERV (endogenous retroviruses), which are very numerous and diverse, HMM seems better at extracting information from such sequences.

Functional biological sequences typically come from families and conservative regions, and many of the sequence analysis are based on identifying the relationship of an individual sequence to a sequence family. Sequences in a family will have diverged from each other by duplication or translocation in the evolutionary history, it might be more difficult to analyses them by multiple alignment. Derived from Hidden Markov Models, profile Hidden Markov Models (pHMM) are widely used for searching databases for remotely homologous sequences.

In general, pHMMs are statistical models of multiple sequence alignments[11],[12]. They capture position-specific information about how conserved each column of the alignment is and what residues are likely to be there. There are several forms of pHMM and we use HMMer[13] to build our model of ERV and search for the target sequence. The basic form of a profile HMM is a linear set of Match (M) states, one per consensus column in the multiple alignments. Each M stated emits a single residue with a probability score that is determined by the multiple alignment or what we called "training sequences" in this project. Each match state carries a vector of 4 probabilities for scoring the 4 nucleotides. Each match state has an I and a D state associate with it, with insertion state also carrying 4 emission probabilities. We call a group of three states (M/D/I) at the same consensus position in the alignment a "node". These states are interconnected with arrows called state transition so that there is either a match state or a deletion state in each node. Insertion occurs between nodes, and I states have a self-transition (I to I), allowing one or more inserted residues between consensus columns. The model begins and ends with dummy non-emitting states, B and E.

Parameters of the model are the transition probabilities and emitting probabilities. Generally, we just count up the number of times each transition or emitting is used in the training sequences to get an estimate of parameters. Sometimes, it is also necessary to add some pseudocounts to avoid zero probabilities and it means observed counts of emissions (residues) and transitions (insertions and deletions) in a multiple alignment are combined with Dirichlet priors to convert them to probabilities in an HMM. In order to calculate the probability of sequence given the model, recursive enumeration of possible sequences under certain rules is needed. The simplest model implies a geometric distribution over insertion length, though biologically it is not a realistic model, it is computationally easy to realize.

To score a match to a hidden Markov Model, we can either use Viterbi equations to get the most probable alignment of a sequence x together with its probability $P(\pi,x|M)$, or use forward algorithm to calculate the full probability of x summed over all possible

paths P (x|M). In HMMER, we use two scoring criteria: E-value and bit score. The bit score is defined as:

$$S = \log_2 \frac{\quad}{\quad}$$

It reflects the ratio of probability that query sequence is a significant match to the probability that the null model is a match. In practice, the score is rescaled to make calculation easier and result readable.

The e-value is calculated from bit score and tells how many false positives we would have expected to see at or above this it. It measures the significance of the bit score, and unlike the bit score, its value is related the length of query sequence.

Actually, HMMer bit scores are relative to two null hypotheses. The first is the null model built into the profile HMM in hmmbuild command. The second is calculated on the fly for each alignment. Sometimes, we might get negative score but good e-value, it means that though the bit score is bad, it is still better than expected by chance and therefore suggestive of distant homology.

Methods

The available full-length platypus chromosomal sequences were obtained from UCSC Genome Browser[14]. Unassembled contigs were discarded. A summary of our database can be seen on table 1. Platypus and opossum LTR sequences were obtained from Repbase[15]. RepeatMasker annotated files for the platypus chromosomes were used as benchmarks to test our method's performance and were also obtained from UCSC Genome Browser. Sequences, otherwise noted, were aligned using standard Clustalw algorithm and the profile HMMs were built from these alignments using the HMMer software package. Two pHMMs were built for each training set according to both a global and local alignment reasoning, allowing for multiple hits and calibrated with random generated sequences of mean and standard deviation close to the full length of each individual model.

	Name	Length (Mb)		Name	Length (Mb)
Analysed chromosomes	Chr 1	47.4	Partially analysed / discarded from final analysis	Chr 7	38.9
	Chr 2	53.3		Chr 10	10.9
	Chr 3	57.9		Chr 11	6.62
	Chr 4	57.3		Chr 12	15.4
	Chr 5	23.9		Chr 18	6.43
	Chr 6	15.8		Chr X1	44.2
				Chr X2	5.49
				Chr X3	5.78
				Chr X5	27.0

Table 1 – Summary of analyzed data.

Results

The platypus pHMMs were built based on an alignment of seven available LTR sequences in RepBase, submitted by A. F. Smit and J. Jurka. Table 2 summarizes this training set. The alignment was trimmed to a smaller region of about 500 bp that encompassed the best aligned region. The platypus pHMMs for both local and global alignment were built and ran against each chromosome individually. Both pHMMs had the length of 430 states and were calibrated with 5000 random generated sequences of 400 bp in length. Table 3 presents a comparison between local and global alignment approaches and a performance analysis on the method based on benchmarking with RepeatMasker annotation for the studied chromosomes.

Sequence name	Family	Length (bp)
PlatERVK1_LTR	ERVK	389
PlatERVK2_LTR	ERVK	492
PlatERVP1_LTR	ERV1	382
PlatERVR1_LTR	ERV1	515
PlatERVR2_LTR	ERV1	473
PlatLTR20B	ERV3	713
PlatLTR35C	ERV3	1005

Table 2 – summary of the training set used to build the platypus-based profile HMM.

	Chr 1		Chr 2		Chr 3		Chr 4		Chr 5		Chr 6	
RepMask	135		115		168		137		38		50	
Method	hits	true	hits	true	hits	true	hits	true	hits	true	hits	true
Local	21	7	11	3	25	10	23	8	10	2	10	7
Global	13	4	5	1	11	3	11	2	3	2	6	2

Table 3 – Summary of hits for the playpus-based pHMM. Hits columns represent the total hits obtained by the method; True columns represent the hits that were validated by RepeatMasker annotation with a 100bp tolerance interval for start and end points. RepMask row indicates all LTR-based annotated features. This includes ERV loci, solo LTRs, Gypsy elements, among others.

Due to the large amount of training data for the opossum pHMMs, we decided to separate the sequences according to the fasta header information. From RepBase, most of the sequences contained family related information, which deemed them belonging to either ERV1, 2 or 3 family. Relying on this classification, we aimed to build four pHMMs – one for each family and a fourth one for non-classified LTR sequences. However, it was impossible to build an informative pHMM for these families, with the exception of ERV 3, since the sequence divergence within the same family was too high to ensure a valid model. Therefore, we decided to build several, smaller pHMMs, based on true genetic distance between all the opossum LTR sequences. The sequences were aligned using a fast clustalw algorithm and the Kimura distance[16] matrix between all

sequences was obtained and exported using the ape package for the statistical software R[17].

In order to detect meaningful distances that would allow us to build sequence clusters, we trimmed down the distance matrix by removing all Kimura distances greater than 0.8, and subsequently deleting all rows and columns which had less than 5 results below the cutoff. Strangely, the new distance matrix presented already formed sequence groups, arranged along the matrix diagonal. Figure 1 in Appendix A displays a colored overview of the detected sequence groups.

Overall, seventeen groups were used to build the same number of pHMMs based on opossum LTRs, named op01 to op17 for local alignment pHMMs and op01g to op17g for global alignment pHMMs. A summary of these pHMMs and their characteristics can be seen in Table 4.

HMM model	Sequences in training set	HMM Length (states)	Calibration dataset length (bp)
Op01	10	782	750
Op02	7	1031	1000
Op03	7	929	900
Op04	9	778	750
Op05	5	828	800
Op06	7	797	800
Op07	7	6904	5000
Op08	8	276	250
Op09	8	326	300
Op10	9	656	650
Op11	6	722	700
Op12	7	574	550
Op13	10	871	850
Op14	8	1125	1000
Op15	7	640	650
Op16	7	846	850
Op17	9	793	800

Table 4 – Summary of the training set size and calibration parameters for the 17 pHMMs built based on opossum (*Monodelphis domestica*) data from RepBase.

We then proceeded to run these pHMMs on several platypus chromosomes and assess the results. pHMMs op07 and op07g were build on a small cluster of whole ERV loci and turned out to take too much computing time in order to provide results within the allotted time, thus being discarded from the final analysis.

Table 5 shows a summary of results for all the pHMMs built based on the opossum LTR data based on a local alignment criteria, while Table 6 provides the same overview for global alignment pHMMs.

	Chr 1		Chr 2		Chr 3		Chr 4		Chr 5		Chr 6	
Model	hits	true	hits	true	hits	true	hits	true	hits	true	hits	true
Op01	28	0	15	0	21	0	15	0	15	0	4	0
Op02	3	0	8	0	7	0	4	0	2	0	0	0
Op03	14	0	17	0	21	0	16	0	1	0	3	0
Op04	13	0	7	0	12	0	6	0	1	0	0	0
Op05	95	0	108	0	107	0	91	0	44	0	20	0
Op06	5	0	4	0	9	0	12	0	2	0	1	0
Op08	13	0	10	0	16	0	15	0	3	0	5	0
Op09	12	0	12	0	16	0	15	0	1	0	3	0
Op10	3	0	4	0	4	0	4	0	3	0	0	0
Op11	5	0	9	0	10	0	14	0	7	0	4	0
Op12	2	0	2	0	0	0	2	0	1	0	1	0
Op13	32	0	29	0	37	0	41	1	15	0	10	0
Op14	33	0	37	0	52	0	43	0	25	0	13	0
Op15	75	0	75	0	83	0	93	0	41	0	18	0
Op16	11	0	16	0	17	0	18	0	9	0	3	0
Op17	26	0	20	0	27	0	33	0	16	0	10	0

Table 5 – Summary of hits for the opossum-based pHMMs with a local alignment strategy. Hits columns represent all the hits obtained for the method while True columns represent those hits validated in RepeatMasker.

	Chr 1		Chr 2		Chr 3		Chr 4		Chr 5		Chr 6	
Model	hits	true	hits	true	hits	true	hits	true	hits	true	hits	true
Op01	13	4	5	1	11	3	11	2	3	2	6	2
Op02	3	0	1	0	1	0	1	0	1	0	1	0
Op03	1	0	1	0	1	0	1	0	1	0	1	0
Op04	1	0	1	0	1	0	1	0	1	0	1	0
Op05	1	0	1	0	1	0	1	0	1	0	1	0
Op06	5	0	6	0	5	0	3	0	3	0	1	0
Op08	1	0	1	0	1	0	1	0	1	0	1	0

Op09	1	0	1	0	1	0	1	0	1	0	1	0
Op10	1	0	1	0	0	0	1	0	1	0	1	0
Op11	1	0	1	0	1	0	1	0	1	0	1	0
Op12	1	0	1	0	1	0	1	0	1	0	1	0
Op13	1	0	1	0	1	0	1	0	1	0	1	0
Op14	1	0	1	0	1	0	1	0	1	0	1	0
Op15	1	0	1	0	1	0	1	0	1	0	1	0
Op16	6	0	10	0	5	0	14	0	5	0	1	0
Op17	1	0	1	0	1	0	1	0	1	0	1	0

Table 6 - Summary of hits for the opossum-based pHMMs with a global alignment strategy. Hits columns represent all the hits obtained for the method while True columns represent those hits validated in RepeatMasker.

Discussion

RepeatMasker annotation in the Platypus genome for endoretroviral elements is, as of date, still quite poor. While we can benchmark our pHMM models built on available platypus LTRs on this annotation, the same is not valid for the comparative opossum pHMMs. Thus, we evaluate our meaningful hits based on phylogenetic relationships to the training data sequences.

In order to trim down our results, we consider a meaningful hit those which e-value is below 10^{-5} and have a high positive bit-score, of at least 10-15. However, there are some particular cases, namely in the opossum pHMMs, where some of the hits, albeit having very low e-values (at most 10^{-6}), also possess strongly negative bit scores. As stated in the HMMer manual, these hits may well be meaningful, as they can represent sequences somewhat phylogenetic distant, yet related, that were detected by a strict pHMM built on closely related sequences. The latter may well be the case of some of our opossum models, since all these pHMMs were built based on short genetic Kimura distances. Throughout the analysis of the results, we will keep an eye for these particular events and relax the bit score criteria for opossum-based models.

The following discussion focuses only on the global alignment pHMMs, since the local alignment ones provide mostly short hits with very low bit scores and high e-values. Although the number of overall hits and RepeatMasker-confirmed hits is larger for the local alignment approach, this is mainly due to the small size of those hits, many of them contained within a single hit from the global alignment approach. The majority of these hits are CT and GA repeats, or sporadic LTR fragments. As a complement to the results shown here, Appendix B comprises the result files for the pHMM runs, which include genomic coordinates, bit scores and e-values for each hit. Appendix C comprises the Kimura distance matrices built with the Phylip package[18] for each of the global pHMM training sets and respective hits. Phylogenies were built using the Phyml[19] software package using the following parameters: HKY substitution model, with full parameter estimation and slow topology search (SPR moves).

A) Platypus pHMM

Chromosome 1 - The global pHMM detected 13 hits, 4 of them verified by RepeatMasker – one belonging to the LTR35 family and three others to the ERVR1 family. A phylogenetic analysis adds extra detail to this annotation, clustering most of the hits around the LTR35 node and one hit close to the LTR20 node. The amount of hits in the phylogenetic vicinity of LTR35 with high bootstrap values on the internal nodes, while keeping small branch lengths, indicates that there are more, non-redundant, copies of this family in the platypus genome than currently annotated. A total of 3 results were excluded from the final phylogeny due to high e-values and/or low bit scores. Figure 1 below details the phylogenetic outcome of the platypus pHMM on chromosome 1.

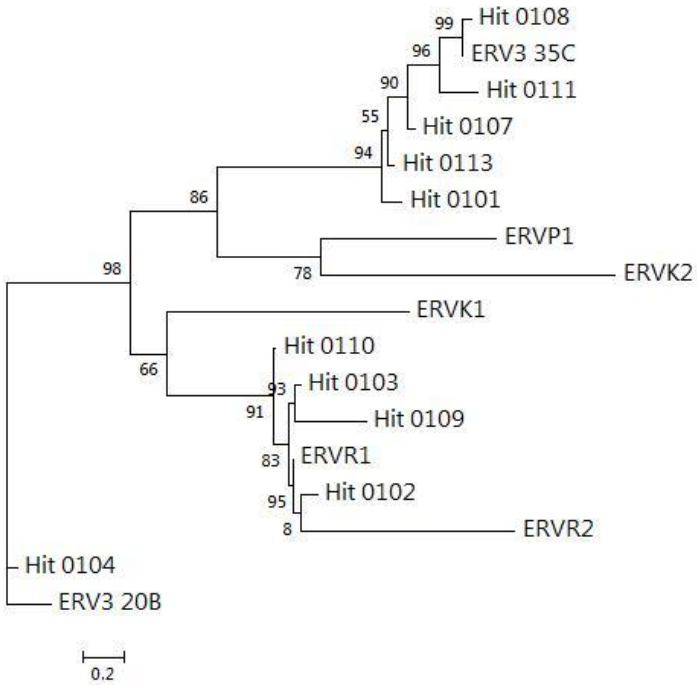

Figure 1 – Phylogenetic tree for the Chromosome 1 hits with the platypus-based pHMM (global alignment)

Chromosome 2 – The phylogenetic tree of results from this chromosome has a poorer resolution. Hit_0202 was confirmed by RepeatMasker as a true hit from the ERV R2 family. However, Hit_0204 strongly correlates with LTR35, indicating that this may be a missed annotated element in the chromosome. Hit_0201 appears to be either from or closely related to the LTR20 family, although the low bootstrap value of the internal node connecting these two leaves may cast some doubts as to the classification. A total of two hits from the original five were discarded from the final analysis due to poor bit scores and/or e-values.

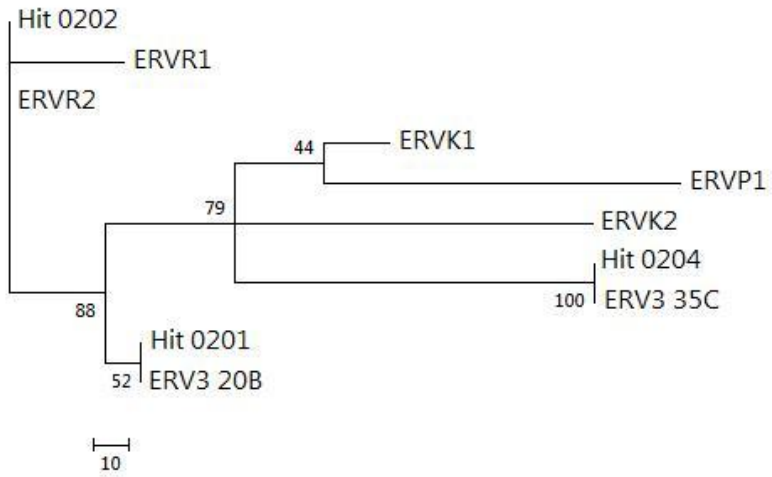

Figure 2 - Phylogenetic tree for the Chromosome 2 hits with the platypus-based pHMM (global alignment)

Chromosome 3 – Most of the phylogenetic relationships between hits and training data sets in this chromosome are fuzzy. RepeatMasker validated 3 out of our 11 hits, two of them belonging to the ERVR1 family while another belongs to the ERVR2 family. This is confirmed by the phylogenetic tree that shows Hit_0308 correlating with ERVR2 and Hits 0310 and 0304 correlating with ERVR1. LTR35 appears to have matches in our hits, namely Hit_0309 (bootstrap value of the internal node to ERV3 35C is 100). Hit_0301 and Hit_0307 are definitely related, but the unresolved node connecting them to the LTR35 leaf disallows a direct phylogenetic relationship to this family. Two results were excluded from the final analysis due to poor bit scores and/or e-values.

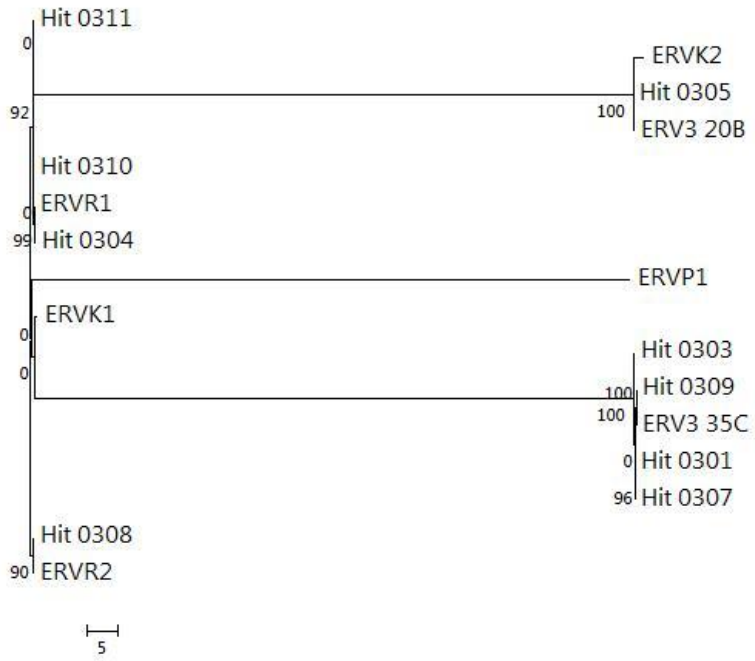

Figure 3 - Phylogenetic tree for the Chromosome 3 hits with the platypus-based pHMM (global alignment)

Chromosome 4 – Two out of 11 hits are confirmed by RepeatMasker in this chromosome, both belonging to the ERVR1 family. However, the phylogeny only seems to validate one of these hits, Hit_0409, while Hit_0406, although being placed near the ERVR1 leaf, has a poor bootstrap value for the connecting internal node. Other meaningful matches not present in the RepeatMasker annotation are Hits 0407 and 0411 that can be related to the LTR20 family; and Hits 0404 and 0410 that can be related to the LTR35 family. Two hits were discarded from the phylogenetic analysis due to low bit scores and/or high e-values.

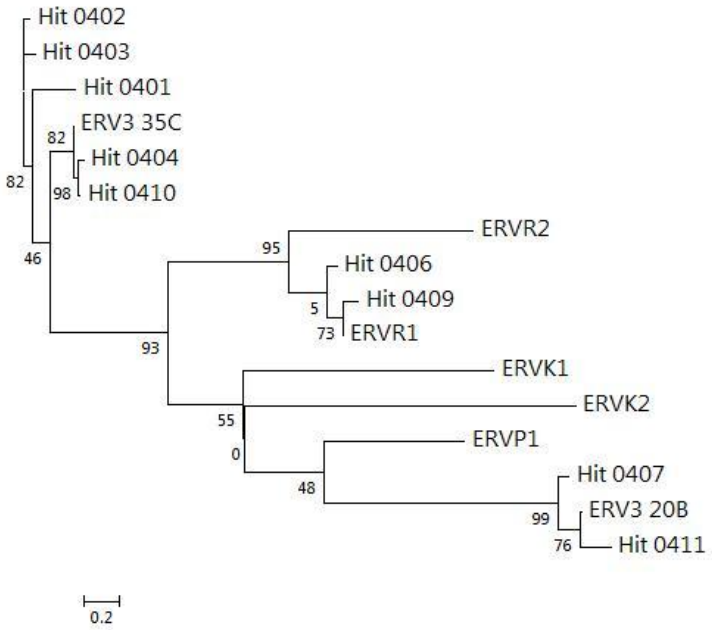

Figure 4 - Phylogenetic tree for the Chromosome 4 hits with the platypus-based pHMM (global alignment)

Chromosome 5 – Two out of three hits for this chromosome were validated in RepeatMasker, both belonging to the ERVR1 family. In our phylogeny, these hits correspond the Hit_0502 and Hit_0503. A good match was also found between LTR35 and Hit_0501, which was not present in the RepeatMasker annotation.

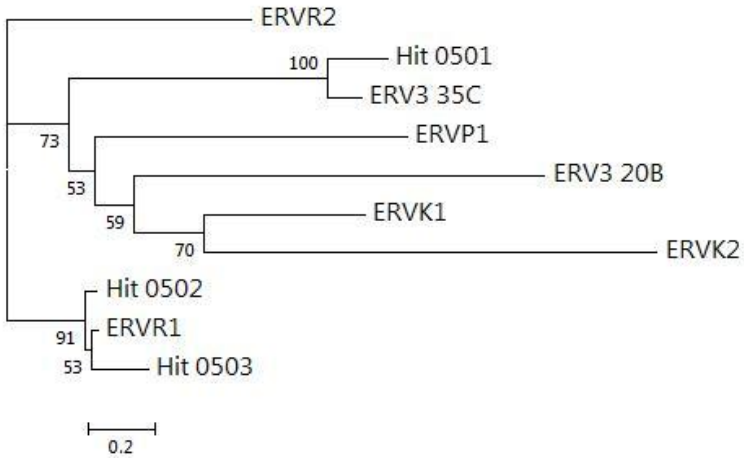

Figure 5 - Phylogenetic tree for the Chromosome 5 hits with the platypus-based pHMM (global alignment)

Chromosome 6 – RepeatMasker annotation confirmed two out of six of our hits in the last of the analyzed chromosomes. One hit belonged to the ERVR1 family while the other was placed in the ERVR2 family. In our phylogeny, these hits correspond to

Hit_0605 and Hit_0601, respectively. However, we have three matches that weren't present in RepeatMasker annotation, all of them clustering very well in the LTR20 vicinity. One of our hits was discarded due to poor bit score value.

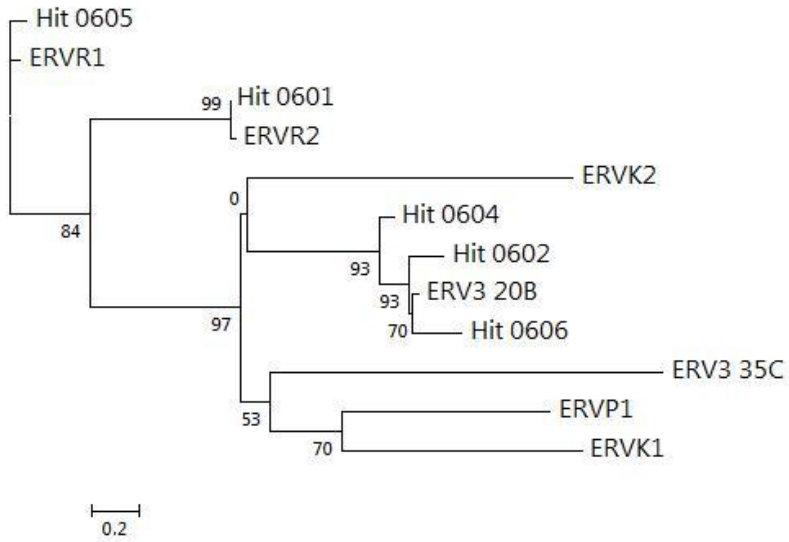

Figure 6 - Phylogenetic tree for the Chromosome 1 hits with the platypus-based pHMM (global alignment)

Summarizing, the pHMMs built on available platypus data managed to recover ERVR1 and ERVR2 data with good accuracy. However, the model failed to detect any LTR belonging to the ERVK or ERVP families, which may indicate that a separate model for these families is necessary to improve sensitivity. LTR35 and LTR20 were quite often detected only after a phylogenetic analysis since the RepeatMasker annotation seldom attributed our hits to these families. The latter results show that a phylogenetic checking of our pHMM method can be a useful tool to validate results whenever RepeatMasker annotation is lacking or not available at all.

B) Opossum pHMMs

In an attempt to detect novel endoretroviral LTRs, we ran several pHMMs built on opossum data, as described previously. RepeatMasker was unable to validate any of our results, with the exception of a single hit in the chromosome 3, using the model op13 in local alignment mode. This hit was catalogued as belonging to the LTR81 family, a family described as being very old and widespread among the mammalian species. Thus, phylogenetic relationships became our available tool for result validations. Since the scores for the local alignment mode were, in general, quite poor, we will only present, as before, discussion of the global alignment pHMMs, separated by model for convenience. Due to unresolved technical problems, we were unable to obtain a phylogeny for the op11 model results.

Model op01 – After examining the phylogenetic relationships between the hits obtained using this model, we have considered all of them as being false positives due to the extreme distance between them and our training dataset. Although many of the hits cluster together, this may only indicate an internal relationship due to the fact that most of them have long regions of CT and/or GA tandem repeats.

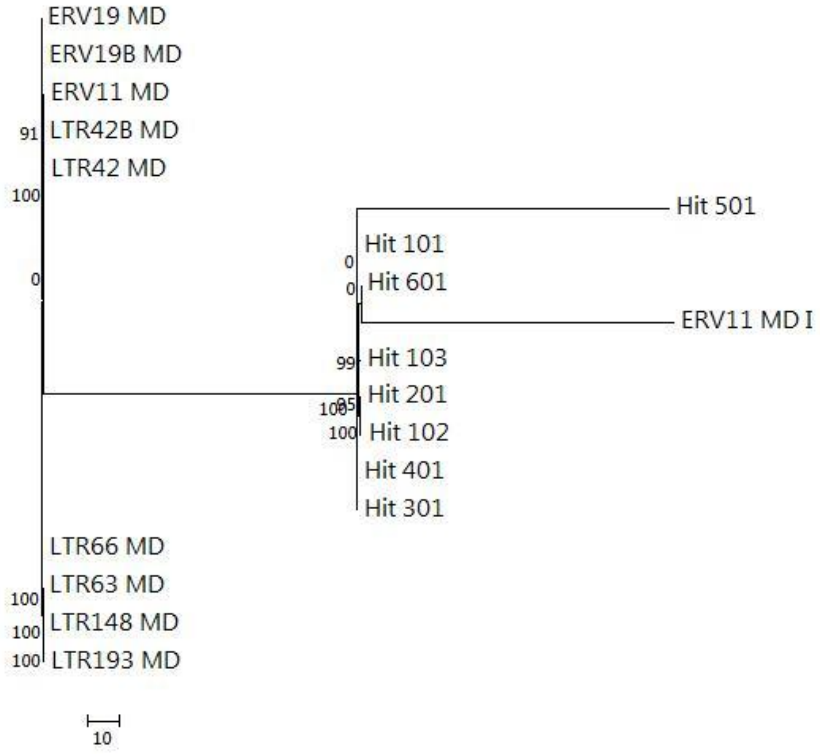

Figure 7 - Phylogenetic tree for the hits obtained with the opossum-based op01 model (global alignment)

Model op02 – This model presents very similar results to the previous one, where all the hits cluster together and apart from the training data, the shortest Kimura distance

between a sequence from the training data and a hit being 1.691 (between Hit_201 and LTR68). Again, all the hits present long CT and/or GA tandem repeats which may be reason behind of their detected phylogenetic relationship.

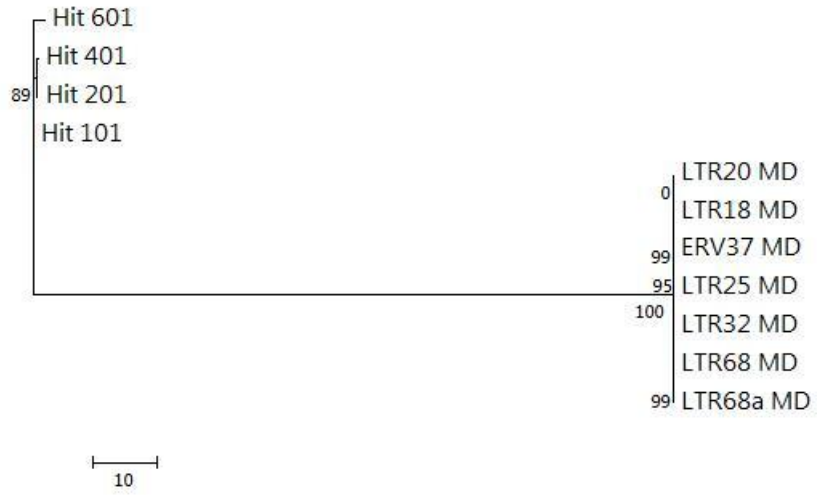

Figure 8 - Phylogenetic tree for the hits obtained with the opossum-based op02 model (global alignment)

Model op03 – This model is a particular case where we have two hits, one in chromosome 1 (Hit_101) and another in chromosome 4 (Hit_401) that appear to be phylogenetically closer to the training set than in previous examples, but the use of an unrooted tree for analysis prevents us from taking further conclusions. Both these hits are cases where we have strongly negative bit scores and low e-values: these may be significant hits that are phylogenetically related to our training set, although falling outside of it due to the use of a strict model, a fact that is reinforced by the observation that none of the hits have a Kimura distance to a sequence from the training set greater than 2.

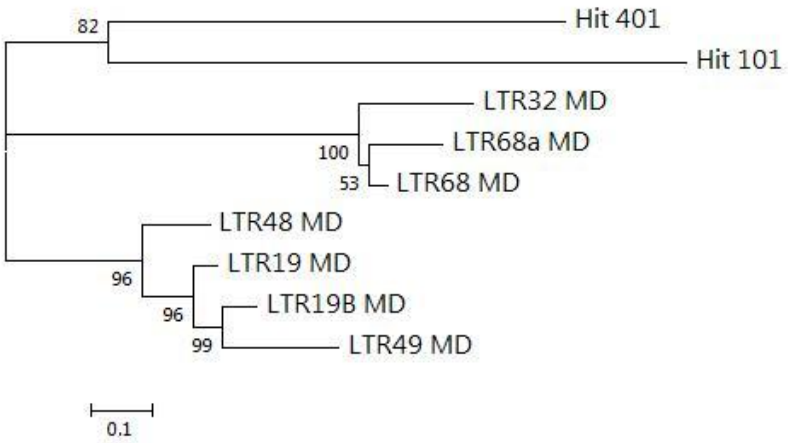

Figure 9 - Phylogenetic tree for the hits obtained with the opossum-based op03 model (global alignment)

Model op04 – This model also provided significant hits (e-value lower than 10^{-6}) with strongly negative bit scores. As the previous model, all these hits form a phylogenetic group that is distant from our training data set. Phylogenetic relationships within the hit space are apparent(Kimura distances < 1.5) but there is a clear separation between the hits and the training data set. For instance, the closer hit to the training data, Hit_501, possesses a large area covered by CT rich repeats. Interestingly, the first 400bp of this sequence are deprived of such repeats and strongly conserved along the mammal lineage.

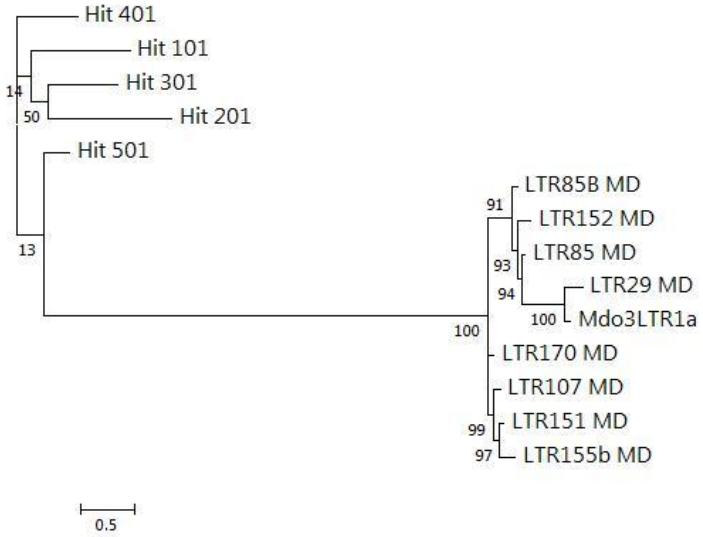

Figure 10 - Phylogenetic tree for the hits obtained with the opossum-based op04 model (global alignment)

Model op05 – This model provided quite a number of strong hits, with high bit scores and very low e-values. Although none of them clustered within the training data set, there seems to be strong phylogenetic relationships within the space of hits. The fact that RepeatMasker did not present any of these hits in the LTR annotation lead us to closely examine them in the UCSC Genome Browser. There, we observed that the majority of the hits are catalogued as intervals comprising low complexity repeats of CT and GA, as well as some elements of LINE and SINE repeats. The phylogenetic tree below translates the relationships found with this model. Analyzing the distance matrix, we can conclude that while there are some strong phylogenetic relationships within the hit space (some pairs possess Kimura distances below 1.5), the distance from the hits to the closer sequence from the training set exceeds 3, and we can reason that the entire hit space may be a group of closely related false positives rich in tandem repeats.

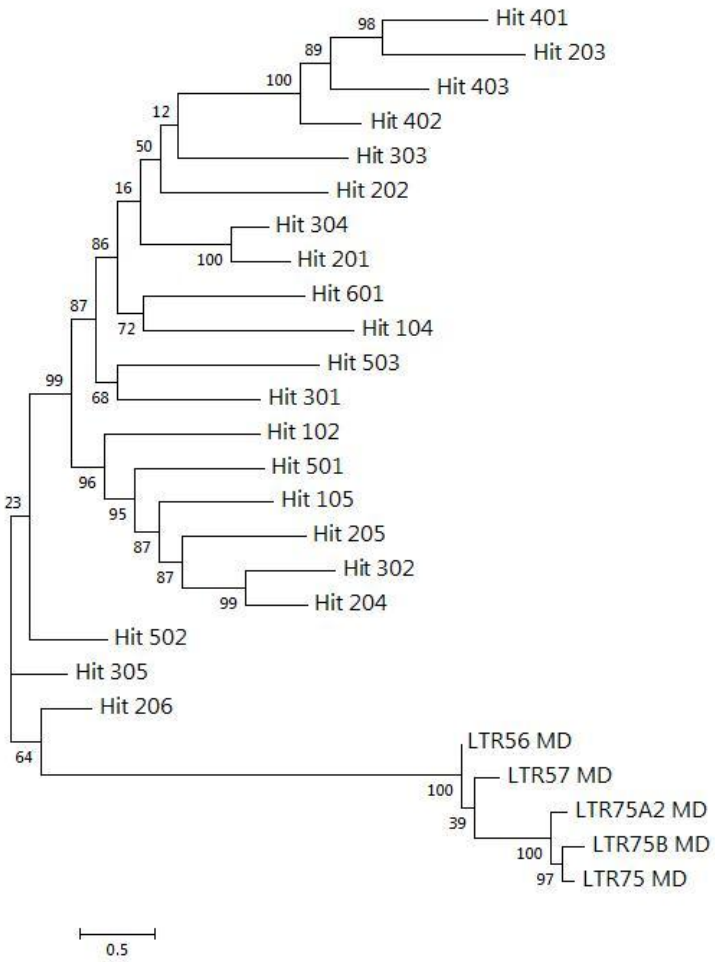

Figure 11 - Phylogenetic tree for the hits obtained with the opossum-based op05 model (global alignment)

Model op06 – This model provided only one significant hit below the e-value cutoff in all studied chromosomes and is, again, a hit with a strongly negative bit score. The phylogenetic relationship between the hit and the training data set is not apparent, however, and the Kimura distances from this sequence to the training data are too high (all greater than 2) to consider this hit as a true positive. Closer examination of the sequence reveals it to be heavily populated by short GA repeats.

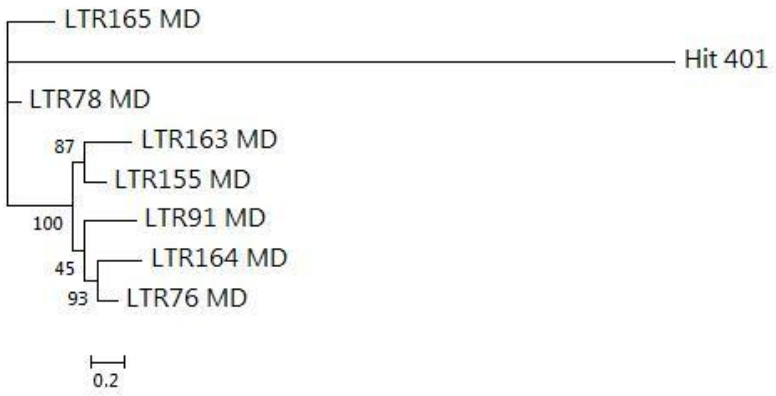

Figure 12 - Phylogenetic tree for the hits obtained with the opossum-based op06 model (global alignment)

Model op08 – Two hits from this model, Hit_101 and Hit_601, from chromosomes 1 and 6, respectively, appear to be phylogenetically related between themselves, with a Kimura distance of 1.174. The closest hit to the training data set is Hit_201, with all Kimura distances to the training sequences under 1.2, despite the low resolution of its internal node. A closer inspection of this sequence in UCSC Genome Browser shows that it has a small AT rich but otherwise it is deprived of annotation.

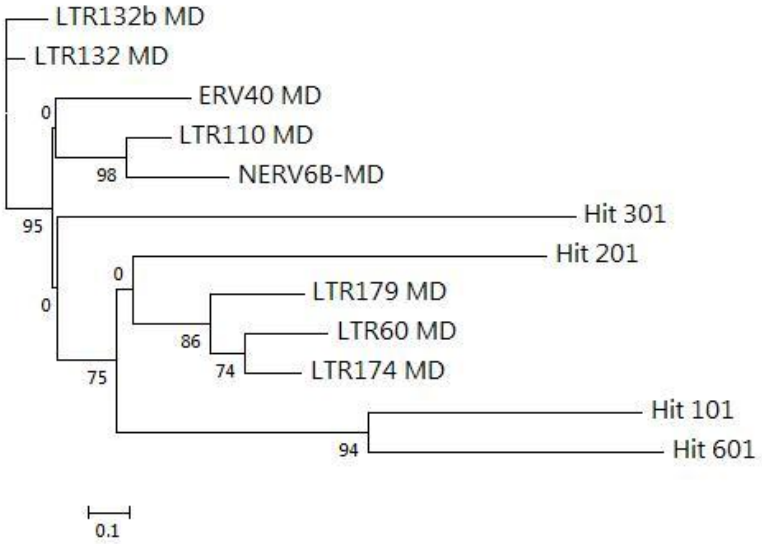

Figure 13 - Phylogenetic tree for the hits obtained with the opossum-based op08 model (global alignment)

Model op09 – This model only provided one hit in chromosome 6. Although it resembles a false positive, analysis of the distance matrix indicates that the Kimura distance between Hit_601 and LTR62 is 1.249, only slightly higher than the maximum distance between two sequences of the training set, 0.945 between LTR62 and NERV2A1. This does not exclude the possibility of Hit_601 being related to the

training sequences. Interestingly, this region possesses regions of high conservation between platypus, human and mouse.

Figure 14 - Phylogenetic tree for the hits obtained with the opossum-based op09 model (global alignment)

Model op10 – The results from this model cluster together with good bootstrap values for the internal nodes, suggesting a phylogenetic relationship between them. Low e-values and strong negative values for the bit score for all these hits may suggest a distant phylogenetic relationship to our training data that is, however, not that apparent in the phylogeny. The Kimura distance matrix confirms a close phylogenetic relationship between Hit_201 and Hit_401 and the LTR51 group as being the most similar to our hits, with an approximate genetic distance to Hit_201 of 2.396.

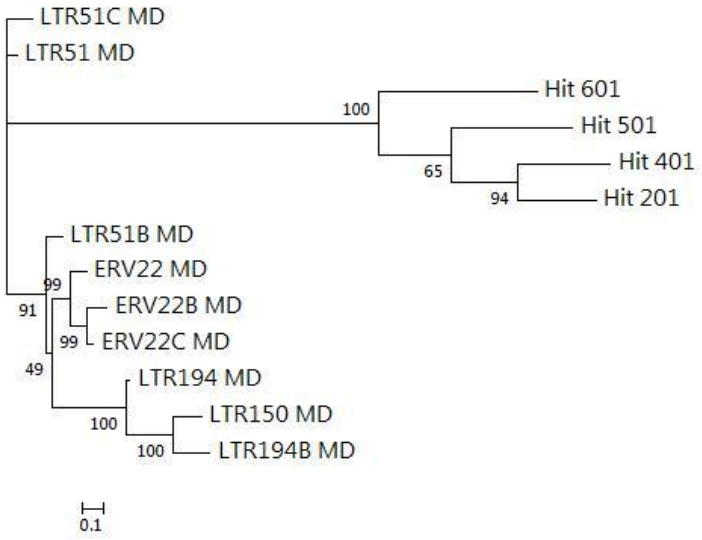

Figure 15 - Phylogenetic tree for the hits obtained with the opossum-based op10 model (global alignment)

Model op12 – This model provided one hit in each chromosome, four of which cluster together with very high bootstrap values, indicating a possible phylogenetic relationship. However, we believe that the hits for this model are false positives, due to the poor resolution of the tree. Analysis of genetic distances does not improve the scenario, only providing support for the relationship between the top group of hits formed by Hit_601, Hit_501, Hit_201 and Hit_301.

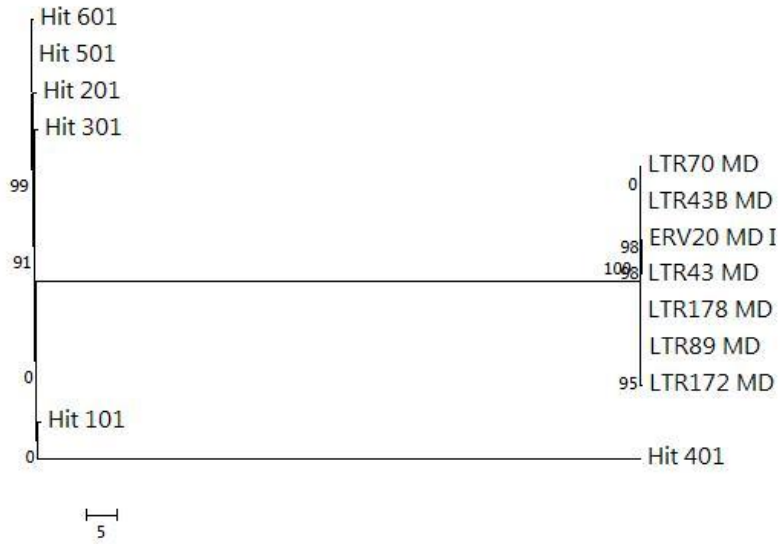

Figure 16 - Phylogenetic tree for the hits obtained with the opossum-based op12 model (global alignment)

Model op13 – This model provided two hits, one for chromosome 5 and another for chromosome 6, that cluster strongly together, but fail to provide any phylogenetic information relating them to the training data. Again, these hits fall in the strongly negative, low e-value category. Kimura distances indicate that LTR83A2 is the closest training sequence to the hits, having distances of 1.763 and 1.906 to Hit_501 and Hit_601, respectively.

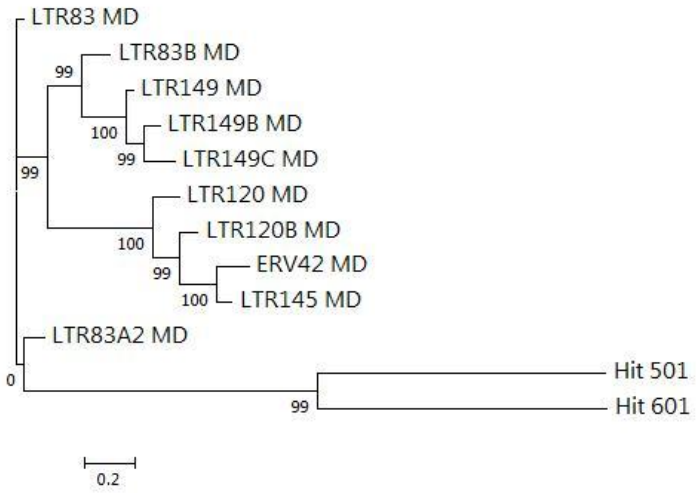

Figure 17 - Phylogenetic tree for the hits obtained with the opossum-based op13 model (global alignment)

Model op15 – This model was quite proficient in providing significant results. We had around a total of 50 results for all 6 chromosomes, with good phylogenetic resolution between most of them. The relationship to the training data is unclear though, as the bootstrap values for the connecting nodes are of 96 between the major hit group and the training data – minor hit group, and of 53 between the minor hit group and the training data. Looking in detail at these sequences in the genome browser, we found out once again that most of them are catalogued as being a mix of SINEs, LINEs and CT/GA rich low complexity repeats, within a genomic interval of 640 bp. Analysis of the Kimura distances between hits and sequences from the dataset reveals that it is impossible to find a genetic distance between a hit and a sequence from the training set below 3.0, hindering the possibilities of a clear phylogenetic relationship.

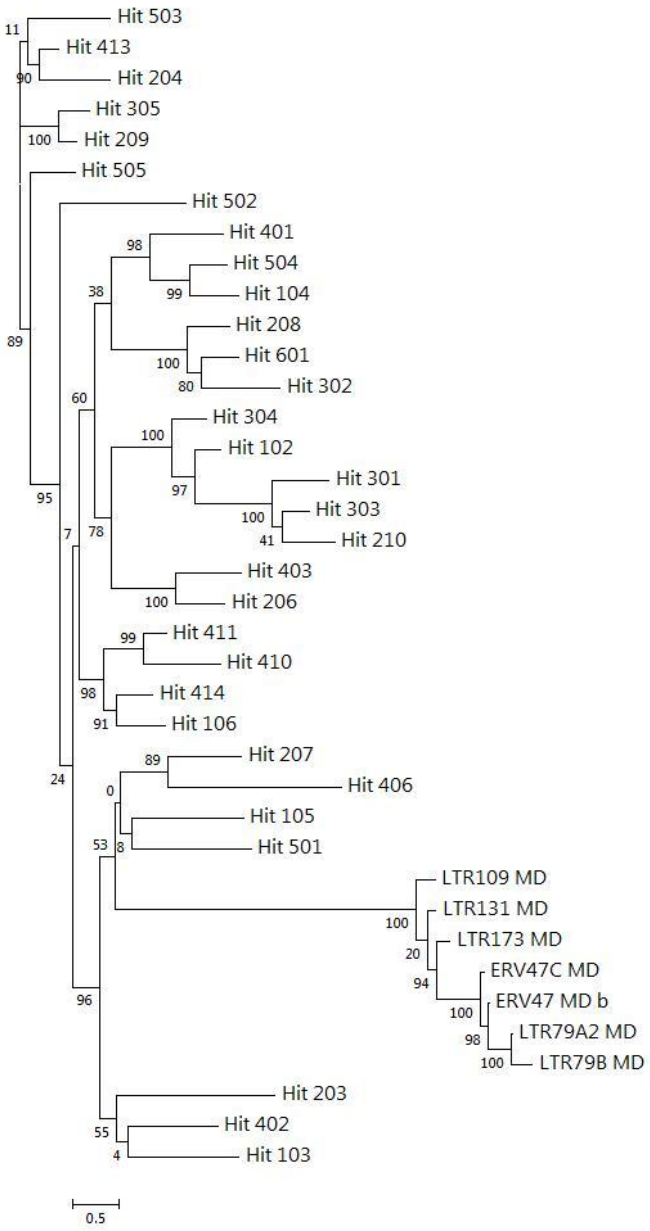

Figure 18 - Phylogenetic tree for the hits obtained with the opossum-based op15 model (global alignment)

Model op16 – This model presents probably the clearest phylogenetic relationship between hits and training data so far. Hit_601 falls close to the LTR31 group, and the remaining hits are not too distant from Hit_601. UCSC Genome Browser details this hit as having two SINE repeats followed by a GA tandem repeat within the 846 bp length of the sequence. Hit_601's Kimura distances to the ERV9 group are all below 1.5, indicating a possible phylogenetic relationship to that group. Hit_401, however, is further distanced to both Hit_601 and the ERV9 group, with distances averaging 2.1.

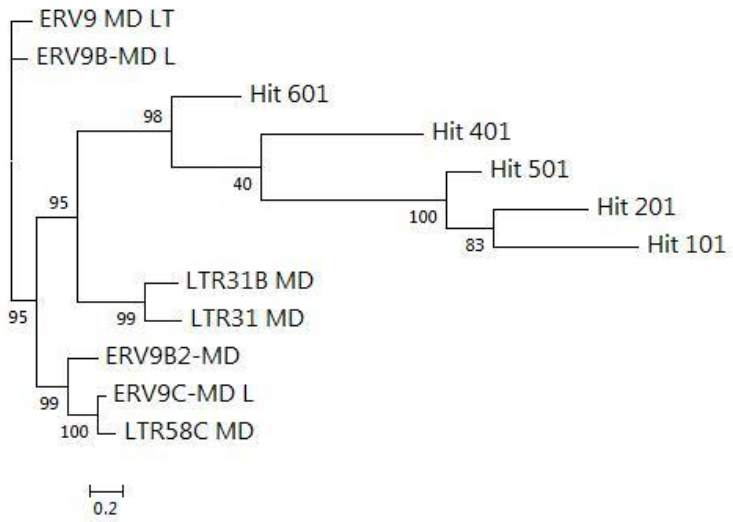

Figure 19 - Phylogenetic tree for the hits obtained with the opossum-based op16 model (global alignment)

Model op17 – This model also provides two hits that not only fall within the training data, but appear to be closely related to the LTR99 and LTR96 groups. Hit_301 and Hit_601 also cluster together with an internal bootstrap value of 100, indicating a strong phylogenetic relationship between them. The Hit_601 is shown in UCSC Genome browser as comprising two GA rich regions in the 3' half of the 793 bp-long sequence, while Hit_301 presents a similar structure, adding two small repeats labeled MonoRep401 in the 5' half. Kimura distances between Hit_301 and Hit_601 is 0.998, while the shortest distance to a sequence from the traning data, LTR6C is 1.756, which does not set aside a phylogenetic relationship between the two hits and sequences from the training data.

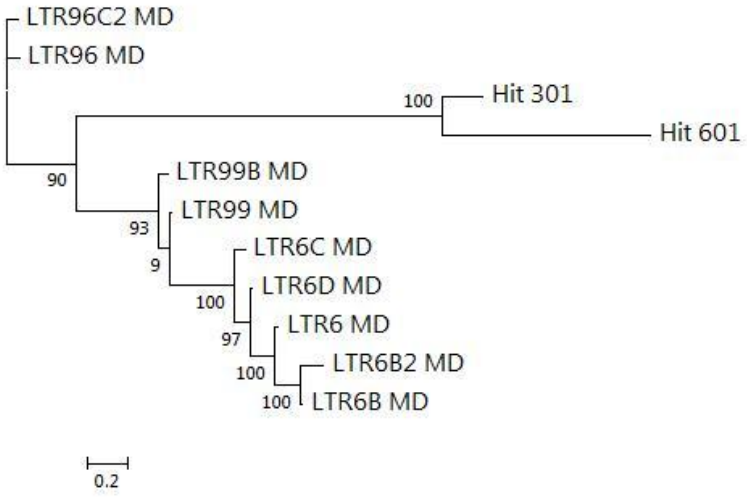

Figure 20 - Phylogenetic tree for the hits obtained with the opossum-based op17 model (global alignment)

Summarizing, we get very few reliable hits with the opossum models. Some we can readily dismiss as being false positives by looking at the phylogenetic relationships between the hits and the training data, while roughly two thirds of the models present significant hits with strong negative bit scores, an artifact that can be derived from the tightly built pHMMs. Recalling, these pHMMs have on average 7 to 8 sequences on their training data, and all of them are closely related (Kimura distance < 0.8), which may lead to somewhat inflexible models.

Conclusions

Overall, there was no clear benefit of using a local alignment strategy as opposed to a global alignment one. The majority of local alignment hits are either small tandem repeat sequences of less than 100bp or parts of hits that were recovered by the global alignment. The platypus-based pHMM was able to successfully recover a good number of true hits, mostly in the ERVR family. While, it failed to recover any hit from the ERVK or ERVP family, the LTR 20 and LTR 35 hits proved to be a valid addition to the current RepeatMasker annotation of the studied chromosomes. A separate pHMM built for those families that could not be retrieved, albeit with the risk of being too specific, could be worth a try.

The comparative pHMMs based on the opossum data, however, proved to be much less efficient, at least when judging by the genetic distances to the sequences in the training set. A change in strategy when building these pHMMs may be in order, allowing for more flexibility by adding more sequences and relaxing their relatedness. Recalling, we used small closely related groups of no more than 10 sequences with a Kimura distance smaller than 0.8.

In order to assess the specificity of our models, we ran the pHMMs against our own training sets and observed the values. The platypus-based pHMM has good scores for detecting all the training set sequences with the exception of ERVP1. Detailed results can be seen in Table 7. From here we can conclude that the model should work well for all sequences with the exception of ERVP1, due to a high degree of dissimilarity between this sequence and the rest of the training set. The bit scores aren't exceptionally high, which should allow for a reasonable degree of flexibility.

Sequence	Score	E-value
ERVR1	176.0	$7.2x10^{-53}$
ERVR2	158.8	$1.1x10^{-47}$
LTR 20B	167.2	$3.3x10^{-50}$
LTR 35C	147.4	$2.9x10^{-44}$
ERVK1	79.8	$6.8x10^{-24}$
ERVK2	116.8	$4.9x10^{-35}$
ERVP1	3.1	$1.3x10^{-4}$

Table 7 – Bit scores and e-values for the training set sequence. Higher scores and lower e-values improve the sensitivity of the model for a specific family.

The same analysis was repeated for all opossum-based models. Careful examination of the bit-score and e-value results reveals that, all models except op02, op03 possess very high bit scores for the majority of the sequences and very low ($< 10^{-80}$) e-values. Since the value of the average bit score is directly proportional to the strictness of the model, this reflects our previous assumption that the opossum-based models are indeed not

very flexible and the building strategy should be changed to allow for a more broad range of training sequences. The full testing tables can be seen in Appendix D.

Most of the unknown hits appear to be heavily rich in short tandem repeats. This was a recurring problem when using pHMMs, and while the training set sequences are not especially enriched in these elements, they proved to be a major factor to take into account when analyzing the results and picking out false positives. The small number of hits in the global alignment strategy allowed us to use manual characterization of most of the hits – a process that, however tedious, proved to be quite interesting and sometimes essential to discriminate results. While doing so, we discovered that the LTR 20 elements of platypus occur in pairs almost exclusively, with a small 10-30bp space of non-LTR sequence between both copies.

Unlike what is done for protein sequences, HMMER does not use mixture Dirichlet priors for nucleotides. Instead, it uses plus-one (Laplace) priors for match and insert emission priors, which provide only rudimentary prior knowledge in pseudocount and for this reason, we need a lot of data in the alignment to get good estimate of the parameters[20]. While in fact we only have a small training set --less than 10 sequences for each group, and it might be a cause for poor detection levels.

Changing the entry probabilities and exit probabilities manually (which is set by default in HMM, independent of the training data), the number of hits we get varies a lot. While in fact it is difficult to set these parameters empirically, it might not be a bad idea to try HMM in a sliding window on the query sequence. On the other hand, most of our false positive hits are successive GA, CT repeats, and we can easily eliminate them using the sliding window.

We also planned on building a third set of pHMMs based on very old, mammalian-wide conserved LTR families. The fact that the only recognized hit on the local alignment opossum-based pHMMs belongs to one of these families (LTR81), indicates that they may be quite useful in detecting LTR elements in the platypus genome.

References

[1] O'Brien SJ, **The platypus genome unraveled**, *Cell 2008 Jun 13;***133(6)**:*953-5*

[2] Gifford R, Tristem M, **The evolution, distribution and diversity of endogenous retroviruses**, *Virus Genes***26(3)**:291-315

[3] Coffin JM, **Structure and classification of retroviruses**, *in* Levy JA, **The Retroviridae (1st ed.)**, New York Plenum Press, pp. 26-34

[4] Mougel M, Houzet L, Darlix JL, **When is it time for reverse transcription to start and go?**, *Retrovirology 2009 Mar 4;* **6:**24

[5] Schön U, Diem O, Leitner L, Günzburg WH, Mager DL, Salmons B, Leib-Mösch C, **Human endogenous retroviral (HERV) LTR sequences as cell type-specific promoters in retroviral vectors**, *J Virol 2009 Sep 9 (Epub)*

[6] Kalyanaraman A, Aluru S, **Efficient algorithms and software for detection of full-length LTR retrotransposons**, *J Bioinform Comput Biol. 2006 Apr;* **4(2)**:197-216

[7] Smit AFA, Hubley R, Green P, **RepeatMasker Open-3.0**, 1996-2004 <http://www.repeatmasker.org>

[8] McCarthy EM, McDonald JF, **LTR_STRUC: a novel search and identification program for LTR retrotransposons,** *Bioinformatics 2003 Feb 12;***19(3)**:362-7

[9] Sperber GO, Airola T, Jern P, Blomberg J, **Automated recognition of retroviral sequences in genomic data – RetroTector**, *Nucleic Acids Res 2007;***35(15)**:4964-76

[10] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG, **ClustalW and ClustalX version 2**, *Bioinformatics 2007;***23(21):**2947-48

[11] Rabiner LR, **A tutorial on Hidden Markov Models and selected applications in speech recognition**, *Proceedings of the IEEE February 1989;***77(2)**:257-86

[12] Eddy, SR, **Profile Hidden Markov Models**, *Bioinf Rev. 1998;***14(9)**:755-63

[13] HMMer software package, <http://hmmer.janelia.org>

[14] Karolchik D, Kuhn, RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, Miller W, Pedersen JS, Pohl A, Raney BJ, Rhead B, Rosenbloom KR, Smith KE, Stanke M, Thakkapallayil A, Trumbower H, Wang T, Zweig AS, Haussler D, Kent WJ, **The UCSC Genome Browser Database: 2008 update,** *Nucleic Acids Res. Jan 2008;***36**:D773-9

[15] Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J, **Repbase Update, a database of eukaryotic repetitive elements**, *Cyt Gen Res 2005:***110**:462-67

[16] Kimura M, **A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences,** *J Mol Evol 1980;***16:**111-120

[17] R statistical software package, <http://cran.r-project.org>

[18] Felsenstein J, **PHYLIP(Phylogeny Inference Package)**, *Auth Distributed 1993*, Dept of Genetics, Univ of Washington, Seattle

[19] Guindon S, Gascuel O, **A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood**, *Systematic Biol 2003;***52(5):**696-704

[20] Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D, **Dirichlet mixtures: a method for improved detected of weak but significant protein sequence homology**, *Comput Appl Biosci. Aug 1996;***12(4):**327-45