Endogenous retrovirus

The nature of endogenous retrovirus

In the human genome, various (pro-)retroviral elements are present in large numbers. In fact, about 5-8 % of the genome consists of such elements [1]. These are so-called endogenous retroviruses which are proviral versions of exogenous retroviruses that have been integrated in the host genome via infection of germ cells. [2] [3] Retroviruses are more or less structurally alike in that they consists of three or four genes – called *env*, *pro*, *pol*, and *gag* – flanked by two identical non-coding regions called long terminal repeats (LTRs). At insertion time, these repeats are identical in both ends of the virus, as they arise due to the replication mechanism of retroviruses where the reverse transcriptase utilizes one the same template for both ends. As the 5' LTR and 3' LTR are identical at insertion time, they can thus be used as references to each other when it comes to integration time estimation. [4]

The abundance of the endogenous retroviruses in mammalian genomes leads to speculation of whether they have a function or not. Several investigations of different viruses have been carried out and some of them yield a positive result with regard to having a physiological function. Time integration can be important in the total assessment of the evolutionary role of the virus: Say a virus that is estimated to being integrated at a certain point in time but is not present in all species speciated since then and is hence lost from the population – such a virus would be less likely to have a vital function. Evidence shows that the LTRs can act as promoters for unrelated genes in their proximity, which means that they are an evolutionary force. [3]

Although some of the viruses play a role in our living, many are merely remnants with both stop codons created by substitution events and large deletions. [2] The large proportion of ERVs in our genome is both due to various infections by different viruses but also the amplification of the integrated viruses either by effect of itself or by help from proteins encoded by similar viruses.

When estimating the integration time, one must take into account that the 5' and 3' LTR evolves with different rates. That is due to the fact that the 5' end has more of a function: The 5' end is involved in several events, hereunder transcription regulation, initiation and termination, [3] whereby the genes are expressed. Thus, the 5' LTR is object to fewer substitutions than the 3' end. This time estimation method is though somewhat biased, as the large number of copies in each virus family makes it very likely that homologous recombination events should take place. [5]

Time estimation

Extract from genome

Many retroviruses have already been catalogued and their approximate positions are known. [7] For the time estimation we wanted to accomplish, we chose ERV3 and ERV-WE1 which are both very well-studied viruses. We search them in the genome

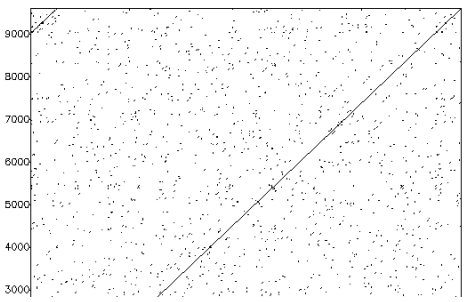

Figure 1: ERV3 plotted against itself (window size 8; threshold 35). The ends show high similarity indicated by the continuous lines in the left uppermost and right lower corners.

browser while adding an additional track from the former retrosearch database. [7] With the RepeatMasker (used to recognize repeated elements based on a database [8]) in the genome browser, we identified both the 5' and 3' LTR and the intra-LTR region and extracted the DNA sequences in FASTA format. Subsequently, we dotplotted[1] each virus against itself to confirm the LTRs' homogeneity.

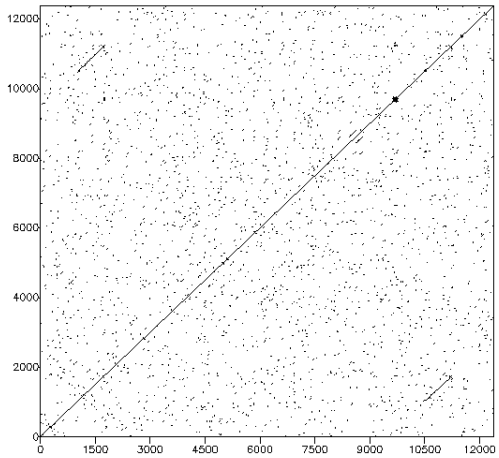

Figure 2: ERV-WE1 plotted against itself (window size 8; threshold 35).

BLAST to find homologs

With or sequence in hand, we needed to investigate whether the virus was present in other species. For this purpose, we blasted the dna sequence against the whole BLAST database to get as many species represented as possible.

BLAST

The basic principle for BLAST (Basic Local Alignment Search Tool) is to search sequences based on segments of the query sequence. These segments are of a constant size and the alignment with the search sequences must have a score above a certain threshold to be considered a match. When a match is found, the segment is extended in both directions to make the longest sequence that is well aligned. This approach makes BLAST somewhat error prone but also fast when one needs to search a lot of sequences. [6]

Orthologs and paralogs – making the distinction

Homologous sequences are sequences that are very similar and thus are likely to have functional similarities. A pair of homologous sequences is for instance a certain retrovirus in one species and its counterpart in another species. When looking for homologs across species one should though be aware of the fact that they can be derived from several events: they can have emerged from the exact same sequence when the speciation event; but also they can have arisen as a duplication event took place either before or after speciation in which case they are out-paralogs and in-paralogs respectively. [9] Also, with our knowledge of variety of retrovirus in the mammalian genomes, there could be homologous sequences that didn't even arise from the same virus ancestor. In the case of integration time estimation, only the orthologs are interesting, as those are the originally same virus and thus the one that has integrated.

[1] Dotplot makes use of a window size and a threshold and some score function. The window size is a measure of how many nucleotides that are compared at the time and the threshold is a score that should be reached before a line between two dots is drawn. The window size and threshold are chosen empirically.

To make sure that we had the orthologous sequences, we performed two tests. One being a dotplot of the human sequence against the monkey with each end of the sequences extended by 1000 bases to see whether the surrounding genetic material were also similar implying that the integration place is identical. The other test was a histogram of our BLAST hits to see whether other sequences could be assumed just as good a match as the chosen one. Using the multiplication of query cover (in percent) and similarity (in percent), we only had one hit of more than 90% and the rest were less than 70% after which we concluded that the chosen sequence in all likelihood was the correct one.

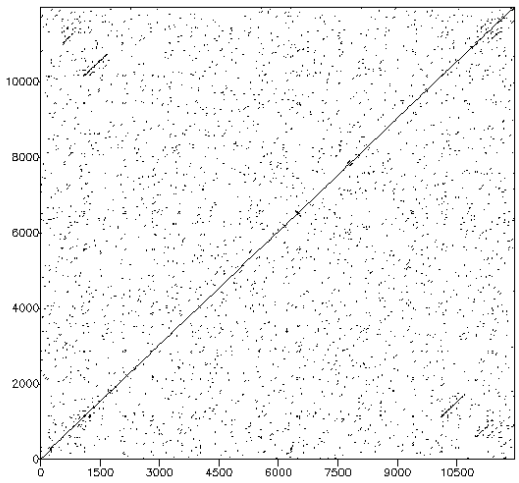

Figure 3: Human ERV3 plotted against chimp ERV3 with 1000 extra bases in each end to confirm same insertion place in each species and thereby orthology.

Species selection

When choosing the species to work with, we had in mind that the age of split and the pre-estimated integration time shouldn't be far apart, *i.e.* integration time 100 Mya and speciation 6 Mya. As we chose two viruses that had been integrated somewhere after the new world and old world monkeys split, we could choose freely among the old world monkeys and primates, emphasizing a broad specter of these. Thus for ERV3, we chose *macaca mulatta* and *pan troglodytes* and for ERV-WE1, we chose additionally three species, namely *gorilla gorilla, pongo pygmaeus* and *hylobates pileatus* and not macaca.

LTR and intra-LTR

We need to split the LTR-regions from the rest of the virus to do time integration estimation. Before we can crop out the LTR-regions for all the species selected, we need to align them. ClustalW is used as alignment tool in MEGA. This alignment method consists of three main stages: pairwise alignment, guide tree calculation and progressive alignment. In the first step all pairs of sequences are aligned separately in order to calculate their pairwise distances and obtain a distance matrix. The second step is to calculate a guide tree from the distance matrix using the Neighbor-Joining method. In the final multiple alignment process sequences are progressively aligned according to the branching order in the guide tree. [10]

The LTR-regions align very well and because we know the LTR-boundaries for the human virus from Genome Browser, we can crop out the LTR-regions for all the species.

Substitution models

For phylogenetic analysis of our data, we need to determine how our data evolved i.e. which substitution model that describes our data. A substitution model specifies the way characters are permitted to evolve between states as well as the relative rate of different kinds of evolutionary change. All models are continuous-time Markov models which mean they describe a process in which the probability of an event happening in some time window is dependent only on the state at that time and independent of how it came to be in that state.

We used a ”Find Best-Fit Substitution Model (ML)”- test in MEGA to see which substitution model fits our data the best. ML in phylogenetic involves searching for the tree that has the highest probability of giving rise to the observed data. An evolutionary tree is needed for evaluating the fit of substitution models to the data, and MEGA5 automatically infers the tree by the Neighbor-Joining algorithm that uses a matrix of pairwise distances estimated under the Tamura-Nei model for nucleotide sequences. [11] Branch lengths and substitution rate parameters are then optimized for each model to fit the data.

The test result is shown in the figure below:

Model	#Param	BIC	AICc	lnL	Invariant	Gamma	R	Freq A	Freq T	Freq C	Freq G
T92	11	2694,8	2627,1	-1302,5	n/a	n/a	3,7394	0,28	0,28	0,22	0,22
T92+G	12	2694,8	2621	-1298,4	n/a	0,43464	3,9997	0,28	0,28	0,22	0,22
T92+I	12	2695	2621,2	-1298,6	0,586177	n/a	3,981	0,28	0,28	0,22	0,22
K2+G	11	2696,5	2628,8	-1303,4	n/a	0,36112	4,0898	0,25	0,25	0,25	0,25
K2	10	2697,9	2636,4	-1308,2	n/a	n/a	3,7469	0,25	0,25	0,25	0,25
T92+G+I	13	2702,9	2622,9	-1298,4	0,218594	0,73356	4,0041	0,28	0,28	0,22	0,22
K2+G+I	12	2704,6	2630,8	-1303,3	0,274256	0,70239	4,0977	0,25	0,25	0,25	0,25
K2+I	11	2704,9	2637,2	-1307,6	0,075689	n/a	3,7609	0,25	0,25	0,25	0,25
HKY	13	2707,6	2627,6	-1300,7	n/a	n/a	3,7353	0,263	0,297	0,2421	0,1978
HKY+G	14	2707,6	2621,5	-1296,7	n/a	0,43529	3,989	0,263	0,297	0,2421	0,1978
TN93	14	2715,7	2629,6	-1300,7	n/a	n/a	3,7351	0,263	0,297	0,2421	0,1978
HKY+I	14	2715,7	2629,6	-1300,7	0,00001	n/a	3,7353	0,263	0,297	0,2421	0,1978
HKY+G+I	15	2715,7	2623,5	-1296,7	0,213687	0,72302	3,9937	0,263	0,297	0,2421	0,1978
TN93+G	15	2715,7	2623,5	-1296,7	n/a	0,43427	3,9878	0,263	0,297	0,2421	0,1978
TN93+I	15	2721,4	2629,1	-1299,5	0,165923	n/a	3,7644	0,263	0,297	0,2421	0,1978
TN93+G+I	16	2723,9	2625,5	-1296,7	0,215513	0,72532	3,9923	0,263	0,297	0,2421	0,1978
GTR+G	18	2736	2625,3	-1294,6	n/a	0,44925	3,9655	0,263	0,297	0,2421	0,1978
GTR	17	2737,4	2632,9	-1299,4	n/a	n/a	3,1493	0,263	0,297	0,2421	0,1978
GTR+I	18	2740,9	2630,2	-1297	0,312043	n/a	3,2059	0,263	0,297	0,2421	0,1978
GTR+G+I	19	2744,2	2627,3	-1294,6	0,162744	0,65467	3,9681	0,263	0,297	0,2421	0,1978
JC	9	2769	2713,6	-1347,8	n/a	n/a	0,5	0,25	0,25	0,25	0,25
JC+G	10	2770,4	2708,9	-1344,4	n/a	0,53125	0,5	0,25	0,25	0,25	0,25
JC+I	10	2776,3	2714,7	-1347,3	0,064057	n/a	0,5	0,25	0,25	0,25	0,25
JC+G+I	11	2778,6	2710,9	-1344,4	0,109603	0,68345	0,5	0,25	0,25	0,25	0,25

Figure 4: Result for ”Find Best-Fit Substitution Model” for ERV3 LTR-regions

The goodness-of-fit of each model is measured by the Bayesian information criterion, corrected Akaike information criterion and the log likelihood. The preferred model is the one that scores the lowest in all three criteria.

The advantage of AIC and BIC are that they can be used to compare both nested and nonnested models. AIC is calculated as: [12]

$$\mathrm{AIC}=2\ln(\mathrm{L})+2\mathrm{k}$$

where *k* is the number of free parameters and L is the maximum likelihood value of the data. Since the preferred model is the one with the lowest score, AIC not only rewards goodness of fit, but also includes a penalty for the increasing number of parameters.

MEGA5 uses the corrected AIC (AICc [13]) which is AIC with a correction for finite sample sizes:

$$\mathrm{AICc}=\mathrm{AIC}+\frac{2\mathrm{k}(\mathrm{k}+1)}{\mathrm{n}-\mathrm{k}-1}$$

where *n* is the sample size (number of observations). AICc is therefore AIC with a greater penalty for extra parameters.

The definition of BIC is: [12]

$$\mathrm{BIC}=-2\ln(\mathrm{L})+\mathrm{k}^{*}\ln(\mathrm{n})$$

The BIC generally penalize free parameters more strongly than AIC. Because real data often has a natural log of n>2, BIC should tend to choose simpler models than AIC.

The last test we used to choose the best-fit model of evolution for our data set were the Likelihood Ratio Test: [12]

$$\mathrm{LR}=2|\ln(\mathrm{L}_1)-\ln(\mathrm{L}_0)|$$

where L_1 is the maximum likelihood under the more parameter-rich complex model, and L_0 is the maximum likelihood under the less parameter-rich simple model (the null hypothesis). When the models compared are nested, twice the log-likelihood difference between the two models are expected to fit a chi-square distribution with ρ degrees of freedom, where ρ is the difference in number of free parameters between the two models. The model with more parameters will always fit at least as well as the model with fewer parameters. Whether it is significant better is determined by deriving the p-value of ρ.

Several hypotheses about the data set can be tested in this manner. We can test if all base frequencies are equal. Is there a transition/transversion bias? Are all transition rates equal? Are there invariable sites? Is there rate homogeneity among sites? An example of testing the last hypothesis is shown here. We compare T92 and T92+G:

$$2*(-1298.4)-(-1302.5)=8.2$$

From the chi square distribution table with one degree of freedom we get a p-value<0.01, which mean we reject the simple model. Therefore we can conclude there is a gamma distribution among sites. Testing the other hypothesis too, shows us that the best substitution model to describe our data is T92+G+I. Tamura 92

extends Kimura's two-parameter method (which distinguish between transition and transversion) to the case where a G+C content bias exists. The rate matrix for T92 is as follows [10]:

	A	G	T	C
A	*	$\kappa(1-\pi_{GC})/2$	$(1-\pi_{GC})/2$	$(1-\pi_{GC})/2$
G	$\kappa\pi_{GC}/2$	*	$\pi_{GC}/2$	$\pi_{GC}/2$
T	$(1-\pi_{GC})/2$	$(1-\pi_{GC})/2$	*	$\kappa(1-\pi_{GC})/2$
C	$\pi_{GC}/2$	$\pi_{GC}/2$	$\kappa\pi_{GC}/2$	*

Table 1: Rate matrix for T92 model

There is only one base frequency: π_{GC}. π_A and π_T are 1- $\pi_{GC}/2$ respectively. The model includes a transition:tranversion bias κ. The higher the value of κ, the higher the rate of transitions relative to transversions.

T92 has two parameters: the base frequency π_{GC} and the transitions:transversion bias κ. The rest of the parameters in the substitution model-test are due to the individual branches, which can be calculated as 2*n*-3, where *n* is the number of sequences. In this example we get 2*6-3 = 9 branches plus 2 = 11 parameters. Gamma distribution and invariable sites each contribute one parameter. Thus T92+G+I has 13 parameters.

We did the same for another virus; ERVW-1. The "Find Best-Fit Substitution Model"-test showed this result:

Model	#Param	BIC	AICc	lnL	Invariant	Gamma	R	Freq A	Freq T	Freq C	Freq G
K2+G	19	3631,4	3501,8	-1731,9	n/a	0,12845	5,8302	0,25	0,25	0,25	0,25
K2+I	19	3638	3508,5	-1735,2	0,673016	n/a	5,6122	0,25	0,25	0,25	0,25
K2+G+I	20	3639,6	3503,2	-1731,5	0,402173	0,35999	5,8995	0,25	0,25	0,25	0,25
T92+G	20	3639,9	3503,5	-1731,7	n/a	0,12411	5,8918	0,2422	0,2422	0,2578	0,2578
T92+I	20	3640,8	3504,4	-1732,1	0,762479	n/a	5,8781	0,2422	0,2422	0,2578	0,2578
HKY+G	22	3647,9	3497,9	-1726,9	n/a	0,12735	5,8619	0,2473	0,2372	0,2944	0,2211
T92+G+I	21	3648	3504,8	-1731,3	0,402757	0,35066	5,9649	0,2422	0,2422	0,2578	0,2578
TN93+G	23	3655,8	3499	-1726,4	n/a	0,12447	5,8637	0,2473	0,2372	0,2944	0,2211
HKY+G+I	23	3656,2	3499,4	-1726,6	0,350009	0,30661	5,9211	0,2473	0,2372	0,2944	0,2211
TN93+I	23	3656,4	3499,6	-1726,7	0,760378	n/a	5,8509	0,2473	0,2372	0,2944	0,2211
TN93+G+I	24	3664,3	3500,7	-1726,3	0,260942	0,23849	5,8992	0,2473	0,2372	0,2944	0,2211
GTR+G	26	3680,6	3503,3	-1725,6	n/a	0,12386	5,8879	0,2473	0,2372	0,2944	0,2211
GTR+G+I	27	3689,1	3505,1	-1725,4	0,257183	0,23509	5,9182	0,2473	0,2372	0,2944	0,2211
HKY+I	22	3693	3543	-1749,4	0,359581	n/a	5,3264	0,2473	0,2372	0,2944	0,2211
K2	18	3700,3	3577,5	-1770,7	n/a	n/a	5,2076	0,25	0,25	0,25	0,25
T92	19	3710,5	3580,9	-1771,4	n/a	n/a	5,2128	0,2422	0,2422	0,2578	0,2578
HKY	21	3717,6	3574,4	-1766,1	n/a	n/a	5,2158	0,2473	0,2372	0,2944	0,2211
TN93	22	3725,5	3575,5	-1765,7	n/a	n/a	5,2171	0,2473	0,2372	0,2944	0,2211
GTR+I	26	3734,4	3557,2	-1752,5	0,277119	n/a	5,3087	0,2473	0,2372	0,2944	0,2211
GTR	25	3750,3	3579,8	-1764,8	n/a	n/a	5,2284	0,2473	0,2372	0,2944	0,2211
JC+G	18	3787,7	3664,9	-1814,4	n/a	0,14869	0,5	0,25	0,25	0,25	0,25
JC+G+I	19	3796,2	3666,6	-1814,3	0,280905	0,29309	0,5	0,25	0,25	0,25	0,25
JC	17	3849,2	3733,2	-1849,6	n/a	n/a	0,5	0,25	0,25	0,25	0,25
JC+I	18	3858	3735,2	-1849,6	0,00001	n/a	0,5	0,25	0,25	0,25	0,25

Figur 5: Result for "Find Best-Fit Substitution Model" for ERV-WE1 LTR-regions

Likelihood Ratio tests comparing different hypothesis reveals that HKY+G is the best substitution model for this data set. The rate matrix for HKY is shown below: [15]

	A	G	T	C
A	*	$\kappa\pi_G$	π_T	π_C
G	$\kappa\pi_A$	*	π_T	π_C
T	π_A	π_G	*	$\kappa\pi_C$
C	π_A	π_G	$\kappa\pi_T$	*

Tabel 2: HKY rate matrix

HKY allows unequal base frequencies ($\pi A \neq \pi G \neq \pi T \neq \pi C$) and it distinguishes between the rate of transition and transversions (κ).

Phylogeny

For phylogenetic analysis we use the T92 model to build a phylogeny for the ERV3 LTR's under the Neighbor-Joining method using default parameter values and 100 bootstrap replicates.

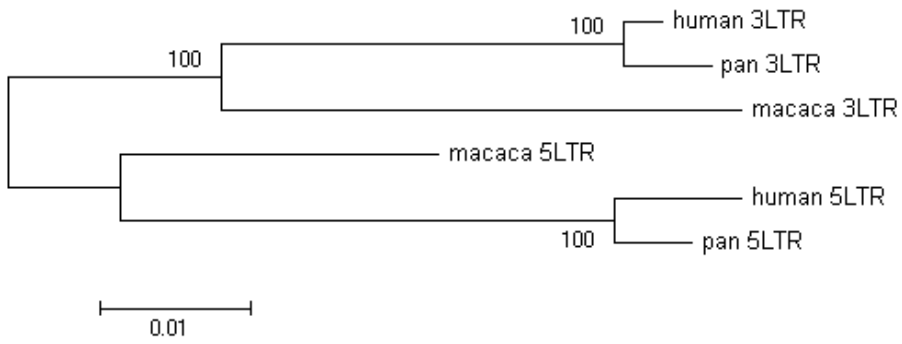

Figure 6: Phylogeny for ERV3 LTRs

The phylogeny correctly separates the 3' and 5'LTRs into two clusters and places pan and human for each LTR as closest related.

We did the same for the ERV-WE1 virus under the HKY model:

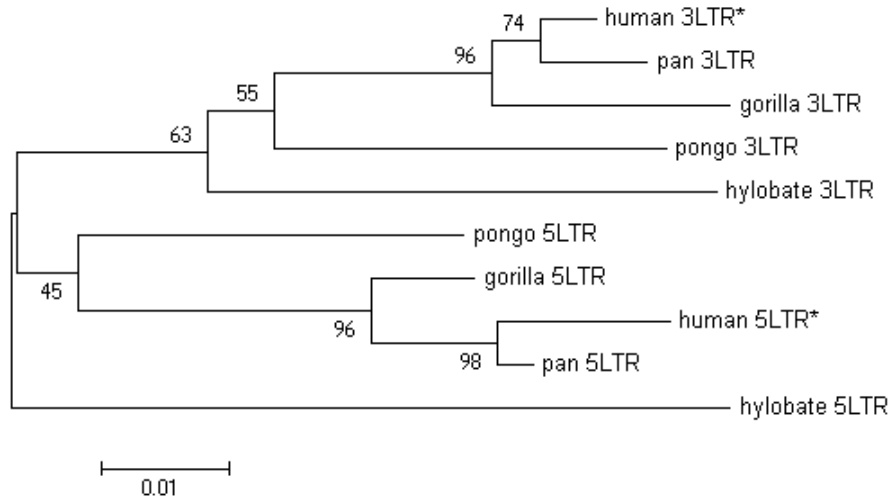

Figure 7: Phylogeny for ERV-WE1

Integration

The goal of this project was to estimate the integration time of one or more retroviruses. This can be looked upon from two different angles. [2] First, an approximate insertion time can be estimated by deducing it from the variety of species that the virus in question is present in. For instance, as we've found the ERV3 in both human and macaque, it is well justified to say that it was inserted before the speciation of these species, *i.e.* before the split of the old world monkeys and apes. As the same virus isn't seen in new world monkeys, we can assume a limit back in time as well, being the new world/old world monkey split.

This can however be a faulty assumption as the virus can potentially have been inserted before the split but afterwards being lost in new world monkeys. This estimate gives us somewhere between the emergence of Simiiformes and the emergence of Catarrhini (36–50 Mya and 20–38 Mya, respectively, according to fossil records) [10] – potentially 20-50 Mya.

As this is a very rough estimate, another method was used. Recall from the introduction that the 3' and 5' LTR evolves with different rate. With this knowledge and likewise knowledge of speciation times, it is possible to calculate the integration time of the retrovirus.

To do this, one must assume a local molecular clock for each side of the phylogeny. As we can calculate the nucleotide substitution rate for the branches until the most recent common ancestor of human and macaque, we will have to assume that this rate (independently for 5' and 3' LTR sequences) is the same as the million years prior to speciation.

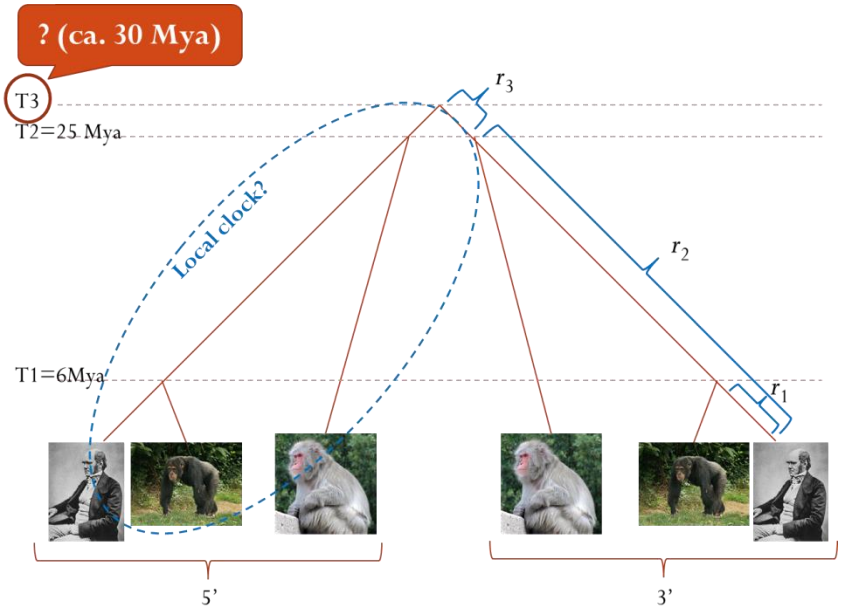

Figure 8: Phylogeny outlined with approximate speciation times. The dashed ring indicates that we would like a local molecular clock for our method to be valid: If r_1 = r2, we can assume that $r_1 = r_3$.

The molecular clock hypothesis

In general, the molecular clock hypothesis is the assumption that the nucleotide substitution rate is constant over time. Since the first use of this in 1962 by Zuckerkandl and Pauling, [11] the modifications of the hypothesis has been numerous. Two dominant branches of the clock thought are the relaxed and the local clock (with the original clock hypothesis referred to as global).

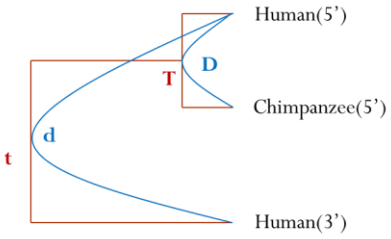

Figure 9: In the molecular clock hypothesis, the proportion between T and D is the same as between t and d, where T or t is time and D or d is distance. In this example, there isn't taken into account that the 3' and 5' LTRs evolve differently.

For the two sets of LTRs, we tested the molecular clock hypothesis in MEGA5. Results are given in Figure 10. The parameters for clock model are *n*-1 and for non-clock model they are 2*n*-3, where *n* is the number of sequences. The clock model is the null hypothesis when we do a likelihood ratio test. In the example, the clock hypothesis is confirmed.

	lnL	Parameters
With Clock	-1304.602	7
Without Clock	-1302.497	11

Figure 10: Results from test of molecular clock hypothesis for the set of ERV3 LTRs in MEGA.

Rather than using a global molecular clock (see Figure 9), we must take into account that the 5' and 3' LTRs evolve separately. As a minimum, we need information about the substitution rate of the 5' LTR obtained from the distance between two species and their speciation time and the same for the 3' LTR.

Integration times

Calculation of the integration time can then be performed using equation

$$T_3 = \frac{D(5'_{human}, 3'_{human})}{\frac{\frac{D(5'_{human}, 5'_{macaque})}{2T_2} + \frac{D(5'_{human}, 5'_{chimpanzee})}{2T_1}}{2} + \frac{\frac{D(3'_{human}, 3'_{macaque})}{2T_2} + \frac{D(3'_{human}, 3'_{chimpanzee})}{2T_1}}{2}} \quad (1.$$

$$T_3 = \frac{D(5'_{human}, 3'_{human})}{\frac{\frac{D(5'_{human}, 5'_{macaque})}{2T_2} + \frac{D(5'_{human}, 5'_{chimpanzee})}{2T_1}}{2} + \frac{\frac{D(3'_{human}, 3'_{macaque})}{2T_2} + \frac{D(3'_{human}, 3'_{chimpanzee})}{2T_1}}{2}} \quad (1)$$

Where the T_1, T_2 and T_3 are speciation times (as outlined in Figure 8) and $D(i,j)$ is the distance between i and j (referring to LTR sequences). The distances are calculated by MEGA (maximum likelihood) and given in **Error! Reference source not found.** (for ERV3). Our results are not completely as expected, as the distance between the 3' LTR sequences should be larger than between the 5' LTR sequences which is not entirely the case. The distances and substitution rates calculated for Syncitin 1 are given in Table 5.

	1	2	3	4	5
1. human 3LTR					
2. pan 3LTR	0.009				
3. macaca 3LTR	0.063	0.069			
4. human 5LTR	0.090	0.093	0.103		
5. pan 5LTR	0.087	0.089	0.097	0.014	
6. macaca 5LTR	0.077	0.079	0.075	0.061	0.061

Table 3: Distance tabel for ERV3

MCL	Distance	Speciation time (Mya)	rate
D55HM	0,114	25	0,00228
D55HC	0,014	6	0,00117
D33HM	0,068	25	0,00136
D33HC	0,009	6	0,00075
D35HH	0,099	?	

Table 4: Rate calculations for ERV3

Comparison	Distance	Speciation time (Mya)	rate
D55HC	0,016	6	0,00133
D55HG	0,030	8	0,00188
D55HO	0,081	14	0,00289
D55HGi	0,108	16	0,00338
D33HC	0,013	6	0,00108
D33HG	0,029	8	0,00181
D33HO	0,058	14	0,00207
D33HGi	0,072	16	0,00225
D35HH	0,075		

Table 5: Rate calculations for ERV-WE1

From these rates, we calculated the speciation time for Syncitin 1: Both using an average of all rates (5' and 3' separately) and the calculated rates based on only human and one more species. The values for the independent rates are plotted against the speciation time for human and the species, the rate was calculated from, in Figure 11. The results show a tendency to be more ancient the more recent the common ancestor lived. Also one result (calculated on basis of human and gibbon ape) lies under the identity line, suggesting that the integration took place after speciation which is highly unlikely since the integration of one virus only happens on an evolutionary time scale and then the chance of the same (perhaps slightly evolved) virus inserting at the same position in some 3 billion base pair long genome is next to nothing. The clear tendency of a decreasing slope indicates that our assumption of a local molecular clock is faulty. This can also be seen from the distance matrix as the average distance between human and gibbon is larger than between the two human LTRs and in that the rates shows a tendency of increasing with the speciation time rising. Hence, based on our results, the selective pressure is different in the different species.

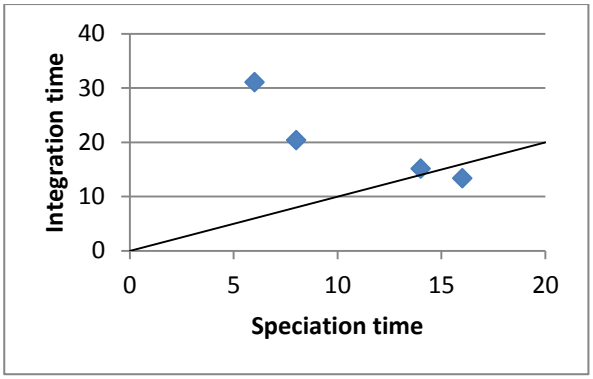

Figure 11: Integration time plotted against speciation time for human and the species the rate used for integration time calculation was calculated from.

Bibliography

[1] International Human Genome Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature,* pp. 860-921, 2001.

[2] N. de Parseval and T. Heidmann, "Human endogenous retroviruses: From infectious elements to human genes," *Cytogenetic and Genome Research,* 2005.

[3] J. Mayer and E. Meese, "Human endogenous retroviruses in the primate lineage and their influence on host genome," *Cytogenetic and Genome Research,* 2005.

[4] A. W. Dangel, B. J. Baker, A. R. Mendoza and C. Yung Yu, "Complement component C4 gene intron 9 as a phylogenetic," *Immunogenetics,* pp. 41-52, 1995.

[5] T. Kijima and H. Innan, "On the Estimation of the Insertion Time of LTR Retrotransposable Elements," *Molecular Biology and Evolution,* 2010.

[6] I. Lobo, "Basic Local Alignment Search Tool (BLAST)," *Nature Education,* 2008.

[7] P. Villesen, L. Aagaard, C. Wiuf and F. S. Pedersen, "Identification of endogenous retroviral reading frames in the human genome," *Retrovirology,* 2004.

[8] C. M. Bergman and H. Quesneville, "Discovering and detecting transposable elements in genome sequences," *Briefings in Bioinformatics,* 2007.

[9] M. Remm, C. E. Storm and E. L. Sonnhammer, "Automatic Clustering of Orthologs and In-paralogs from pairwise species comparison," *Journal of Molecular Biology,* 2001.

[10] P. Perelman, "A Molecular Phylogeny of Living Primates," *PLoS Genetics,* 2011.

[11] E. Zuckerkandl and L. Pauling, Horizons in Biochemistry, M. Kasha and B. Pullman, Eds., New York: Academic Press, 1962, pp. 189-225.

[12] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, "Basic Local Alignment Search Tool," *Journal of Molecular Biology,* 1990.