AMERICAN UNIVERSITY OF BEIRUT

EVOLUTION OF THE RETROTRANSPOSITION ACTIVITY OF LINE-1 IN THE HUMAN GENOME

by
SAWSAN SAMI WEHBI

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Department of Biology
of the Faculty of Arts and Sciences
at the American University of Beirut

Beirut, Lebanon
July,2020

AMERICAN UNIVERSITY OF BEIRUT

EVOLUTION OF THE RETROTRANSPOSITION ACTIVITY OF LINE-1 IN THE HUMAN GENOME

by

SAWSAN SAMI WEHBI

Approved by:

H. Dohna

Dr. Heinrich Burggraf zu Dohna-Schlobitten Advisor
Assistant Professor, Biology

Colin Smith

Dr. Colin A. Smith Member of Committee
Professor, Biology

Dr. Zakaria Kambris Member of Committee
Associate Professor, Biology

Date of thesis/dissertation defense: August 4,2020

AMERICAN UNIVERSITY OF BEIRUT

thesis, dissertation, project release form

Student Name: Wehbi___________Sawsan_____________Sami_______
Last First Middle

✓ Master's Thesis Master's Project Doctoral Dissertation

I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes

after : **One ✓ year from the date of submission of my thesis, dissertation, or project.**
Two ---- years from the date of submission of my thesis, dissertation, or project.
Three ---- years from the date of submission of my thesis, dissertation, or project.

_________________________________13/August/2020_____________

Signature Date

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr Heinrich zu Dohna for his infinite patience and for teaching me how to systematically think not only about scientific issues but life issues as well. You once told me to “Stop worrying about the world, the world will be fine” and I have been trying to do so every day since. I would also like to thank my committee members, Dr Colin Smith and Dr Zakaria Kambris, for being part of this journey and taking the time to provide additional guidance for this project.

Thank you, dad, for giving me the smart genes. With every new scientific discovery, my heart breaks a little that you are not here with us to celebrate them. To my mom, thank you for tolerating me and continuously supplying me with endless love even when I least deserve it.

Finally, I want to express my gratitude to Dr Barbara McClintock, without whom this thesis would not even exist. Thank you for being ahead of your time and believing in your scientific findings when no one else did. I will forever strive to be a strong-minded woman just like you once were.

AN ABSTRACT OF THE THESIS OF

Sawsan Sami Wehbi for Master of Science
Major: Biology

Title: Evolution of the Retrotransposition Activity of LINE-1 in the Human Genome.

LINE-1 (Long Interspersed Nuclear Elements, L1) retrotransposons are the only autonomously active transposable elements in the human genome. Elevated retrotransposition activity of L1s are generally thought to be detrimental to the host. This study performed several analyses to investigate the evolution of the retrotransposition capacity among full-length human L1s. The first analysis showed that the rates at which new L1s emerge are positively correlated with retrotransposition activity values reported for cell-culture based assays, indicating that the retrotransposition activity values measured in cell culture can be considered as valid proxies for human germline retrotransposition capacity. An analysis that estimated the evolution of retrotransposition activity along the phylogenetic tree by sampling over various ancestral state configurations, revealed an evolutionary trend towards lower activity states. A penalized regression model was constructed to identify L1 sequence positions that might directly alter retrotransposition activity, yet the identified positions were not associated with any known biological L1 features. Analysis of sequence variation within L1 loci uncovered single nucleotide polymorphism (SNP) depletion within L1 transcription factor binding (TFB) sites, indicating evolutionary conservation of these sites by the host. Previous studies have shown that some of these TFB sites are essential for successful retrotransposition of human L1s. Together, these findings point towards the potentially ambivalent, yet balanced nature of the host-transposon relation.

CONTENTS

Chapter

LIST OF FIGURES

LIST OF TABLES

Table Page

ABBREVIATIONS

%: percent

BF: bayes factor

BiSSE: binary state speciation and extinction

DNA: deoxyribonucleic acid

L1: LINE-1, Long Interspersed Nuclear Elements 1

LTR: long terminal repeat elements

MCMC: monte carlo markov chain

ORF: open reading frame

RNA: ribonucleic acid

SINE: short interspersed nuclear elements

SNP: single nucleotide polymorphism

TE: transposable elements

TFB: transcription factor binding

UTR: untranslated region

qHL: rate of transition from high activity states to low activity states

qLH: rate of transition from low activity states to high activity states

CHAPTER I

INTRODUCTION

Transposable elements (TEs) are mobile DNA segments that comprise more than half of the human genome (Burns & Boeke, 2013). They are classified as short interspersed elements (SINEs), long interspersed elements (LINEs) or long terminal repeat elements (LTR elements). L1s are a class of LINEs and are the only active autonomous retrotransposons in humans. These transposable elements use a "copy" and "paste" mechanism to create new insertions (Burns & Boeke, 2013). This retrotransposition mechanism depends on transcription to produce mRNA-like intermediates that are then reverse transcribed and inserted in the genome within 15~16 bp to a target sequence containing TTAAAA sense hexanucleotides (Jurka, 1997). Mobile L1s are bicistronic; their 5' cistron ORF1 encodes an RNA-binding protein and their 3' cistron ORF2 encodes a large protein with endonuclease and reverse transcriptase domains (Dai et al., 2014). These cistrons are flanked by 5'UTR and 3'UTR (Dai et al., 2014), dividing the L1 sequence into four distinct functional regions.

De novo insertions created by successful retrotransposition events can be disruptive to the host genome. Some specific insertions have been linked to diseases such as hemophilia and thalassemia (Hancks & Kazazian, 2016). In addition, overall high rates of retrotransposition may increase the risk of proliferation and metastasis of epithelial cancer (Rangasamy et al., 2015) and have been associated with the psychiatric disorder schizophrenia (Bundo et al., 2014). These results suggest that retrotransposition activity is disadvantageous and would be expected to be reduced over time by host-level selection. Random mutations of L1 sequences would be expected to decrease the retrotransposition activity rather than increase it, which would also lead to a decrease of a newly inserted L1's activity over time. Transposon frequency in host genomes depends on selection forces acting on both, the host and the transposon itself. Transposon-level selection favors higher retrotransposition rates (Le Rouzic, Boutin, & Capy, 2007), whereas host-level selection which governs post-insertion allelic variation is often assumed to act against retrotransposition activity. This difference in selection levels is consistent with macroevolutionary patterns that suggest an co-evolutionary arms race between L1s and their hosts' defense mechanisms (Jacobs et al., 2014).

Despite this, L1s appear to have become an integral part of their hosts' developmental process. In mice, for example L1 transcription regulates chromatin accessibility during embryogenesis, which appears to be integral for proper mouse embryo development (Jachowicz et al., 2017). In addition, L1 retrotransposition can also introduce somatic mosaicism (Kano et al., 2009). Neuronal mosaicism due to L1 retrotransposition contributes to neuronal diversity and has been recently suggested to play a role in learning and memory (Kurnosov et al., 2015). These results challenge the view that retrotransposition is always deleterious for the host.

Given that transposons have disruptive and beneficial effects on their host, the host-transposon relation might be better viewed as more of an ambivalent compromise and less of an arms race (Castro-Diaz et al., 2015). The host silences the TEs to avoid genomic chaos but endures occasional insertions. Although these insertions are potentially dangerous, they may also increase genomic plasticity and introduce new genes through molecular domestication. It has been proposed that species harboring active TEs have more dynamic genomes which may increase the species' adaptability to environmental change, enhancing its chances of survival (Ricci et al., 2018). These dynamic genomes may domesticate certain TE insertions into beneficial host genes. Several functional human genes can be traced back to TE-derived coding sequences. In fact, TE-derived sequences have contributed to almost 4% of human protein-coding genes (Nekrutenko & Li, 2001). One of the most well-characterized examples of TE-derived essential human genes are the recombination-activating genes RAG1 and RAG2 that play a key role in adaptive immunity (Sinzelle, Izsvák, & Ivics, 2009).

Variation between L1 sequences at different loci and within the same locus contributes to extensive individual variation in retrotransposition activity (Beck et al., 2010; Brouha et al., 2003; Seleme et al., 2006). These sequence variations may affect the retrotransposition process on multiple levels such as transcriptional regulation and translation levels. Mutations in a L1 might modify the amino acid sequence of proteins encoded by the L1, and even alter the L1's interaction with transcription factor binding (TFB) sites (Sun et al., 2018). Such mutations that can be directly associated with changes in the retrotransposition activity levels are yet to be characterized. Identifying these types of mutations might allow predicting L1 retrotransposition activity based on their sequence, might reveal which functional aspects of L1 sequences (e.g. TFB sites, ORFs) influence the retrotransposition activity, and improves our understanding of the interrelation between L1 sequence variation and the evolution of the L1 retrotransposition capacity.

On a long evolutionary timescale, all transposon insertions will eventually become dormant. For example, none of the L1 insertions that predate the divergence of humans from other primates are still active (Boissinot, Chevret, & Furano, 2000). However, on a smaller timescale, the evolution of individual L1 sequences and the fluctuation of their corresponding activity levels is still unknown. Considering random mutations often lead to loss of function, one would assume that the microevolution of L1s will tend towards lower activity levels. Yet, given that L1 sequences and their retrotransposition can be beneficial at certain circumstances, the possibility of a short-term increase in the L1 retrotransposition activity with time can not be ruled out. Examining the microevolution of L1 retrotransposition activity would shed further light on the host-transposon relationship.

Investigating the evolution of the retrotransposition activity of L1s was only made possible due to the development of the retrotransposition activity assay (Moran et al., 1996), which allowed the exogenous quantification of TE retrotransposition capacity. This technique was fundamental in expanding our knowledge in the mechanisms and consequences of retrotransposition in mammalian cell cultures (Rangwala & Kazazian, 2010). Several variations of this technique are available, perhaps the most commonly used is the enhanced green fluorescent protein (EGFP) retrotransposition assay (Beck et al., 2010; Brouha et al., 2003; Seleme et al., 2006). This assay depends on the construction of TE-containing plasmids with an EGFP cassette inserted in the 3'UTR region of the TE sequence. The expression of the EGFP acts as a visual marker of the TE's movement and can be quantified using flow cytometry. The measurement of the amount of green fluorescent emitted after transfection is considered directly proportional to the rate of retrotransposition within the cultured cell lines (Ostertag et al., 2000). However, it is uncertain whether the numerical values obtained by the retrotransposition assays reflect the in-situ germline retrotransposition activity in humans that is relevant for evolutionary change.

A phylogenetic approach can be applied to analyze the extent to which laboratory-measured values reflect the in-situ human germline retrotransposition capacity. In a phylogenetic tree of L1s found in the human reference genome, each external branch represents a successful germline insertion event that has been identified and sequenced. Higher frequencies of successful germline insertions would lead to faster rates of emergence of new L1s in the population. Such higher rates of emergence of new L1s would be reflected on the phylogenetic tree as extensive bifurcations. Within the context of comparative analysis, observed bifurcation rates on the phylogeny are often referred to as diversification rates. A branching pattern in which L1s with high laboratory-measured retrotransposition activity values are associated with high L1 diversification rates on the phylogeny, would indicate that the retrotransposition assay measurements are representative of the L1 germline insertion rates within the human genome.

Exploring the evolutionary mechanisms that govern the human L1s broadens our understanding of the influence TEs have on biological systems. The general aim of this study was to understand how the extensive individual variation of human L1s reflects the within-locus and between-locus evolution of L1 sequences and their retrotransposition activity. The terms “allelic” and “non-allelic” were used to distinguish between different L1s within the same loci and L1s on different loci, respectively. The post-insertion evolution of L1 retrotransposition activity in human genomes was investigated. Specifically, this study addressed the questions whether the L1 retrotransposition activities measured in cell culture are also reflected in L1 diversification patterns in human genomes, whether there is an evolutionary trend in retrotransposition activity among L1s, and which features of L1 sequences are associated with retrotransposition activity.

CHAPTER II

MATERIALS & METHODS

A. Sequence Collection and Alignment

All of the sequences used in this study were obtained from previous studies that published full-length L1 sequences and their corresponding retrotransposition activity values (Brouha et al. 2003, Seleme et al. 2006, Beck et al. 2010). The retrotransposition activity values were calculated as percentages in reference to a naturally occurring active L1 (i.e L1RP or L1.3).

The L1 sequences were grouped into two sets. The first set contains 52 allelic L1 sequences from three hot L1 loci which were originally obtained by Seleme et al. (2006). The term "hot" refers to L1s with at least one third the activity of the reference L1. The second set contains 158 non-allelic sequences that were studied by both Brouha et al. (2003) and Beck et al. (2010). The L1 sequences from Brouha et al. were extracted from an alignment published in their supporting information (Brouha et al., 2003). The L1 sequences from Beck et al. were obtained by identifying from the supporting information (Beck et al., 2010) the sequences flanking L1 insertions, and locating these flanking sequences in the hg19 reference genome.

Both sets of sequences were aligned using the multiple pairwise alignment software, MUSCLE (Edgar, 2004) on Molecular Evolutionary Genetics Analysis (MEGA) software (Kumar, Stecher, & Tamura, 2016). The consensus sequence of the ancestral L1PA2 was added to both alignments, as an outgroup.

B. Tree Reconstruction

The phylogenetic trees were reconstructed from the alignments using MrBayes v.3.2 (Ronquist et al., 2012) under a general time reversible substitution model. Rate variation between individual base pairs were modelled with a combination of a gamma distribution for rate variation and a proportion of invariant sites. Two Monte-Carlo Markov chains MCMC were run until the average standard deviation of split frequencies became lower than 0.05. A sample frequency of 100 was set to generate samples from posterior distributions of trees and substitution rate parameters, given the alignment data.

The consensus trees of the allelic and non-allelic sequences were determined through the majority rule and illustrated on Figtree v1.3.1. (Rambaut, 2009) and R. Both trees were rooted using L1PA2 which was then dropped from the trees.

C. Binary State Speciation and Extinction (BiSSE) model

The non-allelic tree was used to analyze the association between the diversification rate of L1s and the retrotransposition activity using the R package “diversitree” (FitzJohn, 2012). The L1 retrotransposition activities were coded as a binary character, where L1 alleles with retrotransposition activities lower or equal to 25% were classified as low and otherwise as high. A Binary State Speciation and Extinction (BiSSE) model, which combines the features of the constant-rates birth-death model with the two-state Markov model, was used. According to the BiSSE model, the two character-states (in this case high and low retrotransposition activity), can alternate along the tree based on the Markov process, and the birth and death rates can depend on the character-states. The birth rates correspond to the rate of emergence of new L1s whereas the death rates correspond to the rate at which L1s are removed from the population. Two models were fit to the data, a null model that restricted the birth rates of the high and low activity states to be equal and an alternative model that had no constraints on neither death nor birth rates. The fit of these two models was compared using a likelihood ratio test.

The non-allelic tree was modified because the BiSSE model requires bifurcating ultrametric trees. Polytomies were converted into a sequence of dichotomies by randomly grouping branches in a polytomy into pairs that form a bifurcating node and separating each bifurcating node by a branch that is small compared to the branch lengths in the input tree. The resulting tree was then forced to be ultrametric using the “force.ultrametric” function in the R package “phytools” (Revell, 2012). To determine whether the results are robust with respect to the tree modifications, six different variations of the process that generated an ultrametric bifurcating tree were run. The conversion to bifurcating trees was performed in three different ways, each using a different length value (10^{-6}, 10^{-5}, 10^{-4}) for the branches that separate bifurcating nodes. The three ways to generate a bifurcating tree were combined with two different methods for forcing an ultrametric tree (options “extend” and “nnls” in the function “force.ultrametric”). A one-sample t test was performed to test whether the change in edge lengths through forcing ultrametry differed between tree tips corresponding to high and low activity-state L1.

D. Transition Model Estimation

The L1 retrotransposition activities were coded as a binary character, where L1 alleles with retrotransposition activities lower or equal to 25% were classified as low and otherwise as high. The evolution of the transition rates between these two states along the phylogenetic trees were estimated using BayesTraits, which implements a Bayesian MCMC to estimate posterior distributions of evolutionary rates between discrete characters along a phylogenetic tree (Meade & Pagel, 2016).

The posterior distribution of trees generated by MrBayes was used as input data for the transition model analysis. Three different uniform prior distributions were used for the rate parameters and the analysis below was run for each distribution. The midpoint of these priors was the maximum likelihood estimate for the model with equal transition rates between the two activity states. The range widths of the priors were 200%, 100% and 50% of the midpoint. In order to investigate the direction of L1 evolution on both allelic and non-allelic sequences, two different models were constructed. The complex model, which has no rate constraints, allows the rates for the transitions from high to low retrotransposition to differ from the reverse rates whereas the simple model restricts these rates to be equal. A steppingstone sampler of 100 stones each run for 1,000 iterations was used to estimate the log marginal likelihood of each model. The log Bayes Factor (BF), which is calculated using the log marginal likelihoods of the two models, was used to infer the best fitting model (Kass & Raftery, 1995).

E. Penalized Regression Model

A penalized regression model was fitted to uncover a possible relationship between the retrotransposition activity and single nucleotide variant positions in an alignment of full-length L1. The alignment included all sequences from the three studies with published retrotransposition activities (Brouha et al. 2003, Seleme et al. 2006, Beck et al. 2010). A binary dummy variable was created for each combination of nucleotide and alignment position among variable columns in the alignment. In addition, a study ID was included as predictor variable to account for variability between studies.

The lambda minimizing the prediction error (lambda=2.7) was estimated using the cross-validation method. In order to obtain a conservative set of non-zero coefficients, the lambda minimizing the prediction error was then increased incrementally until the number of non-zero coefficients generated by the model plateaued, leading to a lambda of 4.

The alignment positions were classified into two groups according to their regression coefficients assigned by the penalized model. Positions with a zero regression coefficient were assumed to be independent of the retrotransposition activity whereas positions with a non-zero regression coefficient were assumed to be activity-affecting sites.

F. Analyzing the Characteristics of L1 Activity-Affecting Sites

The relationship between activity affecting sites and non-synonymous positions in the L1 open reading frames was investigated using a Fisher exact test.

Variable positions in the alignment were either classified as retrotransposition-affecting or non-affecting, based on the coefficients of the penalized regression. Each position in the alignment was furthermore classified as either synonymous, non-synonymous or both. The non-synonymous positions in the alignment were determined for each sequence separately according to their reading frames. An alignment position can therefore be synonymous in one sequence and non-synonymous in another. This classification was carried out treating the first two nucleotides in a codon as non-synonymous while the last nucleotide in a codon as synonymous, regardless of the amino acid it encodes. The Fisher exact test was performed to test for independence between these two classifications.

G. Mapping Transcription Factor Binding (TFB) Sites

Several TFB regions that correspond to different cell lines and transcription factors, were mapped to the L1HS consensus sequence by Sun et al. (2018). Activity affecting sites that fall within these regions were calculated independently for each cell line and transcription factor. A Fisher's exact test with a false discovery rate correction was iterated case by case in order to infer a possible association between activity affecting sites and transcription factor binding sites across different cell lines and different transcription factors.

The TFB sites were mapped to their genomic positions on hg19, with respect to each L1 sequence in the data.

H. Mapping Single Nucleotide Polymorphisms (SNPs)

The activity affecting sites that correspond to the alignment were mapped to their respective genomic position for each L1 sequence on the human reference genome hg19. These genomic positions were then compared to the SNPs within the variable positions of the L1 ranges (flanks included) in the 1000 human genome database.

A multiple logistic regression was used to determine whether SNPs from the 1000 Genome project (Sudmant et al., 2015) were associated with activity affecting sites or transcription factor binding sites. The absence/presence of SNPs on a particular genomic position was annotated as a binary response value, whereas a similar binary classification was used to set the absence/presence of genomic activity affecting positions as a predictor variable. The trinucleotide environment, coding ranges, mean of the read coverages, nonsynonymous positions, variable flank positions and the proportion of mismatch were added as covariates for the regression model.

In addition, a binomial test was used to investigate whether ancestral SNPs that fall on activity affecting sites in the L1 alignment are more likely to increase or decrease the retrotransposition activity.

CHAPTER III

RESULTS

1. L1 Allelic vs. Non-Allelic Phylogenies

The L1 allelic tree (**Figure 2**) resolved the three L1 loci, with high posterior probability of one. The tree topology agrees with the coalescence time calculated by Seleme et al., (2006) for each L1, where AL512428 is the most ancestral and has the longest coalescence time of 590 ±160x10^3 years (Seleme et al.,2006). Only one clade (AC002980) retained high retrotransposition activity values among most of its alleles, while the other two (AL512428 and AC021017) mostly contain alleles with lower activities.

The non-allelic tree (**Figure 3**) shares a similar topology to that estimated by Brouha et al., (2003). Most of the clades were well resolved with high posterior probability values (>80%). Longer and older branches tend to have low retrotransposition activity values whereas shorter and more recent branches tend to have higher values. This correlation between branch length and retrotransposition rates was further investigated using a Binary State Speciation and Extinction (BiSSE) model. The null model with equal birth rates for each binary state (high and low retrotransposition activity) was rejected ($p < 2.2x10^{-16}$, likelihood ratio test), in favor of an alternative model in which the birth rate of highly active L1s (λ_1 =72618) is higher than the birth rate of L1s with low activity levels ($\lambda_0 = 0$). The different methods used to convert the non-allelic phylogenetic tree into an ultrametric bifurcating tree lead to the same conclusion. In addition, there was no significant difference in the branch lengths before and after introducing the branch length adjustments, between the high activity and low activity taxa ($p > 0.05$, one-sample t-test).

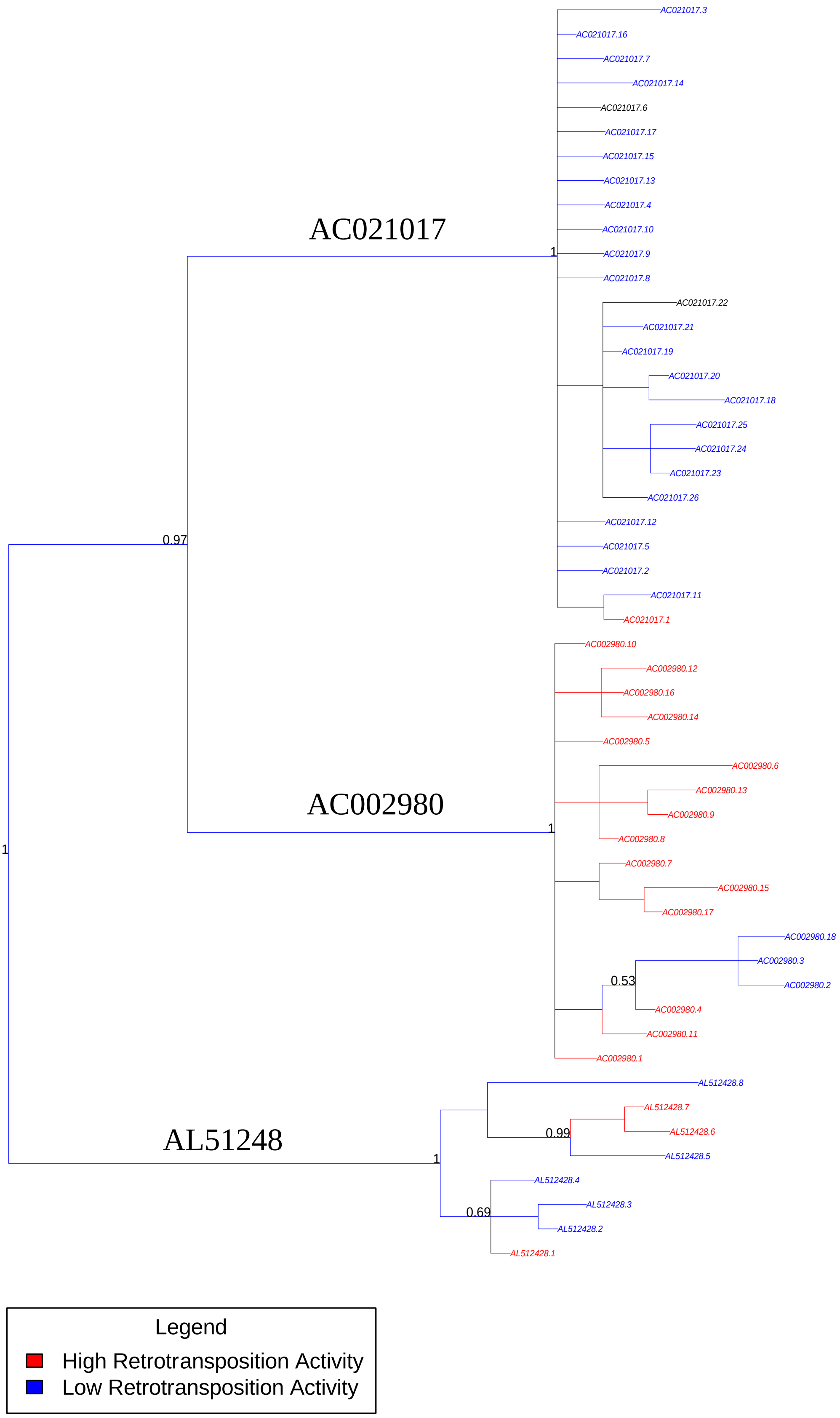
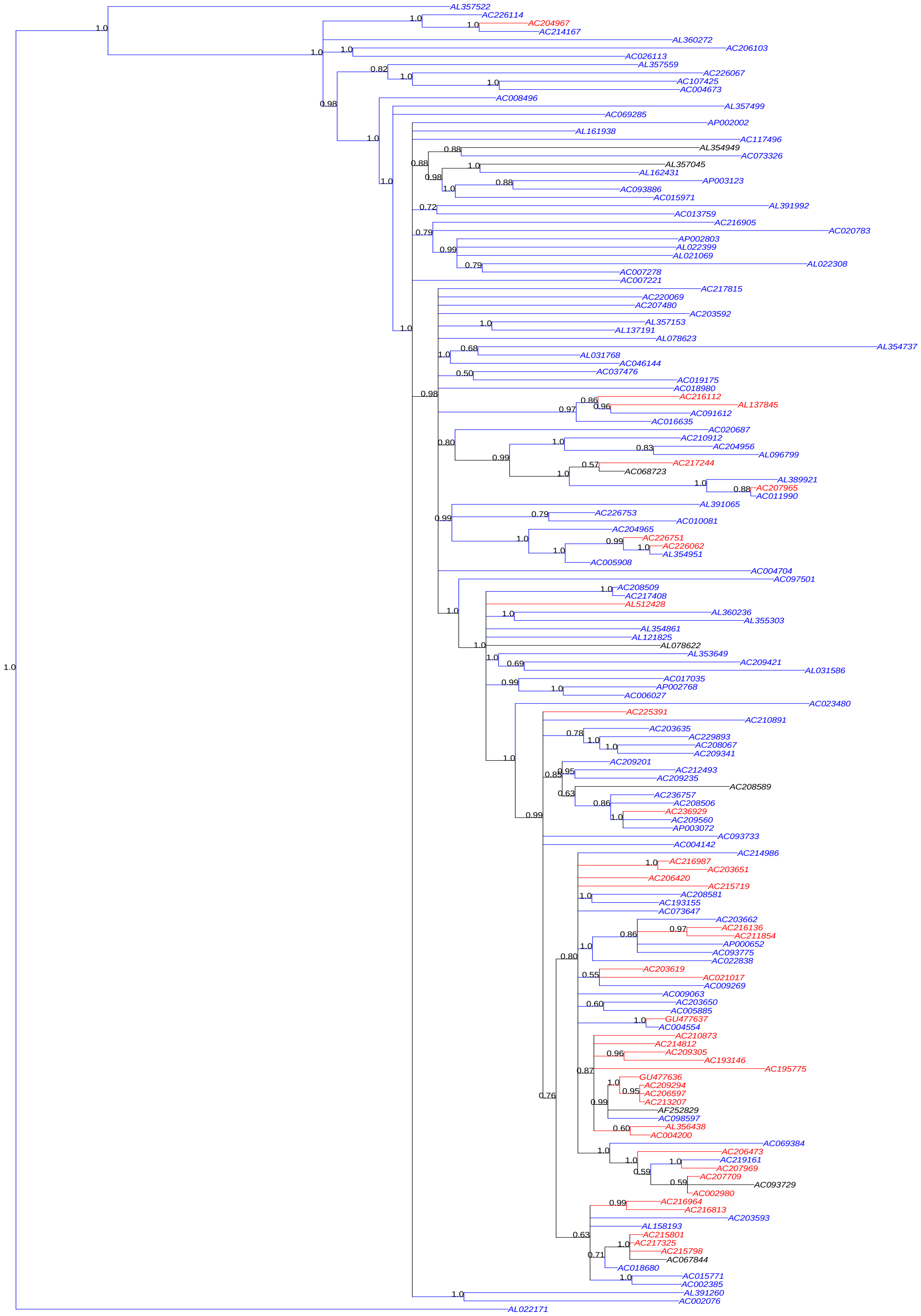

AC021017
AC021017.3
AC021017.16
AC021017.7
AC021017.14
AC021017.6
AC021017.17
AC021017.15
AC021017.13
AC021017.4
AC021017.10
1
AC021017.9
AC021017.8
AC021017.22
AC021017.21
AC021017.19
AC021017.20
AC021017.18
AC021017.25
AC021017.24
AC021017.23
AC021017.26
0.97
AC021017.12
AC021017.5
AC021017.2
AC021017.11
AC021017.1
AC002980.10
AC002980.12
AC002980.16
AC002980.14
AC002980.5
AC002980.6
AC002980.13
AC002980
AC002980.9
1
AC002980.8
1
AC002980.7
AC002980.15
AC002980.17
AC002980.18
AC002980.3
0.53
AC002980.2
AC002980.4
AC002980.11
AC002980.1
AL512428.8
AL512428.7
AL51248
0.99
AL512428.6
AL512428.5
1
AL512428.4
AL512428.3
0.69
AL512428.2
AL512428.1
Legend
🟥 High Retrotransposition Activity
🟦 Low Retrotransposition Activity

Figure 1: **Bayesian MCMC phylogeny based on 52 allelic sequences from 3 hot L1s**, obtained by Seleme et al., (2006). Tree branches were colored by retrotransposition activity levels, where sequences with an activity value of 25% or lower were considered low and labelled blue and sequences with activity values higher than 25% were considered high and labelled red. Taxa with unavailable retrotransposition activity values remained unlabeled (black tips). Tip labels refer to the GenBank accession numbers under which each L1 was published. The major clades are shown on the tree and the posterior probability values are shown at major nodes. The tree was rooted using the consensus ancestral sequence of L1PA2.


AL357522
1.0
AC226114
1.0
AC204967
1.0
AC214167
AL360272
AC206103
1.0
1.0
AC026113
AL357559
0.82
AC226067
1.0
AC107425
1.0
AC004673
0.98
AC008496
AL357499
AC069285
AP002002
AL161938
1.0
AC117496
1.0
AL354949
0.88
AC073326
0.88
AL357045
1.0
AL162431
0.98
AP003123
0.88
AC093886
1.0
AC015971
AL391992
0.72
AC013759
AC216905
AC020783
0.79
AP002803
AL022399
0.99
AL021069
AL022308
0.79
AC007278
AC007221
AC217815
AC220069
AC207480
AC203592
AL357153
1.0
AL137191
1.0
AL078623
0.68
AL354737
1.0
AL031768
AC046144
0.50
AC037476
AC019175
AC018980
0.98
AC216112
0.86
AL137845
0.96
0.97
AC091612
AC016635
AC020687
AC210912
1.0
0.80
AC204956
0.83
AL096799
0.99
AC217244
0.57
AC068723
1.0
AL389921
1.0
AC207965
0.88
AC011990
AL391065
0.79
AC226753
0.99
AC010081
AC204965
1.0
AC226751
0.99
AC226062
1.0
AL354951
1.0
AC005908
AC004704
AC097501
AC208509
1.0
AC217408
AL512428
1.0
AL360236
1.0
AL355303
AL354861
AL121825
AL078622
1.0
AL353649
1.0
AC209421
0.69
AL031586
AC017035
0.99
AP002768
1.0
AC006027
AC023480
AC225391
AC210891
AC203635
0.78
AC229893
1.0
AC208067
1.0
AC209341
1.0
AC209201
AC212493
0.95
0.85
AC209235
AC208589
0.63
AC236757
AC208506
0.86
AC236929
1.0
AC209560
0.99
AP003072
AC093733
AC004142
AC214986
AC216987
1.0
AC203651
AC206420
AC215719
1.0
AC208581
AC193155
AC073647
AC203662
AC216136
0.97
0.86
AC211854
AP000652
1.0
AC093775
0.80
AC022838
AC203619
0.55
AC021017
AC009269
AC009063
0.60
AC203650
AC005885
1.0
GU477637
AC004554
AC210873
AC214812
0.96
AC209305
AC193146
0.87
AC195775
GU477636
1.0
AC209294
0.95
AC206597
0.76
AC213207
0.99
AF252829
AC098597
0.60
AL356438
AC004200
AC069384
1.0
AC206473
1.0
AC219161
1.0
AC207969
0.59
AC207709
0.59
AC093729
AC002980
0.99
AC216964
AC216813
AC203593
AL158193
AC215801
0.63
AC217325
1.0
AC215798
0.71
AC067844
AC018680
AC015771
1.0
AC002385
1.0
AL391260
AC002076
AL022171

Legend

🟥 High Retrotransposition Activity
🟦 Low Retrotransposition Activity

Figure 2: **Bayesian MCMC inference phylogeny based on the 158 non-allelic L1 sequences** obtained by Brouha et al., (2003) and Beck et al., (2010). Tree branches were colored by retrotransposition activity levels, where sequences with an activity value of 25% or lower were considered low and labelled blue and sequences with activity values higher than 25% were considered high and labelled red. Taxa with unavailable retrotransposition activity values remained unlabeled (black tips). Tip labels show the GenBank accession numbers of each L1 was published. All the clades are shown on the tree and the posterior probability values are shown at the nodes of each clade. The tree was rooted using the consensus ancestral sequence of L1PA2.

2. Evolution of the Retrotransposition Activity along the Phylogenetic Trees

Two evolutionary models for the transitions between two different retrotransposition activity levels were fitted to phylogenetic trees of L1. The results obtained from different prior distributions were qualitatively similar and only the numerical values based on the widest prior distribution are reported here.

In the allelic dataset, the two rates of transition were approximately equal where the rate of transition from low retrotransposition activity to high activity (qLH) was 81.17 and the rate of transition from high retrotransposition activity to low activity (qHL) was 80.36 (**Table 1**). The complex model has a slightly lower log marginal likelihood than the simple model which indicates a weaker fit to the data (log BF=-0.1< 2). This means that on the allelic level, a L1 has an equal probability of evolving towards either a higher activity or a lower activity.

However, for the non-allelic two-state phylogenetic tree, the complex model fits the tree best (log BF=3.8), where qHL is almost twofold qLH (**Table 1**). This implies that a reduction in the retrotransposition activity of L1s is twice as frequent as an increase.

Dataset	***Parameter restriction***	***Number of free parameters***	***Transition Rates***	***Log marginal likelihood***

Allelic sequences	Unrestricted rates	2	qLH=81.17
			qHL=80.36
				-34.60

	Restricted rates	1	qLH=qHL=81.70	-34.55

Non-allelic sequences	Unrestricted rates	2	qLH=37.82
			qHL=81.70
				-78.94

	Restricted rates	1	qLH=qHL=51.00	-80.87

Table 1: Comparison of the two different models for each L1 dataset. The model with higher number of parameters is considered the complex model. The reported numerical values were generated using the widest prior distribution. The noted transition rates are the average of transition rates from each iteration in the Monte Carlo Markov Chains.
qHL = rate of transition from high retrotransposition activity state to low activity state.
qLH = rate of transition from retrotransposition low activity state to high activity state.

3. L1 Activity-Affecting Alignment Positions

The retrotransposition activity values predicted by the penalized regression model correlated with the observed retrotransposition activity values (R_2 = 0.81) (**Figure 3**).

Thirty-four alignment positions (out of 2162 variable positions, ~1.6%) had a non-zero regression coefficient for predicting the retrotransposition activity. Due to the format of the binary regression matrix used in the penalized model, some of these alignment positions had two coefficients. Each of these coefficients belonged to a different nucleotide variant on the same position, which yields a total number of 47 activity-affecting nucleotides in the alignment.

The distinction between synonymous and non-synonymous sites is not significantly associated with the distinction between the activity affecting positions and non-affecting positions.

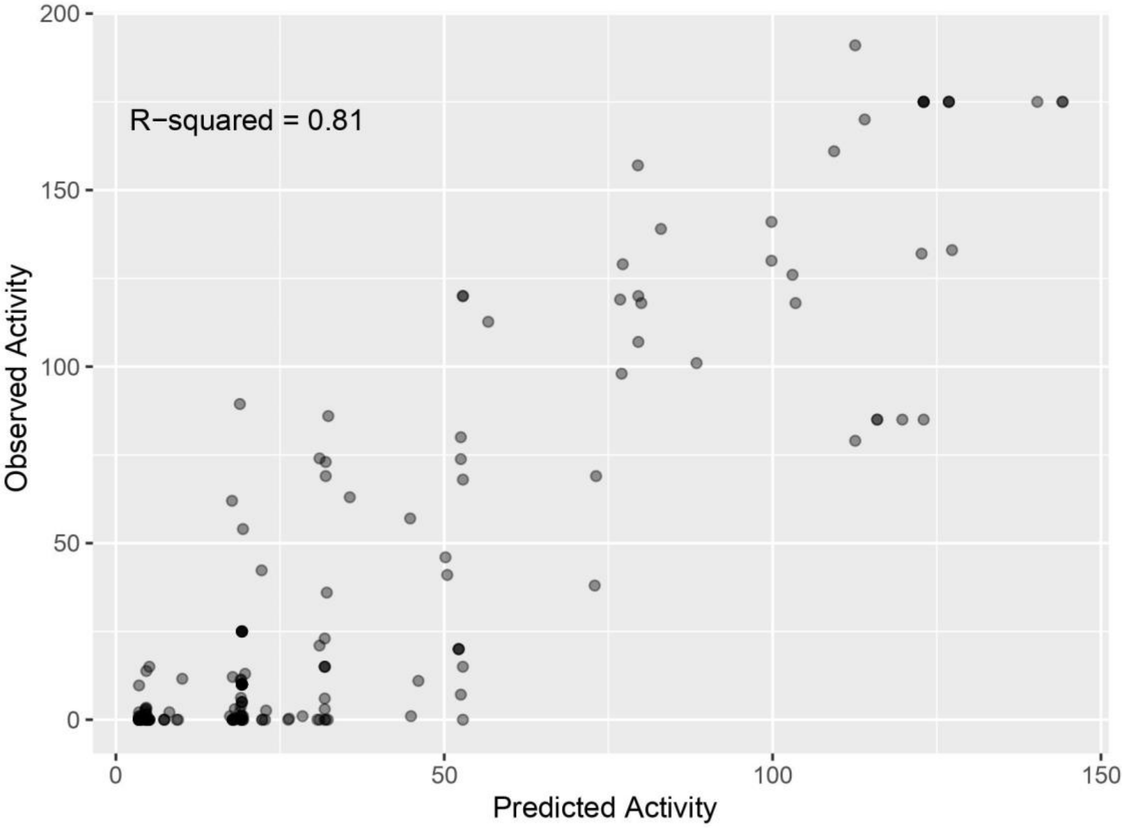


Figure 3: **Scatterplot of the observed retrotransposition activity values and the values predicted by the penalized regression model**.

4. L1 Transcription Factor Binding Sites

After applying the false discovery rate correction, there was no association between transcription factor binding sites and activity affecting positions across all cell lines and transcription factors ($p > 0.05$, Fisher exact test).

All the activity affecting TFB sites found across the different cell lines and transcription factors were concatenated and their association to L1 SNPs was interpreted using the multiple binary regression model. There was a significant negative association between L1 TFB sites and SNPs ($p < 0.05$, logistic regression), where TFB sites were less likely to be polymorphic than expected by chance (**Figure 4B**).

5. L1 Activity-affecting SNPs

Within the variable positions on L1, ten SNPs fell on activity affecting genomic positions which can be traced back to six unique alignment positions (**Figure 4A**). There is no association between L1 SNPs and activity affecting sites ($p > 0.05$, logistic regression). The distribution of SNPs does not significantly vary between the activity affecting positions and non-affecting positions.

Only seven out of these ten positions can be assigned to an ancestral nucleotide, three of which had a positive coefficient. SNPs were not associated with an increase or decrease the activity.

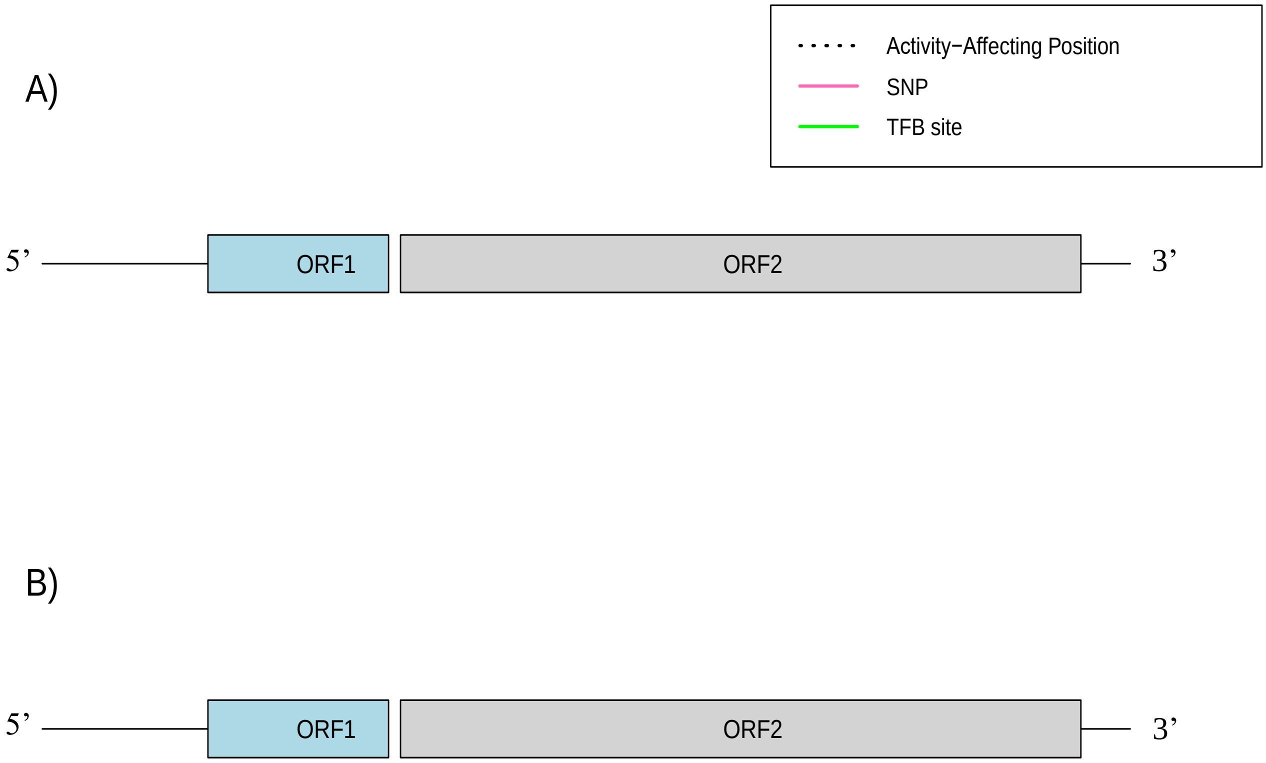


Figure 4: **The distribution of the activity affecting sites, SNPs and TFB sites on the L1 structure.** The blue box indicates the open reading frame 1 (ORF1), the grey box the open reading frame 2 (ORF2) and the lines on both sides the untranslated regions. **(A)** The positions of the activity affecting sites (dotted black) along with the L1 SNPs (pink) and TFB ranges (green) that exclusively fall within these sites. **(B)** All of the known SNPs (pink) that fall within the L1 sequence ranges and the L1 associated TFB sites (green) in different cell lines and across different transcription factors.

CHAPTER IV

DISCUSSION

The major clades in the phylogenetic trees were properly resolved. A biased direction in the evolution of the retrotransposition activity was not detected on an allelic level. However, the retrotransposition activity displayed a higher transition towards lower values within non-allelic L1s. The non-allelic L1 sequences exhibited a positive correlation between L1 diversification rates and their retrotransposition activity. Sites that influenced retrotransposition activity did not show any association with the TFB sites, SNPs or the non-synonymous positions within the L1s. Finally, L1 TFB sites were found to be relatively conserved, with fewer SNPs falling on them than expected by chance.

On the non-allelic phylogenetic tree (**Figure 3**) less active L1s exhibited longer branches whereas more active L1s exhibited shorter branches. According to the BiSSE model that was fitted to the L1 phylogeny, L1s with higher retrotransposition activities branch more frequently on the tree. Since each branch in the phylogenetic tree corresponds to a different germline L1 insertion, the significant correlation between retrotransposition rate and branching rate confirms that retrotransposition activities measured in cell culture assays reflect the rate at which germline L1 insertions appear in the human population. Since L1 germline insertions are hard to measure directly, cell culture-based retrotransposition activity measurements have been used as proxies (Brouha et al.,2003). The results show that these previous interpretations of culture-based activity values are justified, which broadens the possibility of future studies of TE activity.

The analysis of the diversification rates required converting the non-allelic tree into an ultrametric bifurcating tree. In principle, this manipulation of the branch lengths and tree topology could have enhanced the difference in diversification rates between the two activity states. However, the changes on the branch lengths due to the tree manipulation did not differ between the low activity taxa and the high activity taxa. In addition, the results were qualitatively the same for the different tree modifications. This means that the higher branching rate of highly active L1s is robust and unlikely to be an artifact of the tree manipulation

The evolutionary trend towards lower retrotransposition rates was established using the log Bayes Factor (BF). BFs have been criticized for being overly sensitive to the prior distribution (Morey, Romeijn, & Rouder, 2016). In this study, different prior distributions of transition rates were used, where narrower prior ranges lead to lower BF values. However, for each prior distribution, the complex model remained more likely than the simple model (had a higher marginal likelihood). In the context of this analysis, broad priors are preferable since they allow larger differences between the rate parameters. Even though the different prior distributions yielded different log BF values, the broadest range had a log BF value of 3.8 which can be considered as positive evidence for the complex model (Meade and Pagel, 2016). The results, therefore, indicate an evolutionary trend towards lower L1 retrotransposition activity between loci.

The evolution of the retrotransposition activity within loci reflects the selective pressures experienced by the host, whereas variation between L1s on different loci is under both host-driven and transposon-driven selection. In general, highly active L1s are thought to be negatively selected by the host due to their potential to disrupt the genome and positively selected among transposons. Among non-allelic L1s there is a detectable evolutionary bias towards reduced activity states, suggesting that the host-driven selection for lower retrotransposition rate is stronger than transposon-level selection for highly active L1s. One would therefore expect a stronger evolutionary bias towards lower retrotransposition activity states within the same L1 locus, since within-locus allelic variations of L1s are only subjected to host-driven selection. Hence, the absence of a clear trend in the evolution of retrotransposition activity on an allelic level within a locus is most likely due to the small sample size and low number of activity state transitions within the alleles rather than the absence of a negative host-driven selective force. Further sequencing of additional within-locus L1 sequences and measuring their respective retrotransposition activity is required for a comprehensive understanding of the within-locus evolution of L1 retrotransposition activity.

Sites that significantly affect retrotransposition were neither associated with transcription binding sites nor with non-synonymous sites on ORFs. The absence of any significant association is either because the methodology failed to identify the true drivers of retrotransposition activity or because retrotransposition activity is determined by a more complex underlying biology.

The multiple logistic regression model, that detected the negative association between L1 TFB sites and L1 SNPs, included covariates that accounted for variation in SNP occurrence due to variation in detection levels (accounted for by mean read coverage) and variation in mutation rates (accounted for by number of SNPs in flanking regions, proportion of mismatch, and trinucleotide neighborhood). It is possible that there are other unknown and unaccounted for phenomena might have influenced the observation of L1 SNP depletion. However, given that the model accounted for known processes that could influence the mutation or SNP detection rate, it is most parsimonious to interpret the remaining systematic variation in SNP density as a result of differences in purifying selection.

SNPs were depleted on TFB sites in the L1 untranslated regions, indicating that host-level selection conserves TFB sites. A conservation of L1 TFB sites was also observed by Sun et al. (2018), who compared the consensus sequences of the ancestral L1s (L1PA1-L1PA7) with the consensus L1HS. However, this pattern is likely to be caused by transposon-level selection, since some of these TFB sites are essential for the initiation and propagation of L1 transcription, and eventually retrotransposition. For example, mutations in the RUNX3 transcription factor binding sites on the L1HS 5’-UTR lead to a reduction in the transcription and retrotransposition levels measured using cell culture-based assays (Yang et al., 2003). The depletion of L1 SNPs in the 5’UTR promoter region indicates purifying selection on the level of the host, where new SNPs that emerge on TFB sites get weeded out. The maintenance of L1 TFB sites by the host might be due to their generally beneficial role as promoters and enhancers of neighboring host genes (Garcia-Perez, Widmann, & Adams, 2018). In fact, almost 25% of proximal promoter regions as well as the 5’ and 3’ untranslated regions in the human genome contain TE-derived sequences, most of which are clustered around rapidly evolving lineage-specific genes (Jordan et al., 2003). Thus, the conservation of the L1 TFB sites by the host indirectly supports the hypothesis that TEs are capable of donating working regulatory sequences to host genes (Mariño-Ramírez et al., 2005).

The overall evolutionary dynamic that governs the host-transposon relationship remains elusive. Although the “arms race” hypothesis is the prevailing explanation for the nature of this relationship (Jacobs et al., 2014), a more recent view proposes that it is an ambivalent compromise (Castro-Diaz et al., 2015). The term arms race implies that every aspect that is beneficial for one party, should be deleterious for the other. The evolutionary trend among L1s, across different loci, towards lower retrotransposition activity suggests that host-level selection opposes transposon-level selection, which is consistent with the arms race hypothesis. However, the apparent maintenance of the L1 TFB sites by the host as indicated by SNP depletion implies a more nuanced view on the host-transposon relationship. The reduced level of L1 TFB site-polymorphism in the population, along with the conservation of these sites across different ancestral L1s (Sun et al., 2018) shows that they are important for both parties (i.e. the host and the transposon). This observation provides further evidence that the relationship between the L1s and their host genome might not be entirely antagonistic.

REFERENCES

Beck, C. R., Collier, P., Macfarlane, C., Malig, M., Kidd, J. M., Eichler, E. E., … Moran, J. V. (2010). LINE-1 retrotransposition activity in human genomes. *Cell*, *141*(7), 1159–1170. https://doi.org/10.1016/j.cell.2010.05.021

Peter Beerli, Comparison of Bayesian and maximum-likelihood inference of population genetic parameters, *Bioinformatics*, Volume 22, Issue 3, 1 February 2006, Pages 341–345, https://doi.org/10.1093/bioinformatics/bti803

Boissinot, S., Chevret, P., & Furano, A. V. (2000). L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Molecular Biology and Evolution*, *17*(6), 915–928. https://doi.org/10.1093/oxfordjournals.molbev.a026372

Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Morant, J. V., & Kazazian, H. H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(9), 5280–5285. https://doi.org/10.1073/pnas.0831042100

Bundo, M., Toyoshima, M., Okada, Y., Akamatsu, W., Ueda, J., Nemoto-Miyauchi, T., … Iwamoto, K. (2014). Increased L1 retrotransposition in the neuronal genome in schizophrenia. *Neuron*, *81*(2), 306–313. https://doi.org/10.1016/j.neuron.2013.10.053

Burns, K. H., & Boeke, J. D. (2013). Human Transposon Tectonics. *Cell, 149*(4), 740–752. https://doi.org/10.1016/j.cell.2012.04.019.Human

Castro-Diaz, N., Friedli, M., & Trono, D. (2015). Drawing a fine line on endogenous retroelement activity. *Mobile genetic elements*, *5*(1), 1–6. https://doi.org/10.1080/2159256X.2015.1006109

Dai, L., LaCava, J., Taylor, M. S., & Boeke, J. D. (2014). Expression and detection of LINE-1 ORF-encoded proteins. *Mobile Genetic Elements*, *4*(3), e29319. https://doi.org/10.4161/mge.29319

Del Carmen Seleme, M., Vetter, M. R., Cordaux, R., Bastone, L., Batzer, M. A., & Kazazian, H. H. (2006). Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(17), 6611–6616. https://doi.org/10.1073/pnas.0601324103

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

FitzJohn, R. G. (2012). “Diversitree: Comparative Phylogenetic Analyses of Diversification in R.” *Methods in Ecology and Evolution*, in press. doi: 10.1111/j.2041-210X.2012.00234.x.

Garcia-Perez, J. L., Widmann, T. J., & Adams, I. R. (2016). The impact of transposable elements on mammalian development. *Development (Cambridge, England)*, *143*(22), 4101–4114. https://doi.org/10.1242/dev.132639

Hancks, D. C., & Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA*, *7*(1). https://doi.org/10.1186/s13100-016-0065-9

Jachowicz, J. W., Bing, X., Pontabry, J., Bošković, A., Rando, O. J., & Torres-Padilla, M. E. (2017). LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nature Genetics*, *49*(10), 1502–1510. https://doi.org/10.1038/ng.3945

Jacobs, F. M. J., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A. D., Katzman, S.,

… Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*, *516*(7530), 242–245. https://doi.org/10.1038/nature13760

Jordan, I. K., Rogozin, I. B., Glazko, G. V., & Koonin, E. V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in genetics : TIG*, *19*(2), 68–72. https://doi.org/10.1016/s0168-9525(02)00006-9

Jurka J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(5), 1872–1877. https://doi.org/10.1073/pnas.94.5.1872

Kano, H., Godoy, I., Courtney, C., Vetter, M. R., Gerton, G. L., Ostertag, E. M., & Kazazian, H. H. (2009). L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes and Development*, *23*(11), 1303–1312. https://doi.org/10.1101/gad.1803909

Kass, R., & Raftery, A. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution*, *33*(7), 1870–1874. https://doi.org/10.1093/molbev/msw054

Kurnosov, A. A., Ustyugova, S. V., Nazarov, V. I., Minervina, A. A., Komkov, A. Y., Shugay, M., … Lebedev, Y. B. (2015). The evidence for increased L1 activity in the site of human adult brain neurogenesis. *PLoS ONE*, *10*(2), 1–14. https://doi.org/10.1371/journal.pone.0117854

Le Rouzic, A., Boutin, T. S., & Capy, P. (2007). Long-term evolution of transposable elements. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(49), 19375–19380. https://doi.org/10.1073/pnas.0705238104

Mariño-Ramírez, L., Lewis, K. C., Landsman, D., & Jordan, I. K. (2005). Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenetic and genome research*, *110*(1-4), 333–341. https://doi.org/10.1159/000084965

Meade, A., & Pagel, M. (2016). *BayesTraits V3*. (November), 81. https://doi.org/10.1016/S0022-3913(12)00047-9

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. https://doi.org/https://doi.org/10.1016/j.jmp.2015.11.001

Nekrutenko, A., & Li, W. H. (2001). Transposable elements are found in a large number of human protein-coding genes. *Trends in genetics : TIG*, *17*(11), 619–621. https://doi.org/10.1016/s0168-9525(01)02445-3

Ostertag, E. M., Prak, E. T., DeBerardinis, R. J., Moran, J. V., & Kazazian, H. H., Jr (2000). Determination of L1 retrotransposition kinetics in cultured cells. *Nucleic acids research*, *28*(6), 1418–1423. https://doi.org/10.1093/nar/28.6.1418

Rambaut A. 2009. FigTree version 1.3.1 [computer program] http://tree.bio.ed.ac.uk .

Rangasamy, D., Lenka, N., Ohms, S., Dahlstrom, J. E., Blackburn, A. C., & Board, P. G. (2015). Activation of LINE-1 Retrotransposon Increases the Risk of Epithelial-Mesenchymal Transition and Metastasis in Epithelial Cancer. *Current molecular medicine*, *15*(7), 588–597. https://doi.org/10.2174/1566524015666150831130827

Rangwala, S. H., & Kazazian, H. H., Jr (2009). The L1 retrotransposition assay: a retrospective and toolkit. *Methods (San Diego, Calif.)*, *49*(3), 219–226. https://doi.org/10.1016/j.ymeth.2009.04.012

Revell, L. J. (2012) phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*, **3**, 217-223.

Ricci, M., Peona, V., Guichard, E., Taccioli, C., & Boattini, A. (2018). Transposable Elements Activity is Positively Related to Rate of Speciation in Mammals. *Journal of Molecular Evolution*, *86*(5), 303–310. https://doi.org/10.1007/s00239-018-9847-7

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., … Huelsenbeck, J. P. (2012). Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542. https://doi.org/10.1093/sysbio/sys029

Sinzelle, L., Izsvák, Z., & Ivics, Z. (2009). Molecular domestication of transposable elements: From detrimental parasites to useful host genes. *Cellular and Molecular Life Sciences : CMLS*, *66*, 1073–1093. https://doi.org/10.1007/s00018-009-8376-3

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., ... & Konkel, M. K. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75-81.

Sun, X., Wang, X., Tang, Z., Grivainis, M., Kahler, D., Yun, C., … Boeke, J. D. (2018). Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(24), E5526–E5535. https://doi.org/10.1073/pnas.1722565115

Yang, N., Zhang, L., Zhang, Y., & Kazazian, H. H., Jr (2003). An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic acids research*, *31*(16), 4929–4940. https://doi.org/10.1093/nar/gkg663