EVOLUTIONARY INFERENCE IN TRANSPOSABLE ELEMENTS

by

AARON C. WACHOLDER

B.S., Stanford University, 2010

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Ecology and Evolutionary Biology

2017

This thesis entitled:

Evolutionary Inference in Transposable Elements

written by Aaron C. Wacholder

has been approved for the Department of Ecology and Evolutionary Biology

_____

David W. Stock

_____

David D. Pollock

Date_____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Wacholder, Aaron C. (Ph.D., Ecology and Evolutionary Biology)

Evolutionary Inference in Transposable Elements

Thesis directed by Professor David D. Pollock

Transposable elements (TEs) are a large component of many eukaryotic genomes, and the evolution of TEs is closely connected to that of their hosts. Accurate inference of TE evolutionary relationships is essential to understanding the biology and evolution of TE families and the role they play in genome evolution. Additionally, the great quantity of TEs makes them a useful model system for understanding genomic processes such as mutation and recombination, and their utility as a research system also depends on accurate evolutionary inference.

In this dissertation, I describe novel computational methods for evolutionary inference in TEs, applying them primarily to the *Alu* family of primate retroelements. A major task in TE evolutionary study is the classification of elements of a family into subfamilies. I developed the AnTE algorithm, a Bayesian approach to subfamily classification that, in contrast to previous deterministic methods, allows for probabilistic subfamily classification, an important advance due to the high uncertainty involved. I use AnTE to provide a more complete picture of the evolutionary history of *Alu* elements than provided by previous analyses, especially regarding the role of gene conversion. This work suggests that current *Alu* subfamily classification found in widely-used databases such as RepeatMasker and RepBase provides a misleading account of *Alu* evolutionary relationships.

Building on the AnTE research, I developed a Bayesian phylogenetics approach to the detection and characterization of gene conversion events among

TEs in a genome. I use this approach to identify a burst of interlocus gene conversion among *Alu* elements in the gorilla genome, occurring at much higher rates than on any other branch of the Great Ape phylogeny. Abnormally high *Alu* gene conversion rates in gorilla appear to be driven by binding to *Alu* by PRDM9, a rapidly-evolving protein that targets DNA sequence motifs for double-strand breaks in meiosis. These findings indicate one evolutionary pathway for rapid gene conversion in a TE family, and the conversion events identified provide a rich dataset for understanding the dynamics of gene conversion in primates.

ACKNOWLEDGEMENTS

I would like to thank all who made this dissertation possible. First, I offer sincere thanks to my advisor, Dr. David D. Pollock, for giving me an opportunity to study under him. His support, ideas, and advice have contributed greatly to my development as a researcher. I also thank the members of the Pollock laboratory for being a continual source of valuable feedback. I thank the members of my dissertation committee for their help and guidance throughout my Ph.D. work.

I thank the Integrated Quantitative Biology Program, The Computational Biosciences Program at Anschutz Medical Campus, and the Ecology and Evolutionary Biology Department for providing support for my Ph.D. The work presented in this dissertation was funded by the National Institutes of Health (GM083127, GM097521593, 5T15LM009451).

CONTENTS

CHAPTER

TABLES

Table

FIGURES

Figures

CHAPTER I

INTRODUCTION

Transposable elements (TEs) are genomic sequences that can replicate and insert themselves elsewhere in the genome. Many TE families have been successful at replicating, such that large portions of many eukaryotic genomes are TE-derived[1–3]. At least two-third of the human genome appears to be repeat-derived[4], likely primarily from extinct TE families. As ubiquitous inhabitants of eukaryotic genomes, the evolution of TE families is intimately connected to that of their hosts[5,6]. On occasion, TEs can provide major benefits to their hosts[7]; most importantly, TEs can provide the raw material for new functional genes[8,9] and regulatory elements[10–12]. However, TEs can also have many deleterious effects. TE insertions into genes or regulatory elements can eliminate important functionality[13], and recombination between TE copies can also cause harmful genomic rearrangements[14]. TE insertions and TE-associated rearrangements in both somatic cells and the germline are associated with numerous human diseases.[13,15–17]

The study of TE evolution is important for at least two reasons. First, TEs are important genomic actors. As TEs are a large part of many genomes, to understand genome evolution, and genome function, we must understand TE

evolution and function. Second, TEs are a useful research tool even when they are not the main subject of interest. As it is unlikely that a similar TE inserts at the same genomic position multiple times, presence/absence of TE insertions are a strong phylogenetic marker[18–20]. TEs have also been commonly used to study neutral mutation process[21–23], as they are thought to typically evolve neutrally after insertion[24]. In general, TEs are useful for genome evolution research because they serve as a sort of natural experiment. A single active TE might produce numerous copies that insert throughout the genome. As these copies are initially identical, their subsequent evolution can be thought of as many replicates of an experiment differing only by genomic position. The large number of TEs provides a high degree of statistical power to test evolutionary hypotheses and fit complex models.

Despite the ubiquity and utility of TEs, much about their biology and evolution remains poorly understood, even for well-studied families. Early research on the major human TE families *Alu* and LINE1 was dominated by the master element model of TE evolution[25,26], which posits that a tiny number of hyperproductive elements produced all copies of its family in the genome. Though later research demonstrated that this model was incorrect[27,28], no alternative has been proposed to explain all that the master element model attempted to explain. We have only modest understanding, for example, of the proportion of elements that are replicative[29], the typical replicative lifetime of active elements, or the processes that cause succession between different subclasses of a TE family.

The aim of this work is to develop and utilize a toolkit for the evolutionary study of transposable elements. Starting from a set of TE sequences of a given family, I develop computational methods for aligning elements, classifying elements, dating elements, inferring element ancestry, and reconstructing gene conversion events between elements. Though methods exist for most of these tasks, the Bayesian phylogenetic framework I employ provides substantial benefits in accuracy, precision, and conceptual coherency, which are necessary for developing a clear picture of TE evolution. After developing this TE evolutionary toolkit, I use it to investigate TE evolutionary dynamics and the dynamics of nonallelic gene conversion.

My major study subject is the *Alu* family[30], a primate-specific SINE retrotransposon and the most common TE in the human genome. *Alu* elements were derived from a duplication of the ribosomal 7SL RNA gene early in primate evolution approximately 65 MYA[31]. At around 300 bp in length, *Alu* does not code for its own replication but instead relies on the replication machinery of the much larger LINE-1 retroelement[32,33]. *Alu* elements are transcribed by RNA polymerase III, then reverse-transcribed and integrated into the genome by two proteins encoded by LINE-1.[33] *Alu* has been extraordinarily successful in the primate genome, with over 1 million copies that comprise over 10% of the human genome overall[30]. Though most *Alu* and LINE-1 copies are inactive in the human genome, active copies of both remain[28], and *Alu* recombination activity and insertion activity are each associated with human diseases[15,16]. Importantly, large numbers of *Alu*

elements retain their full length, which facilitates evolutionary inference. *Alu* elements appear to engage in high rates of nonallelic gene conversion[34,35] and homologous recombination[14,36], and are a useful model system for understanding these processes in primates.

CHAPTER II

THE MEANING OF THE SUBFAMILY CONCEPT IN TRANSPOSABLE ELEMENT EVOLUTIONARY STUDIES

Given a sequenced genome, the first step to TE analysis is family classification: to identify the TEs in the genome and assign them into families. There are numerous TE family classification methods[4,37] and I do not attempt to improve upon them here. The second step is subclassifcation: given a set of TE sequences from a particular family, divide them into smaller subclasses, typically called "subfamilies" to facilitate further analysis. Subfamily classification is the starting point of most analyses involving TEs, and such classifications are often obtained from the databases RepBase[38] and RepeatMasker[39], the latter of which forms the basis for the repeat track on the UCSC Genome Browser[40]. Despite its use in nearly all TE research, the subfamily concept is rarely defined explicitly. Neither RepBase nor RepeatMasker give a definition on their websites or in their supporting publications[38,39]; nor does the publication describing the popular CoSeg subfamily classification method, which is used by RepeatMasker[27].

The earliest reports on *Alu* subfamily division, by Willard et al.[41], Britten et al.[42] and Jurka et al.[43] also contain no explicit definition of "subfamily." These authors performed subfamily classification by identifying sets of nucleotide differences from the overall *Alu* consensus that were correlated with each other in their *Alu* sequence datasets. The identified nucleotides were called "diagnostic nucleotides" and subfamily consensus sequences were constructed using the set of sequences containing each set of correlated nucleotides. The elements within each subfamily were inferred to be derived from retrotransposition from a single source gene.[42,44] *Alu* subfamilies were given further evolutionary interpretation with the introduction of the master element model[25,26,45], which held that only one or a few "master" elements were responsible for most *Alu* insertion in the genome. In this model, as a master gene experiences mutations at diagnostic nucleotides during its evolution, it creates successive subfamilies, each new subfamily differing from the previous at the most recently mutated site. In this conception of "subfamily", many distinct *Alu* subfamilies share an ancestral locus.

What, then, are subfamilies? Most usage in the *Alu* literature is consistent with the idea that subfamilies consist of all elements that transposed from a common *sequence*, where *sequence* here means an ordered set of nucleotides, not a particular locus in the genome. That is, if two identical replicative loci both generate copies, the copies belong to the same subfamily because they were transposed from an identical sequence, but if a single replicative locus experiences mutation and remains replicative, the elements replicated from that locus before

and after that mutation belong to different subfamilies. This definition is consistent with the practice of distinguishing subfamilies by diagnostic nucleotides, as groups of elements copied from different sequences will systematically differ at all positions that distinguished the replicative sequences. It is also consistent with usage of the subfamily concept in the master element model, which holds that a single replicator can generate many subfamilies as it evolves.

An important implication of the sequence-based subfamily definition is that there is no necessary evolutionary relation between the elements in a subfamily, though in most cases we do expect the relationship to be close. It is tempting to say that a group of elements all copied from the same sequence were all descended from a common ancestor with that sequence. This need not be the case, however, because two distinct replicators, neither on which is ancestral to the other, could converge by mutation to an identical sequence. In this case, their descendants (while they had that sequence) would be classified together despite being more closely related to other elements outside of the class than to each other. This need not be a rare case; if replicative elements are highly similar to each other (which they often are; see Chapter V) then convergence to identity could be relatively common, especially if there are strong sequence constraints on replication such that active elements are restricted to a narrow region of sequence space[28]. A second implication is that there is no guarantee that the evolutionary relationships between subfamilies form a tree structure, despite common practice of assuming such a structure[27,30]. As subfamilies

may have multiple sources, and those sources may themselves have been descended from distinct other sources, a subfamily could have multiple ancestral subfamilies.

It is generally preferred that biological classifications be based on evolutionary relationships, and that taxonomic groups contain all elements descended from a common ancestor[46,47]. Why, then, should we classify elements by replication from a common sequence (ordered set of nucleotides) rather than a common ancestor, when the former does not imply any particular evolutionary relationship? We might consider an alternative approach. Suppose we have a set of TE sequences from a newly sequenced genome, and wish to infer their evolutionary history. We have methods to infer gene trees from a set of extant homologs[48]. We might propose to apply these methods to the sequences in a dataset of TEs, construct a tree, and perhaps identify useful monophyletic groups on the tree as "subfamilies." These subfamilies would, then, be the TE equivalent to species; assignment of an element to a subfamily would imply descent, along with all other subfamily members, from a common ancestor.

The problem with this approach is that a single replicative element can produce numerous identical copies, and those copies can produce further copies. Presented with a large quantity of elements that were identical at insertion, we have no sequence data to use to resolve relationships among these identical elements. This in itself is not a fatal flaw, as phylogenetic methods are capable of dealing with uncertainty[49]. But a tree of human *Alu* elements, for example, would contain around one million leaves[30], and much of its structure would be wasted on

representing the unresolvable relationships between elements that were identical at insertion. Such a tree would be extremely unwieldy for both visualization and analysis.

Subfamily classification, then, is a practical first step to TE analysis. We should group together those elements whose ancestry cannot be distinguished, because there is no benefit to considering them separately. If extant elements differ from the sequence they replicated from only by the accumulation of random mutations after insertion, then these post-insertion differences are not informative as to their origin, and elements copied from identical sequences can be safely grouped together at no cost to our ability to resolve their evolutionary relationships. The task of subfamily classification, in this perspective, is equivalent to the task of separating out informative vs. non-informative variants in each element.

Were transposable element evolution only as complicated as presented so far, imprecise definition of the subfamily concept would perhaps not lead to large errors in analysis. There is an additional complexity, however, which requires precision to navigate, at least in some TE families such as *Alu*. While retrotransposition is the primary means by which *Alu* elements replicate, gene conversion is an important secondary mechanism of *Alu* replication[50]. During a gene conversion event, a DNA strand at a donor locus serves as a template to replace a homologous sequence elsewhere in the genome[51]. Gene conversion occurs between alleles on homologous chromosomes during meiosis, but can also occur ectopically between non-allelic homologs as part of the double strand break repair process[52,53]. There is

considerable evidence that *Alu* elements engage in extensive non-allelic gene conversion.[34,35,54] Thus, while retrotransposition creates new copies of active elements in the genome, gene conversion can create partial or complete copies of elements in other already-inserted elements.

Let us reconsider, in the context of gene conversion, the definition of "subfamily" that I argued corresponds to typical usage of the concept: each subfamily contains all elements replicated from a given sequence (i.e., ordered set of nucleotides). Suppose an element is completely converted by another element. In that case, all information about the sequence that the locus was originally copied from is lost, making subfamily inference under this definition impossible (such complete conversion events have been identified in *Alu* using orthology data[50], but conversion can be inferred from orthology only in limited cases). Similarly, if only half an element is converted, we cannot distinguish using the sequence data of a single genome which half was present at insertion and which was donated. For a TE family evolving in the context of extensive gene conversion, it is not workable to define subfamilies on the basis of transposition from a common sequence. Given the limitations of our sequence data, classification of conversion products must be based on the sources of its component parts, with no distinction between the original transposition source and later invasions.

Existing subfamily classification methods essentially ignore gene conversion, arbitrarily assigning "hybrid" elements that appear to be derived from multiple distinct replicative sequences to subfamilies corresponding to only one of those

sequences (see Chapter V for examples). It is unclear what understanding of the "subfamily" concept justifies this. Much downstream analysis using TE subfamilies relies on the idea that each subfamily member was derived from the subfamily consensus[20,21,45], but this will be true of only part of the hybrid sequence. In any case, given the practical justification for subfamily classification to group together elements that differ only by noninformative variants, we should distinguish apparent hybrid elements from others, as their evolutionary origins can be distinguished. As shown in Chapter V, separating out possible hybrid elements is useful for understanding the evolutionary processes operating in a TE family.

How should we understand subfamilies in TE families characterized by high rates of gene conversion? From a practical perspective, subfamilies should group together elements differing only by site variants that are not informative as to their evolutionary origins. Non-informative variants are the result of random mutations that occur in the element after insertion. Elements should be grouped, then, by their *mutation-reversed* sequence: the sequence the element would have, if no part of the element experienced mutation after replication and insertion into the genome. For an element that never underwent gene conversion, mutation reversal simply produces the sequence it was copied from; these elements are thus classified the same way as in our previous definition. For converted segments of a hybrid element, mutation reversal produces the sequence the homologous segment in the conversion donor was copied from. Subfamilies group together all elements with

identical mutation-reversed sequences (Figure 1).

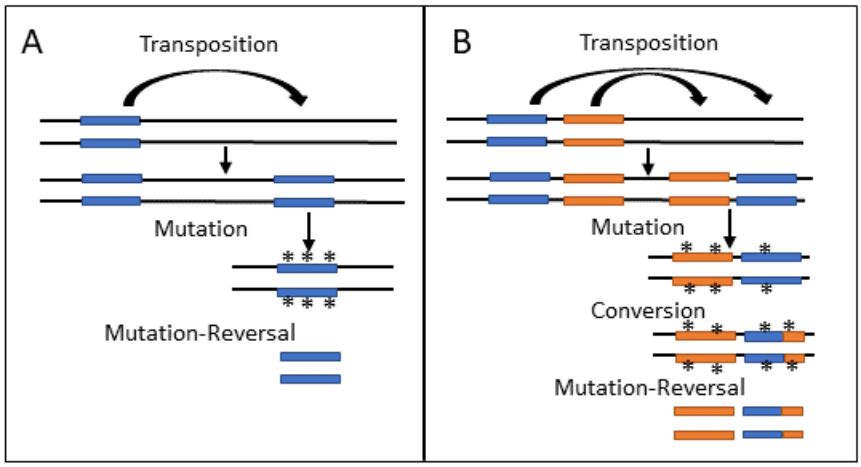

Figure 1: Mutation-Reversal of Transposable Elements.

A) The mutation-reversed sequence of an element that did not experience gene conversion is identical to the sequence it was originally copied from. B) The converted and unconverted components of the mutation-reversed sequence of a gene conversion acceptor are both identical to the homologous sequence of the replicative sequences they were derived from.

Visualization and analysis of TE evolutionary relationships is facilitated by grouping elements into subclasses as a first step in analysis. Ideally, we would group elements by common ancestry, but the limitations of TE sequence data make this approach infeasible in most cases. Instead, I suggest a pragmatic approach: as post-insertion mutations are not informative as to the origins of an element, it is sensible to group together elements that differ only by such mutations. In this way, we can reason collectively about the evolutionary history of elements in each class. Subfamily classification under this interpretation can be accomplished using a probabilistic mutation-reversal algorithm, the subject of Chapter III.

CHAPTER III

THE ANTE METHODOLOGY FOR TRANSPOSABLE ELEMENT ANCESTRAL INFERENCE AND MUTATION REVERSAL

AnTE was originally developed to perform the task of ancestral sequence reconstruction: identifying the replicative sequences in a TE family and probabilistically assigning each element to the sequence (i.e., an ordered set of nucleotides, not a specific locus) it was copied from[55]. This was a new approach to essentially the same task attempted by older TE subfamily classification algorithms such as CoSeg[27] and MASC[44]. It is still most straightforward to understand AnTE as a method for solving the ancestral sequence reconstruction problem, and for most of this chapter we will take this perspective. However, as discussed in Chapter II, it is not feasible to define subfamilies based on replication from a common sequence in a TE family characterized by extensive gene conversion; instead, elements should be classified into subfamilies based on their mutation-reversed sequences. We will see at the end of the chapter that the AnTE algorithm can be used to accomplish mutation-reversal without modifications, but this requires reinterpretation of the results.

Comparison to Previous Subfamily Classification Algorithms

The earliest TE subfamily classification algorithm was MASC[44], using an approach similar to hierarchical k-means clustering of element sequences. The CoSeg algorithm by Price et al.[27] was designed to address limitations in MASC and remains the most popular algorithm for TE subfamily classification. The CoSeg algorithm iteratively identifies sequences in a family or proposed subfamily that contains pairs of sites with nucleotide variants that co-occur more frequently than would be expected by random mutation from the subfamily consensus sequence. This pair of sites is then used to divide sequences into two new subfamilies, which may be further split by the same criteria, and so on to completion. The observation of overrepresented nucleotides at a pair of sites suggests that some sequences currently assigned to a subfamily were produced by a replicative sequence that diverged at these sites prior to replicating. This justifies introducing a new subfamily to contain the descendants of that replicator.

The CoSeg algorithm has two major limitations that motivated the development of an alternative method. First, previous work on SINE elements in human[56] and opposum[57] indicated that, after assignment to CoSeg-inferred subfamilies, positions in many subfamilies differed from the subfamily consensus more than expected from mutation alone. This is a problem for downstream

analysis because it leads to unrealistically high estimates of mutation rates at these positions if the subfamily consensus is assumed to be the ancestor of all elements in the subfamily, and also inflates subfamily age estimates based on molecular clocks. The high rates of variation at some sites could be caused by either unidentified progenitor sequences or gene conversion. A partial explanation for this problem in CoSeg is a rule in the algorithm that forbids using a site to split off a new subfamily if it has already been used earlier by the iterative algorithm, even if there is sufficient evidence of additional subfamily structure. This rule is intended to avoid creating subfamilies that are merely composed of gene conversion products, though there is no guarantee that it does not cause some subfamilies descended from transpositionally-active sequences to be missed as well. As I argued in Chapter II, it is preferable to accept that subfamilies can be composed entirely of gene conversion products, and one benefit of this approach is that it avoids the need for arbitrary rules that can lead to error in age and mutation rate inference.

An additional limitation of previous subfamily algorithms[27] is that they are all deterministic: each element is assigned to a single subfamily. While probabilistic inference is of broad utility in evolutionary study[49], it is especially important for TEs, because TEs are often very similar to each other and so there is often considerable uncertainty in ancestry-descendant relationships. Taking a Bayesian perspective, the AnTE algorithm gives the posterior probability each element is descended from each ancestral sequence (equivalently, the probability each element belongs to each subfamily).

The AnTE Approach

The goal of AnTE is to infer, given a set of extant TEs in the genome, the ancestral sequences that generated those elements. The core of AnTE is a model giving the likelihood of generating the extant TE sequence data given four sets of parameters: 1) a set of ancestral sequences 2) the productivity of each ancestral sequence 3) the time periods in which each ancestral sequence was active and 4) a substitution model describing how elements evolve after insertion. The AnTE algorithm uses Markov chain Monte Carlo (MCMC), a powerful technique for exploring many-dimensional parameter spaces[58], to draw sets of model parameters from the posterior distribution. AnTE output is a set of draws of parameter values from the posterior, from which we can estimate posterior distributions for parameters of interest.

Thus far, this is a standard Bayesian MCMC approach. However, there is one complication. I found that it is not feasible to explore all of sequence space for all possible ancestral sequences within the MCMC. A Markov chain in which there can be any number of ancestral sequences with any sequence does not mix efficiently and therefore produces inconsistent results. Instead, AnTE employs an iterative approach. It starts with a reasonable first guess at what the ancestral sequences are. Then, the Markov chain is run with a fixed set of ancestral sequences based on

this initial guess, while all other parameters are allowed to vary. On the basis of that run, the list of ancestors is refined, eliminating prospective ancestral sequences that do not appear to have actually been ancestral while adding new ones that appear to have support. Further iterations are then run, each time refining the list of ancestors. There are, then, three components to the AnTE algorithm: 1) generation of the initial list of candidate ancestors 2) the MCMC and 3) candidate list refinement. I describe each of these components in turn.

Generating an Initial Set of Candidate Ancestors

In the first published version of AnTE[55], we employed a top-down approach to initial candidate ancestor generation. Essentially, the idea of the top-down approach is to first identify the clearest divisions within the TE family, and then, in each individual subset, the clearest division within that subset, and so on. This iterative splitting approach can be used to generate an initial guess at subfamily structure, to then be refined based on MCMC results. It is conceptually similar to CoSeg[27], which also employs iterative splitting. However, we found the top-down approach to be slow when applied to a large dataset of *Alu* elements, as it can take many iterations to identify small subfamilies.

The most recent version of AnTE instead employs a bottom-up approach. For every element, the *N* most similar elements in the sequence dataset are identified.

The consensus of these N elements serves as a candidate ancestor. The idea underlying this approach is that, if an ancestral sequence produces many copies, those copies will vary at random due to mutations, but a majority of descendants will still have the ancestral nucleotide at all sites. Therefore, consensuses of highly similar sequences will tend to be ancestral sequences. Ancestral sequences that produced many fewer than N descendants will likely be missed by this method, and there may be some false positives, but this is acceptable given that our goal is only to produce an initial good guess as to the ancestral sequences, which will be refined later. I used $N=100$ for all analyses reported here. Using this approach, only three iterations are necessary to identify subfamily structure in *Alu*.

Parameter Estimation Using Markov Chain Monte Carlo

Given a set of candidate ancestral sequences, the ancestry model consists of three sets of parameters. A_j, the productivity of each candidate ancestor j, is proportional to the expected number of copies produced by the candidate and defined such that the sum of the vector is equal to 1. T_j, the time of activity of candidate ancestor j, describes when that candidate was replicative. I assume a single time point of replication (i.e., that all descendants of a given ancestral sequence are the same age). $\boldsymbol{Q}$ is a nucleotide substitution matrix giving the substitution rate between all pairs of nucleotides, distinguishing the hypermutable

C→T and G→A substitution rate at CpG sites from the non-CpG rate of those substitution types.

The likelihood of generating any TE sequence S_i in the sequence dataset $\boldsymbol{S}$, given all parameters, is defined as:

$$L(S_i \mid \boldsymbol{A},\boldsymbol{T},\boldsymbol{Q}) = \sum_{j=1}^{N_c} A_j * P(C_j \rightarrow S_i \mid T_j)$$

where N_c is the number of ancestral candidates, C_j is the *j*th candidate ancestral sequence, A_j is the productivity of candidate *j*, and $P(C_j \rightarrow S_i \mid T_j)$ is the probability of transitioning from sequence C_j to sequence S_i in time period T_j. This sequence transition probability is the product of the transition probabilities at each site between the base in C_j and the base in S_i at that site. The transition probabilities between each pair of nucleotides over time T_j are obtained from the matrix exponential e^{QT_j}. The overall likelihood of the data, $L(\boldsymbol{S} \mid \boldsymbol{A},\boldsymbol{T},\boldsymbol{Q})$, is the product of the likelihood of all sequences:

$$L(\boldsymbol{S}|\boldsymbol{A},\boldsymbol{T},\boldsymbol{Q}) = \prod_i L(S_i|\boldsymbol{A},\boldsymbol{T},\boldsymbol{Q})$$

Using this likelihood function, each set of parameters is sampled using Metropolis-Hastings proposals[59]. All parameters are given uniform priors.

To sample productivity vector $\boldsymbol{A}$, proposals are made in which a randomly-selected donor contributes a random value uniform between 0 and 0.00015 to a random recipient. The $\boldsymbol{T}$ parameters are sampled by two proposal types. In the

first, a candidate ancestor j is selected at random. A random integer n is drawn from 0 to 1000, and T_j is set to n/1000. In the second, candidate ancestors i and j are selected at random, and their associated parameters T_i and T_j are swapped. The substitution rate parameters are sampled by a single proposal, in which the current rate is added to a draw from a normal distribution with mean 0 and standard deviation 0.01. As all proposals are symmetric, the chain satisfies detailed balance[59] if the acceptance probability $A(x, x')$ for the moves from x to x' follows the Metropolis-Hastings acceptance proposal, where $p(x)$ is the likelihood of the set of all parameters x:

$$A(x, x') = \min\left(1, \frac{p(x')}{p(x)}\right)$$

Thus, using this acceptance probability, the stationary distribution of the Markov chain will be identical to the posterior probability distribution.

In each generation of the MCMC, one proposal is made for $\boldsymbol{A}$ and one for $\boldsymbol{T}$, while $\boldsymbol{Q}$ is sampled only once every 1000 rounds. I run the chain for 100,000,000 generations. The first 20,000,000 generations are burn-in and not used to estimate posteriors. After burn-in, I take one parameter set every 1,000 generations for a posterior sample. In each posterior sample, every element is randomly assigned to a candidate ancestral sequence based on its probability of descent from that ancestor given the sampled parameter set. Good mixing is verified by running three replicates with each replicate starting from a random parameter set, and

confirming that the within-replicate variance in likelihood was at least 99% of the overall variance.

Refining the Set of Candidate Ancestors

The next step is to refine the set of candidate ancestors on the basis of the MCMC results. This involves two tasks: first, removing candidates with weak evidence for being true ancestors; second, adding new candidates that appear to be plausible ancestors. A similar strategy is employed for both tasks. The basic idea is to compare the number of elements we expect to mutate towards a particular sequence type, given the parameters derived from the MCMC results, to the number observed with that sequence type. If the frequency of the type can be explained by mutation from descendants of other ancestors, there is no need to infer that the candidate ancestral sequence of that type was ever replicative and so it can be removed. Alternatively, if the frequency is greater than can be explained by mutation, then it justifies adding a candidate ancestral sequence of that type.

The following process is conducted separately for each draw d from the posterior. Each candidate, C_i, is considered for removal in turn. For every other candidate ancestor, $C_{j,j\neq i}$, AnTE estimates, based on the model parameter set, the probability a descendant of that candidate ancestor would mutate to the C_i site variant at all positions at which C_i and C_j differ. I refer to these as C_i-like elements,

as they look like C_i descendants despite not being descended from C_i. Based on the probability each ancestor would produce C_i-like elements, 10 simulations are performed according to the parameter set, counting the total number of C_i-like elements generated in each simulation. I call these counts $H_{d,r}$ for posterior draw d and simulation replicate r. AnTE then counts the number of C_i elements inferred to exist according to the element ancestry assignment associated with each parameter set. If the number of inferred C_i-like elements, plus inferred descendants of C_i (call this sum J_d), is small relative to the expected number of C_i-like elements produced by other ancestors ($H_{d,r}$), then we do not need C_i as an ancestor to explain the sequence data and it can be removed from the list of candidates. For each parameter set, we have ten $H_{d,r}$ counts and 10 J_d counts. If the average J_d value is smaller than at least 5% of the $H_{d,r}$ values, then the candidate is removed. Candidates with an average descendant count of less than 5 are also removed.

After removing candidates with weak support for being ancestral, AnTE then considers new candidates to add. AnTE considers every possible sequence, D_k, one nucleotide away from the post-removal set of ancestral sequences and not already in that set. As above, for every drawn parameter set, 10 simulations are run to generate counts of D_k-like elements we expect to arise by mutation, given the model parameters (label these counts $L_{d,r}$ for draw d and simulation replicate r). AnTE also infers, for every draw from the posterior, the count of D_k-like elements according to the element ancestry assignment associated with that draw (label these counts M_d). If the count of apparent D_i-like elements is too large to be

explained by mutation from ancestors within the model, then a new ancestor with sequence D_k is necessary to explain the sequence data, and so is added to the candidate list. Sequence D_k is added as a candidate if the average M_d value is greater than 1.3 times as large as 95% of the $L_{d,r}$ values, the average D_k is at least 5, and the difference between the average D_k and 95% of $L_{d,r}$ values is at least 10. I find that these thresholds produce reasonable results for the *Alu* dataset described in Chapter V.

Sequence Alignment

AnTE requires as input a dataset of elements in a multiple alignment. In my early AnTE work, with small datasets, I used manual alignment, as I found that common multiple-alignment methods produced errors that resulted in incorrect ancestral inference.[55] This is infeasible for large datasets. Therefore, I developed an alignment approach specifically for use with *Alu* elements, though largely based on previous work on probabilistic sequence alignment.[60] Probabilistic alignment avoids biases from deterministic alignment.[61]

In probabilistic alignment, rather than producing a single optimal alignment, the probability distribution of alignments is approximated by drawing a large number of alignments from this distribution. This allows quantification of alignment uncertainty. Ideally, probabilistic alignment integrates over gap and

scoring matrix parameters as well as the alignments themselves.[60] Due to the large number of sequences in *Alu*, full integration of these parameters was computationally infeasible, so scoring parameters were instead fixed based on estimates from an initial alignment. Probabilistic alignments were conducted using the probabilistic version of the Needleman-Wunch algorithm described by Zhu et al.[60], except with a fixed scoring matrix.

The initial alignment was constructed using a probability of 0.04 for all mismatches, 0.0045 for gap start and 0.009 for gap extension. Preliminary analysis suggested that these values produced reasonable alignments. A sequence T_0 was selected at random from the sequence data to serve as an initial template for alignment. All sequences were aligned to T_0. Then, at each site, the number of each nucleotide variant, insertion, and deletion among all the sequences in the database relative to T_0 were counted. A new template sequence, T_1, was then constructed from T_0 as a starting point. If a plurality of sequences had a particular difference from T_0 at any site, T_1 incorporated that difference. Thus, T_1 represents a consensus of all the *Alu* elements in the dataset. All sequences were then probabilistically aligned to T_1 using the same scoring matrix. A new scoring matrix was constructed from these alignments by setting the probability of any mismatch equal to the observed frequency of a mismatch between T_1 and each sequence in the data, using 1000 draws from the alignment distribution for each sequence. Similarly, gap start and extend probabilities were estimated from the observed frequency of gaps

initiation and continuation among all alignment draws for all sequences. All sequences were then realigned to T_1 using the new scoring matrix.

A full alignment can be constructed by randomly selecting one of the draws from the alignment distribution for each sequence in the data. Multiple alignments can be drawn and analyzed to estimate the effect of alignment draw on results of interest. A single multiple-alignment drawn from the posterior was used for all analyses presented in Chapter V.

Interpretation of AnTE Results

The straightforward interpretation of AnTE is that each ancestral sequence identified represents a replicative sequence; i.e., that there existed at some point in the history of the TE family one or more replicative elements with that exact sequence, and that those replicative elements produced perfect copies (by assumption) through transposition, which then mutated after insertion according to the inferred substitution matrix. The parameters of the model allow us to estimate how productive each ancestral sequence was (though not how that productivity was distributed among different identical replicative loci) and when it was active. In Chapter IV, I use this interpretation of AnTE to analyze the LAVA family of transposable elements, which does not appear to engage in high rates of gene conversion.[55]

In a TE family that does undergo high rates of gene conversion, such as *Alu*, AnTE results require alternative interpretation. The AnTE method identifies a set of sequences such that replication of those sequences followed by mutation of the copies can explain the sequence data. Should a sequence expand by gene conversion to a count greater than can be explained by mutation from descendants of other replicators, that sequence will be interpreted by AnTE as an ancestral sequence regardless of whether it was ever transpositionally-active. I argued in Chapter II that this is appropriate, as TE subfamily classification should separate out types that are products of gene conversion between descendants of different replicative sequences. In this case, the "ancestral sequence" represents not necessarily a transpositionally-active sequence, but a "mutation-reversed" sequence of all elements it is "ancestral" to. Note that this interpretation follows directly from the likelihood function used in the MCMC, which gives the probability of producing the sequence data by mutation of copies of the "ancestral sequences". The "ancestral sequences", then, are by construction the sequences their descendants would be if the post-insertion mutation process were reversed. Essentially, in the straightforward interpretation of AnTE, mutation-reversal is used as a means to identify replicators; in the alternative interpretation, mutation-reversal is itself the goal. The other model parameters, aside from the rate parameters in the substitution matrix $\boldsymbol{Q}$, require reinterpretation as well. The productivity vector $\boldsymbol{A}$ is reinterpreted as giving the expected count of elements that mutation-reverse to

each ancestral sequence. The activity time vector $\boldsymbol{T}$ is reinterpreted as giving the average age of elements that mutation-reverse to each ancestral sequence.

CHAPTER IV

INFERENCE OF LAVA ELEMENT ANCESTRY

LAVAs are a class of element found exclusively in gibbon (Hylobatidae) species, and are composed of portions of other TEs usually found in primate genomes: L1ME5, *Alu*Sz6, and SVA_A[62,63]. The LAVA elements are an attractive system for understanding the evolution of TEs because their recent origin (sometime after the Gibbon divergence from other hominids 15-18 million years ago) and limited diversification[63] make analysis of their relationships tractable. Using AnTE, I evaluated whether the likely number of replicating ancestral sequences in LAVA differed from the number of subfamilies returned by CoSeg[27], whether the subfamilies previously identified are compatible with predicted ancestral relationships, and whether AnTE solved the problem of unrealistically high implied mutation rates at some sites. Finally, I suggest new subfamily designations in the gibbon LAVA TE family based on their probable relationships. The AnTE methodology used to analyze LAVA is an earlier version than the improved method presented in Chapter III; the earlier methodology is fully described in Wacholder et al.[55].

Gibbon LAVA Sequence Filtering and Alignment

LAVA sequences in the Gibbon genome were identified using the probability-based oligonucleotide clustering method *P-clouds* [64]. The published LAVA consensus sequence, which contains only the region 3' of the VNTR [63], was segmented into regions which were used to form clouds. The genome was then searched for locations that matched the cloud data. Identified locations were merged if the distance between them was less than the length of the region in the consensus sequence. This resulted in 1136 sequences will full 3' regions. Sequence for the region 5' of the VNTR was obtained by building clouds from the region upstream of the VNTR in these sequences. Locations matching these clouds were then merged to the 5' sequences to obtain full-length sequences. This process identified 338 sequences with complete 5' regions. Alignments for both the 3' and 5' regions were constructed manually.

Identification of CoSeg Subfamilies and the Problem of Excess Mutations

The CoSeg algorithm was applied to 986 aligned LAVA elements (401 bp) to obtain 14 subfamilies. Some sites showed higher levels of divergence from the CoSeg-defined subfamily consensus sequences than might be expected due to

mutation alone, consistent with previous findings in *Alu*[56]. To determine the plausibility that the CoSeg subfamily consensus sequences represent all of the ancestral sequences of the TEs in the data, I developed a resampling test. Null expectations were obtained by resampling substitutions from the consensus sequence of each subfamily, accounting for variation in mutation rates among sites and mutation types. The substitution resampling process was replicated 1000 times to get a predicted distribution of each nucleotide at each site for each subfamily under the assumption that all differences between ancestors and descendants are due to mutation. The expected sums of deviations from these expectations were compared to the observed deviations from expectation among the real by-site nucleotide distributions in each CoSeg-inferred subfamily.

Applying this test to the LAVA CoSeg subfamilies, I found that in 12 of the 14 CoSeg subfamilies, deviation from expectations exceeded the deviation among any of the 1000 resampling replicates (Figure 2). Thus, we can reject the hypothesis that the sequence data can be explained solely by substitutions from the subfamily consensuses, and infer that there are likely to be many more ancestral sequences

than identified by CoSeg.

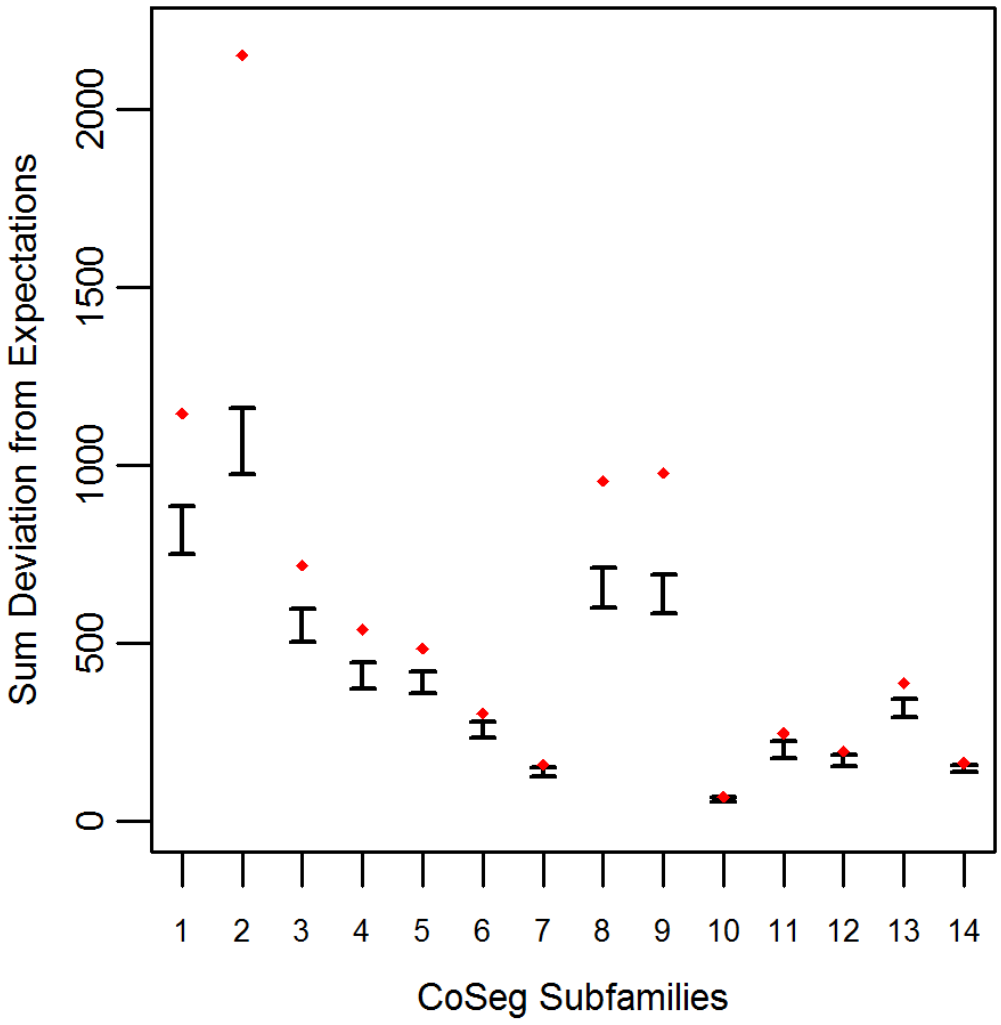

Figure 2*:* Deviation from Expectation in Randomly Sampled CoSeg Subfamilies. For each CoSeg subfamily, the 99% confidence interval is given for the deviation from expectations among 1000 substitution redraws under the hypothesis that all differences between subfamily members and the subfamily consensus are due to mutation, rather than replication. Diamonds indicate the deviation from expectation in the observed substitution data.

Support for a Large Number of Replicative LAVA Sequences

Separate Markov chains were run on LAVA for five different prior distributions of the total number of replicative sequences, set by applying a penalty on each additional ancestor inferred by the model. These penalties consisted of 0, 2, 4, 6, or 8 log points per ancestor. In LAVA, 38-43 (99% credible region) replicative sequences were inferred even under the harsh 8 log penalty, many more than the 14 subfamilies identified by the CoSeg program (Table 1 and Figure 3).

Prior penalty (log)	Number replicative LAVA sequences (99% credible region)	Mutation-only hypothesis p-value
0	60-72	.090
2	50-60	.064
4	44-52	<.001
6	41-47	.004
8	38-43	<.001

Table 1: Number of Replicative Sequences Identified for Different Prior Penalties in LAVA

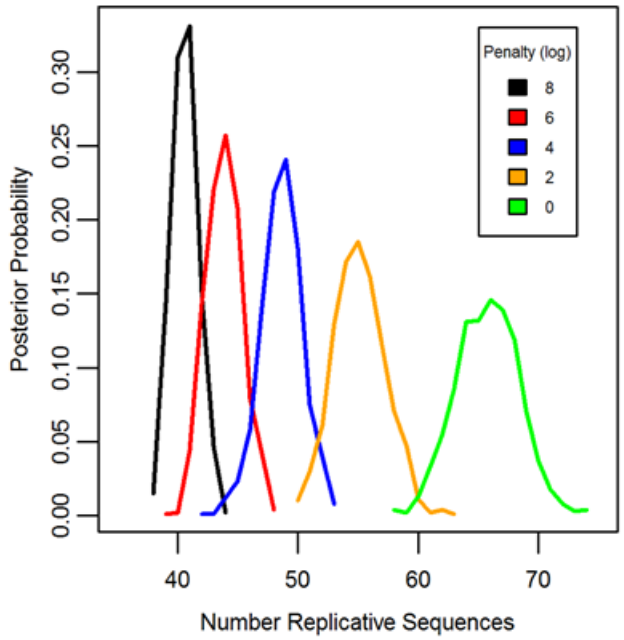

Figure 3: Posterior Distribution of the Number of Replicative Sequences.

Posterior distribution of the number of replicative sequences in LAVA for MCMC runs with different penalties applied to each additional replicative sequence. Higher penalties indicate a prior distribution favoring fewer replicative sequences. Each distribution is an average over 10 replicates.

The same substitution resampling method applied to the CoSeg subfamilies above was applied to the results from each AnTE run, testing whether mutation alone can explain the differences between inferred ancestral sequences and their descendants (Table 1). Based on this analysis, we reject the mutation-only hypothesis for the LAVA runs with 8 ($p<0.001$), 6 ($p=0.004$), or 4 ($p<0.001$) log penalty, inferring that these runs fail to identify some true ancestral sequences. We fail to reject the mutation-only hypothesis for the 2 log penalty run ($p=0.064$) and the 0 log penalty run ($p=0.090$). Thus, we select the results from the 2 log penalty chain as a conservative estimate of the number of replicative sequences in the

history of LAVA, and use it in all further analyses of LAVA. The 99% credible region for the number of replicative elements in the 2 log penalty run is 50-60, suggesting 50 as a reasonable lower bound for the total number of replicative sequences.

A Bushy Network of Related Ancestral Sequences

Network representations of the relationships among the elements of the LAVA families are shown in Figures 4-5. These networks show the predicted ancestral relationships among all sequences with more than 50% probability of being replicative (shown most clearly in Figure 4a). The arrows on the graph indicate the predicted original source of each replicative sequence, with cycles representing uncertainty about the direction of original descendancy. Note that later copies of that sequence may have arisen from other ancestors, including possible back mutation from one of its descendants. Each node in the graph represents a particular sequence, with the diameter of the node proportional to its estimated frequency of replication.

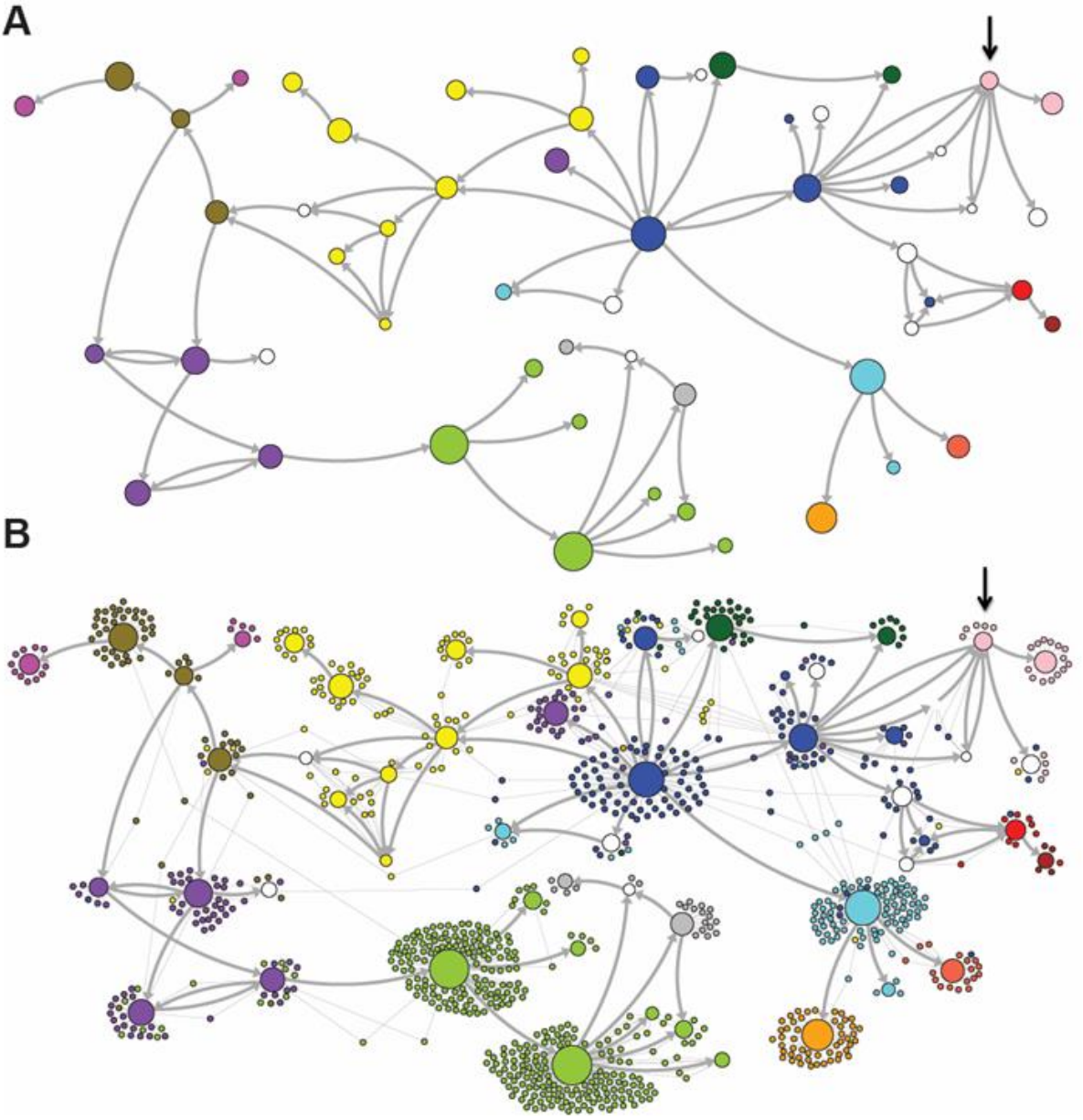

Figure 4: Ancestral Relationships Among LAVA Elements.

The predicted network of LAVA ancestral relationships is shown. A) All sequences that replicated with probability >30% are represented as nodes in the network. Arrows are drawn between sequences if there was at least 5% probability that an ancestral relationship existed between those sequences, with the direction of the ancestor-descendant relationships indicated by the arrows. Sequences are colored based on their CoSeg subfamily assignments (Figure 6). Sequences colored white do not exist in the data, but are inferred to have existed ancestrally. B) The network in A is modified by the addition of all extant TEs in the data added to the network as nodes represented by small dots. Edges are drawn between an element and an ancestral sequence if there was at least 5% probability

the element descended from the ancestral sequence. Nodes are colored based on CoSeg subfamily assignment.

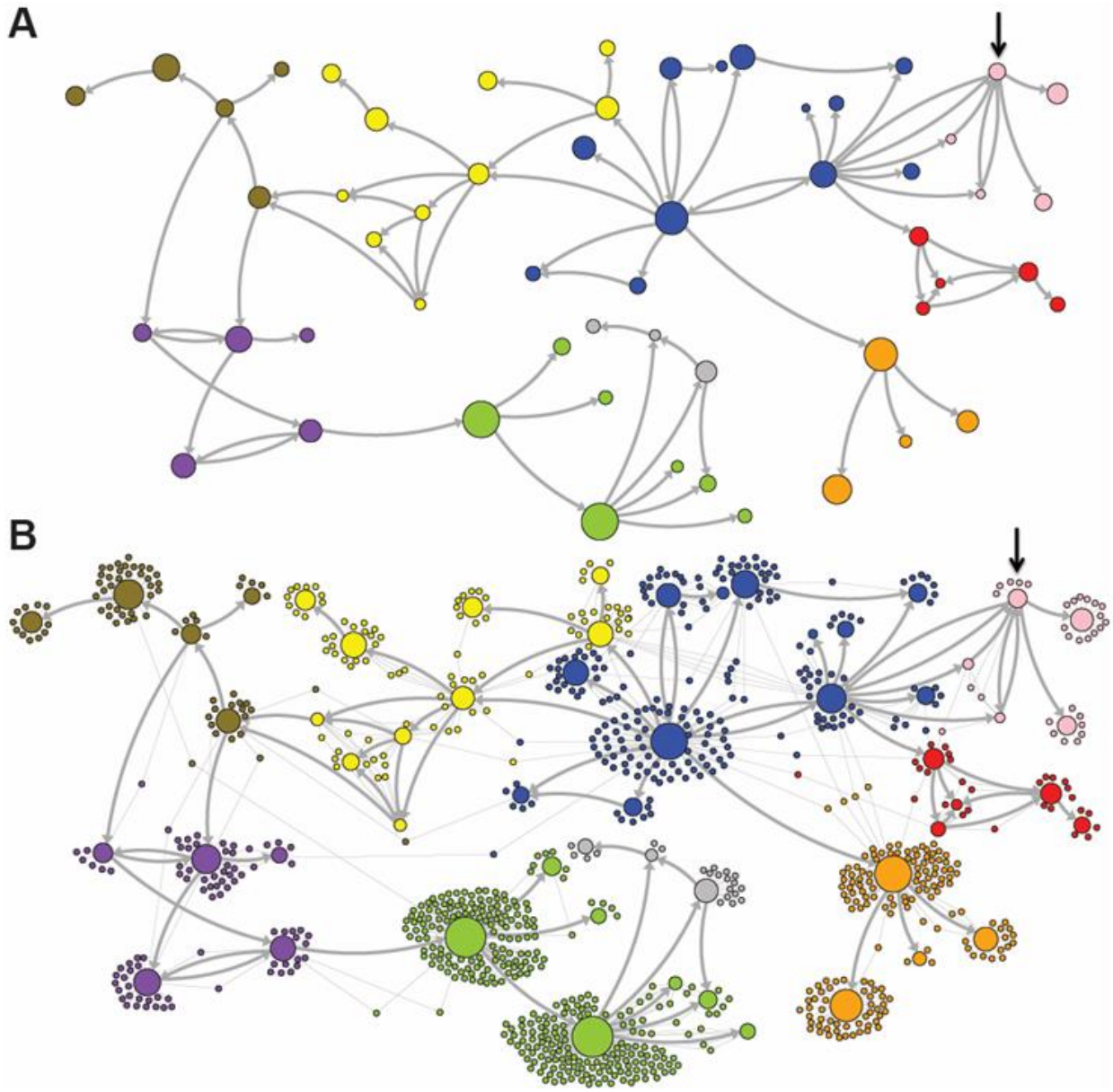

Figure 5: New AnTE Subfamily Assignments for LAVA Elements

The predicted network of LAVA TE ancestral relationships is shown, as in Figure 4. A) All sequences that replicated with probability >30% are represented as nodes in the network, exactly as in Figure 4A except that nodes are colored based on their new AnTE-based subfamily assignments. B) As in Figure 4B, all TEs in the data are added to the network as nodes, represented by small dots, and using the coloring scheme of the new AnTE-based subfamily assignments.

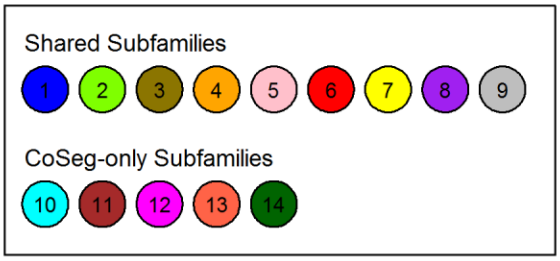

Figure 6: Subfamily Color Legend.

Subfamilies as defined by CoSeg are shown divided into two groups: those that correspond to a new AnTE subfamily (shared subfamilies #1-9), and those that are not classified as AnTE subfamilies (ancestral CoSeg-only subfamilies #10-14). The subfamily colors correspond to coloration in Figure 4-5, and numbering corresponds to information in the tables.

There are four sequences inferred to have at least a 5% probability of being the LAVA root according to the AnTE algorithm. We compared these sequences to the segment of the human genome homologous to the 3' end of LAVA[63]. One of these four plausible root sequences (Figure 4 and 5, marked with an arrow) has only 2 differences from the human sequence among 73 discriminatory sites; among all other candidates with >50% probability of being replicative, there are 4-28 differences (mean 12.1). Thus, the marked sequence is the probable ancestral LAVA, and the inferred root from AnTE is consistent with the homology data.

Revised LAVA Subfamilies

The assignment of CoSeg subfamilies to nodes in the ancestry networks of LAVA (Figure 4) indicates that most CoSeg subfamilies are represented by multiple

ancestral replicative sequences. Although CoSeg subfamilies tend to cluster together in the network, replicative sequences from three LAVA subfamilies (colored in purple, magenta and light blue in the graph) are disjointed, with intervening replicative sequences from other subfamilies (or that are not assigned to a subfamily at all). Additional discrepancies can be found when considering the CoSeg subfamily assignments of all sequences, not just replicative sequences (Figure 4b). Among descendants of all ancestors with CoSeg subfamily assignment, 57 LAVA sequences (6.5%) are assigned to different subfamilies than their most probable ancestor.

Based on this result, and considering the ancestral relationships inferred by the AnTE MCMC, I propose a subfamily organization for LAVA with 9 new subfamilies (Figure 5; see Figure 6 for legend). This subfamily scheme was designed based on the desiderata of a) relatively few subfamilies; 2) matching the CoSeg subfamilies where possible, to facilitate comparison; and 3) minimizing the number of sequences with uncertain subfamily assignment. The low mixing of colors in Figure 4b indicates that these goals have largely been achieved, although there is unavoidable uncertainty at most boundaries between subfamily groups. I emphasize here that the utility of the subfamilies is entirely organizational and aesthetic. I recommend that any analytical inference be carried out on the full ancestral probability network, and that it should sum over all ancestral uncertainty rather than arbitrarily assigning uncertain sequences to one ancestor or another and subsequently treating the assignment as though it were data.

Analysis of 5' Region of LAVA

The LAVA sequence is divided by a VNTR (variable number of tandem repeats) region of up to 2000 bp. My main analysis focused on the region 3' from the VNTR, as many LAVA loci lack all or part of the VNTR and 5' region. The full-length 5' region is around 350 bp, and I found 337 loci with intact 5' regions. Analysis of these sequences revealed three separate clusters defined by presence or absence of two large interior segments of around 100 bp each. I used AnTE to reconstruct the ancestral relationships separately within each of these three clusters. These ancestral networks largely agree with the analysis of the 3' region: the first cluster consists mostly of sequences from the adjacent green, purple, and brown subfamilies from Figure 5 (Figure 7A); the second cluster consists mostly of green and grey subfamilies (Figure 7B), and the third cluster is composed mostly of the older red, yellow, pink, and blue subfamilies (Figure 7C). However, 26 sequences (7.7%) are assigned ancestors on the 5' network that are distantly related to ancestors in the 3' network. A probable explanation for this discrepancy in placement between the 3' and 5' ancestral networks is recombination across the VNTR. Aside from these putative recombinants, the network structure within the three 5' clusters is largely in agreement with the structure of the 3' network (compare Figure 5 and Figure 7).

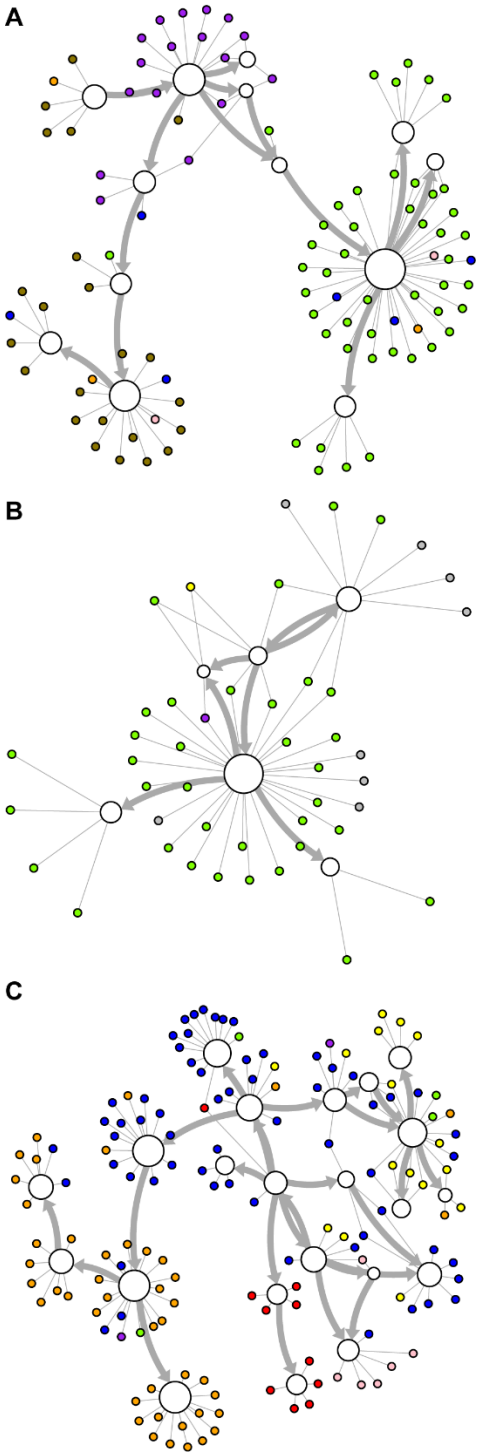

Figure 7: LAVA Ancestry Network Based on 5' Region

The predicted network of LAVA ancestry relationships, as described in Figure 4, but based on the region 5' of the VNTR rather than the 3' region. A) Cluster 1 network B) Cluster 2 network C) Cluster 3 network. Colors of sequences are based on the subfamily assignments shown in Figure 5.

Conclusion

We have confirmed here that the CoSeg subfamily classification method fails to identify many highly-probable ancestral sequences in LAVA and *Alu*Sc, and therefore that CoSeg subfamily consensus sequences are problematic for use as presumed ancestors in divergence and substitution analysis. In contrast, the AnTE method provides a detailed picture of TE evolutionary history, providing ancestral sequences, the times of replicative activity of these sequences, and their replication frequency. The AnTE method enables the probabilistic evaluation of relationships between thousands of elements within subfamilies and between subfamilies.

Despite the assumptions made in creating subfamilies using previous approaches, they have often been used in studies of TE evolution. For example, most methods for estimating the age of subfamilies are based on some measure of divergence between subfamily consensus sequences and the members of the subfamily[65–68]. The findings presented here suggest that this prior widespread use of subfamily consensus sequences as the single ancestral subfamily source sequence to analyze TE mutation patterns[69] has led to over-estimation of substitution rates and TE divergence times, and to incorrect inference of substitution patterns. AnTE can be used to improve such analyses, and may be useful to revise existing subfamily nomenclature based on more realistic estimates of ancestral replication patterns, as I have done with the gibbon LAVA elements. Overall, I expect that

such approaches will be central for evaluating genome structural evolution and using TEs to understand genome-wide mutation processes.

CHAPTER V

THE NETWORK STRUCTURE OF ALU EVOLUTION

I applied AnTE to a dataset of almost 146,865 *Alu* elements, identifying subfamilies defined by mutation-reversal to a particular ancestral sequence. I then analyzed these subfamilies to better understand how gene conversion and replication processes influence the sequence architecture of *Alu*. Previous work indicates that *Alu* engages in extensive gene conversion[70–72], with Roy et al. finding that gene conversion may be responsible for 10-20% of variation among recent *Alu* elements[70]. AnTE offers a systematic approach to identifying *Alu* subfamilies that may be explained by gene conversion, and allows for inference about the evolutionary mechanisms that may contribute to these subfamilies. Comparing our results to the RepeatMasker[73] classification of *Alu*, I find that current classification schemes, by ignoring gene conversion, produce fundamentally misleading results.

The Network Structure of *Alu* Diversity

A sequence dataset was constructed by filtering the *Alu* annotations of the hg38 assembly of the human genome from RepeatMasker[74]. Sequences outside of the 275-325 bp length range were excluded to obtain a set of only full-length *Alu*. A subset of this dataset was selected including all elements with an A at position 78 and a T at position 88; this corresponds mostly to elements in the *Alu*Y and *Alu*Sc subfamilies under traditional classifications. I use this subset of *Alu* because it consists of younger *Alu* elements, for which evolutionary analysis is more straightforward.

The AnTE algorithm identified 295 distinct *Alu* subfamilies, each associated with a particular ancestral sequence. The subfamilies ranged in expected frequency from 5 (the lower limit of detection) to 66,444 for the subfamily associated with the *AluY* consensus sequence. For visualization, the subfamilies were arranged in a network in which each subfamily is represented by a node (Figure 8) constructed by first drawing edges between all nodes representing subfamilies with ancestral sequences that differed by a single site variant. Additional edges were then drawn between all nodes that differed by two site variants but were not connected through a path of single-variant edges. This process was continued through higher numbers of variants until all nodes were connected directly by an edge or via a path along

multiple edges. Seventeen of the identified ancestral sequences are RepBase *Alu*S or *Alu*Y subfamily consensus sequences, as indicated on the network.

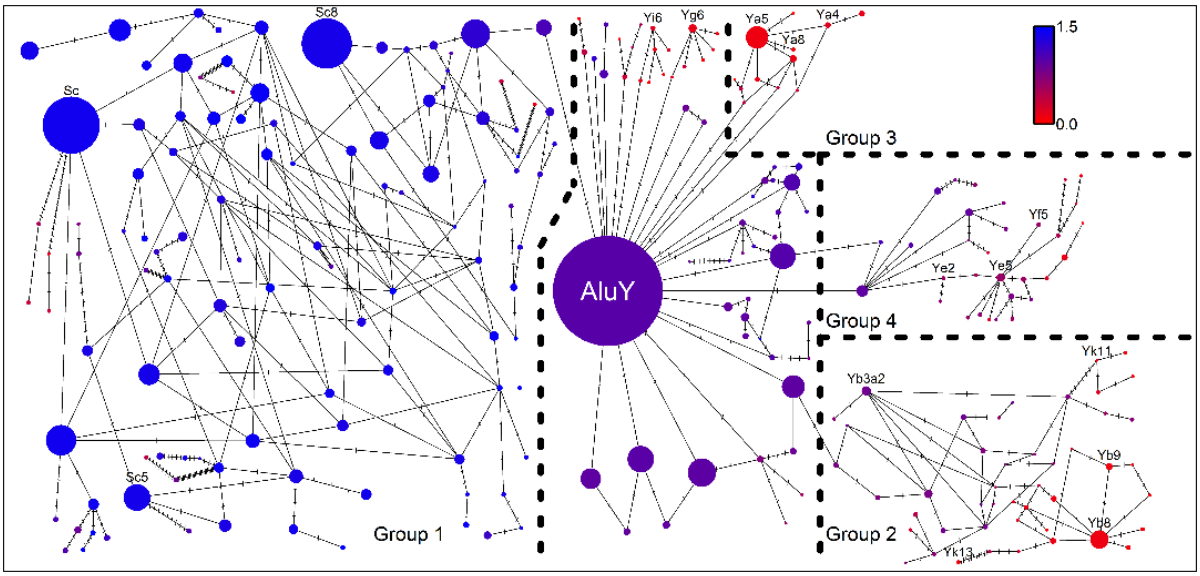

Figure 8: Network Representation of Ancestral Sequences.

Each of the 256 nodes in the network represents the inferred ancestral sequence of at least five *Alu* elements. Size of node is proportional to frequency of elements with that source sequence. Notches on each edge indicate the number of sites that differ between the sequences of the two nodes connected by that edge. If nodes have sequences identical to a RepBase consensus sequence, the *Alu*S or *Alu*Y RepBase identifier is noted. The estimated mean age of each node on the *Alu* network is indicated by color. Dashed lines separate four groups of nodes selected for individual analysis.

The highest-frequency predicted subfamily, by far, is the subfamily associated with the *Alu*Y consensus, which contains 37.4% of the elements in our dataset. I select four groups of subfamilies around the *Alu*Y consensus to facilitate analysis of the different regions in the network; each group is at least two variants distant from the *Alu*Y consensus, relatively well-connected internally and has few

external connections (Figure 8). In addition to the four selected groups, there are many nodes or small groups of nodes branching directly from the *Alu*Y consensus.

I defined the age of a node as the average age of the elements assigned to its represented subfamily by AnTE, with the age of each element estimated by its differences from the *Alu*Y consensus at invariant sites. Ages were scaled relative to the age of the *Alu*Y consensus node, which was set to age 1.0 (Figure 8). The oldest nodes are in Group 1, which has average node age 1.25 and includes all of the RepBase-annotated AluS sequences; it contains only a few nodes that appear to have replicated much later than the majority of elements in the region. Group 2 has an average age of 0.46 but shows a major division by age, with 20 of 46 nodes younger than 0.3 and 21 older than 0.6. We refer to the older subset, with average age 0.71, as Group 2A and the younger subset, with average age 0.14, as Group 2B. Group 3 is composed entirely of nodes younger than 0.4, with average 0.21, while Group 4 contains a wide range of ages, with average 0.61. It thus appears that Group 1 contains the ancestors to the other nodes in the network, while the other three groups were replicative either concurrent with and/or following the expansion of the *Alu*Y consensus sequence.

Visual inspection reveals an abundance of cycles across much of the network. Cycles in the network are suggestive of gene conversion, though they can also be explained by convergent or revertant mutations among transpositionally-active elements (Figure 9). We can estimate an approximate upper bound on the number of cycles expected to be generated by mutation, such that the remainder is most

plausibly explained by gene conversion. In the absence of sequence constraints on replication, we expect no more than 3.4 four-cycles (cycles consisting of four nodes) in the network, much smaller than the 76 four-cycles observed. Sequence constraints on replication increase the probability of cycles generated by mutation, because there are fewer options for mutations that retain replicability and therefore a greater probability of convergence among such mutations. If sequence constraints are such that only site variants present in source sequences are compatible with replication, then we expect no more than 13.2 four-cycles across the network, still much lower than the 76 observed. The interconnected network structure in *Alu* thus appears most plausibly explained as a result of extensive gene conversion.

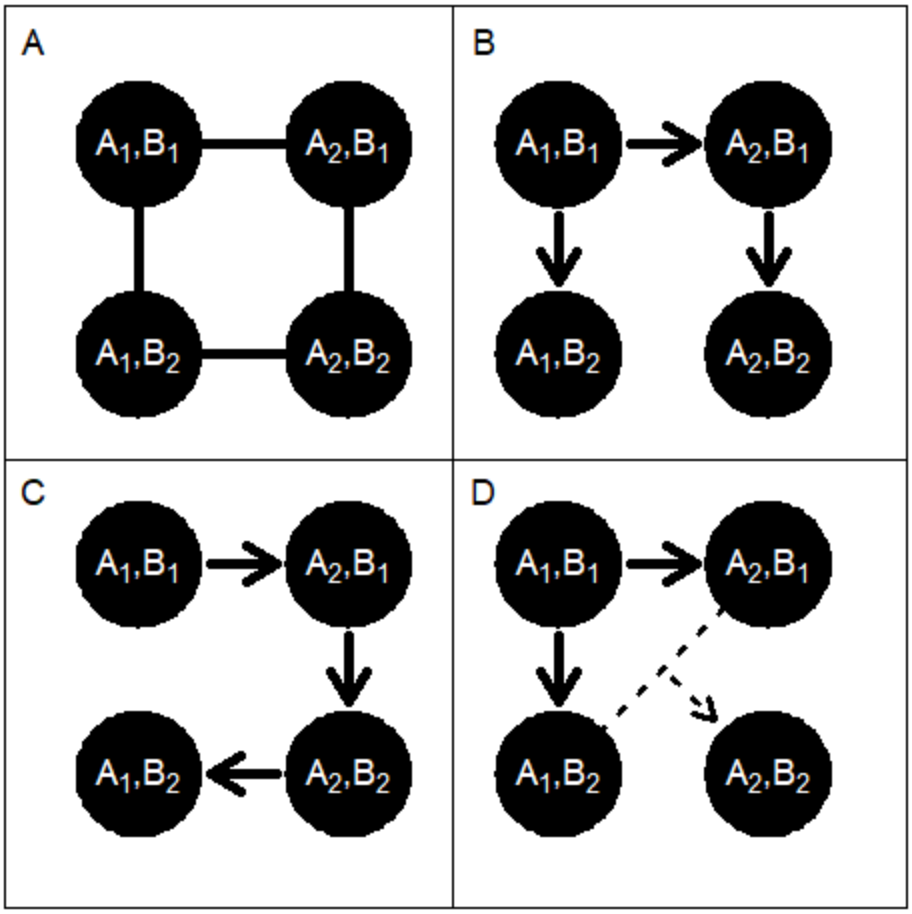

Figure 9: Cycle Generation in the TE Network.

The simplest type of cycle in the network consists of four nodes, each representing one possible combination of two sites with two variants each. Each node is labeled by its variant at variable positions *A* and *B*. As in Figure 8, edges are drawn between nodes that differ by a single variant. A) A generic representation of a four-cycle, differing at positions *A* and *B*. There are multiple mechanisms by which such a four cycle could be generated. B) Generation by convergence: two related replicative sequences both generate new replicative sequences by the same mutation at site B. C) Generation by reversion: Reversion of a mutation at site *A* creates a new replicative sequence that completes a cycle. D) Generation by gene conversion: conversion between two replicative sequences creates a new combination of variants, forming the fourth node in the cycle.

Classificiation of Nodes in the *Alu* Network

Ancestral sequences can represent either sequences that were transpositionally-active, or sequences that were never active, but were created through gene conversion. In general, determining whether ancestral sequences were ever replicative is a challenging problem. However, I expect highly productive transpositionally-active sequences to be present at high frequency, because all copies they produced that did not undergo gene conversion would mutation-reverse to that sequence. Following this reasoning, I divided the nodes in the network into three categories: major replicators, intermediate nodes, and minor replicators. The major replicators were distinguished as those nodes representing sequences that appear to have been transpositionally active due to being high-frequency relative to other nodes in their region. The intermediate nodes, located between major replicators in the network, represent sequences that contain alternative combination of site variants present in the major replicators of their group or the *Alu*Y consensus but no additional variants. These nodes represent potential gene conversion products between copies of the major replicators and *Alu*Y, although a fraction could also be derived by convergent or revertant mutations, as discussed above. Conversion products involving major replicators from different groups are also possible, but, as discussed below, I find little evidence for such products existing at sufficient frequency for identification. Third, minor replicators are nodes

that are relatively low in frequency, but represent ancestral sequences containing variants not present in the *Alu*Y consensus or the major replicators of their region and therefore cannot be explained by gene conversion events among those replicators.

Procedurally, I first classified the highest-frequency node in each group as a major replicator, and then classified as intermediates all other nodes containing only variants present in either the identified major replicator or *Alu*Y. I then selected the next most frequent node and classified it as a major replicator, provided it contained at least 4% of the total frequency in the group, and classified all elements containing only variants present in either major replicator or *Alu*Y as intermediates. I continued this process until the most frequent unclassified node in the group was at less than 4% of region frequency. Remaining nodes were classified as minor nodes. In Group 2, which contains two sets of nodes with disparate ages (Figure 8), I applied this process separately to Group 2A and 2B to identify the major replicators active in each period.

Overall, I classified 13 nodes as major replicators, 88 as intermediates, and 155 as minor (Figure 10, Table 2). Group 4 is characterized by a different pattern than the other regions. In Groups 1, 2, and 3, there are far more intermediate nodes than major replicators, a pattern suggestive of gene conversion, while in Group 4 there are similar numbers of intermediates and major replicators. Group 4 also shows little cyclic structure compared to the other regions. Group 4 is distinguished

from *Alu*Y by a 2 bp deletion at positions 265-266 in the *Alu* sequence, which may inhibit gene conversion with *Alu*Y.

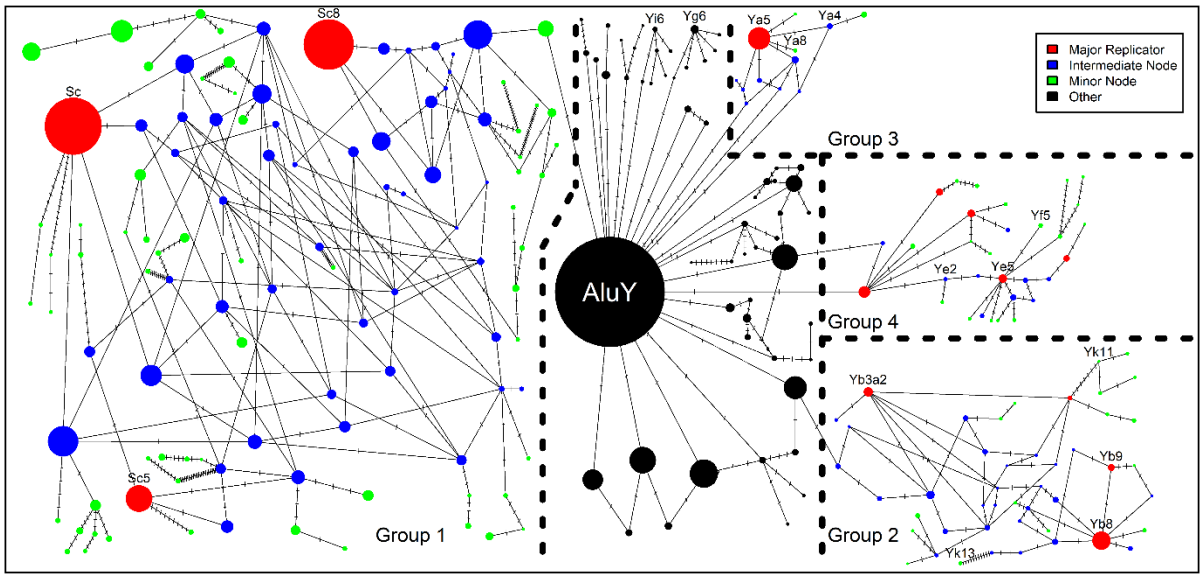

Figure 10: Classification of Subfamilies by Type

A proposed scheme for classifying subfamilies in the Alu network by type, reflecting the possible mechanisms by which each subfamily originated and grew.

Group	Node Count	Mean Age	Major Replicators	Intermediate Nodes	Minor Nodes
1	110	1.25	3	49	58
2	46	0.46	4	25	17
3	31	0.61	5	7	19
4	12	0.21	1	7	4
Full Network	256	0.87	13	88	155

Table 2: Properties of the Major Groups in the *Alu* Network

In Groups 1, 2, and 3, the predicted major replicators and intermediates largely follow the patterns that would be expected if the intermediates were primarily composed of conversion products of copies of the major replicators. As

there are many such possible conversion products I expect many intermediate nodes, and since converted products would be divided between these sequences, I expect each individual intermediate to be much lower in frequency than the major replicators themselves, provided that at least a substantial fraction of elements remain unconverted. Both these expectations are largely met (Table 2). In Group 1, for example, there are 49 intermediates, each a potential gene conversion product from only three major replicators, along with the *Alu*Y consensus. Two of the Group 1 major replicators, corresponding to RepBase consensus sequences *Alu*Sc and *Alu*Sc8, make up 21.7% and 16.3% of the total region frequency, respectively, both considerably above the most frequent intermediate at 5.8%. The third Group 1 major replicator, corresponding to RepBase consensus sequence *Alu*Sc5, is only at 4.5% of Group 1 frequency, less than some intermediates. However, if we consider only intermediates containing variants distinguishing *Alu*Sc5 from the other major replicators, the highest is at 1.7%, so *Alu*Sc5 appears much higher frequency than its own potential conversion products. With only one exception, no major replicator in the first three regions is less than 2.5 times as high frequency as any intermediate that shares a distinguishing variant with it. The one exception is the major replicator corresponding to RepBase consensus sequence *Alu*Yb3a2, which, at 8.7% of Region 2 frequency, is only moderately higher than an adjacent intermediate sequence with 6.7% of region frequency. I refer to this intermediate as *Alu*Yb3a2-249, as it differs from *Alu*Yb3a2 by a single variant at position 249. The

abnormally high frequency of *Alu*Yb3a2-249 may indicate that this intermediate had substantial transposition activity.

Though major replicators are generally much higher in frequency than individual intermediate nodes, intermediate nodes in aggregate make up a comparable proportion of frequency to the major replicators. Overall, major replicators make up 44.5% of regions 1-3 while intermediates make up 41.0%. There appears to be a substantial difference by age in the relative proportion of intermediates and major replicators. In Group 1, with average age 1.25, major replicators make up 42.6% of the group and intermediates 42.4%. Among the Group 2 nodes older than 0.5, 28.5% of frequency is made of major replicators and 61.6% of intermediates; even excluding the high-frequency intermediate AluYb3a2-249, intermediates make up 42.9% of frequency. In contrast, among Group 2 nodes younger than 0.5, major replicators make up 71.3% of the frequency and intermediates only 17.1%. Similarly, in Group 3, with average node age 0.21, major replicators make up 80.2% of the frequency and intermediates 14.3%. The tendency of younger groups to have much lower frequency among intermediates suggests that gene conversion products accumulate over long time periods.

While there are many intermediates between major replicators within Group 1-3, there are few nodes in the network that could be explained by gene conversion between copies of major replicators from different regions but not by conversion within groups. Such a node would contain, for example, variants present in Group 1 major replicators but not Group 2 major replicators, as well as the reverse, and only

contain variants present in at least one of the two groups. Although seven such nodes exist in the network, all seven contain only a single variant not present in the major replicators of their own group, suggesting that these may be the result of homoplasy rather than conversion. In any case, the total frequency of these nodes is only 0.2% of the full network, compared to 20% for within-region intermediates. It thus appears that gene conversion between copies of distant replicative sequences is rare.

The 98 minor nodes indicate the existence of numerous replicative sequences less productive than the major replicators, and widely spread throughout the network. Twelve of thirteen major replicators, and 30 of 88 intermediate nodes, are closer to at least one minor node than any other major replicator or intermediate. The minor nodes are generally younger than their closest intermediate or major replicator: in 51 of 54 cases where the 95% credible region for the age of the minor node is outside the 95% credible region of its closest intermediate or major replicator, the minor node is younger. This age pattern is expected if intermediate nodes are often ancestral to minor replicators, and the large number of minor replicators adjacent to intermediate nodes suggests that many intermediate node elements were replicators. In many cases, the average age gap between the minor nodes and their closest intermediate or major replicator is large. The mean age gap, 0.22, corresponds to 4.5 expected additional substitutions across the entire *Alu* element between their average times of ancestral activity. Similarly, there are an average 6.0 differences between minor nodes and their closest intermediate or

major replicator. This result suggests that, in many cases, conversion products became active long after the peak activity of the replicators from which they derived.

Explaining Intermediate Nodes in the Alu Network

Although intermediate nodes between the major replicators represent potential gene conversion products, gene conversion is not the only process than can explain intermediate nodes. One possibility is that some intermediate nodes represent points on a linear evolutionary trajectory between two major replicators. Potentially, as in the master element model of TE evolution[25], a single replicative locus could slowly evolve from one major replicator to another, producing numerous copies at every step along the way. Alternatively, a succession of distinct loci could be involved. In either case, the remaining intermediate nodes could be explained by gene conversion between the copies generated at each evolutionary step.

If some intermediate nodes are steps along an evolutionary trajectory, we expect a temporal gradient between nodes associated with the earliest and latest stages of the trajectory. Surprisingly, there appears to be no substantial age gradient across most of Group 1 (Figure 8). Region 1 major replicators are very close in age despite substantial differences in sequence. *Alu*Sc8 elements are 98.5% as old as *Alu*Sc elements (95% credible region: 97.9%-99.1%), corresponding to an expected

additional 0.4 substitutions across the entire *Alu* sequence for an *Alu*Sc element relative to an AluSc8 element. As *Alu*Sc and *Alu*Sc8 differ at 6 positions, for a single master replicator to be responsible for both *Alu*Sc and *Alu*Sc8 it would need to experience substitution at approximately 15 times the expected rate. Similarly, *Alu*Sc5 elements are 99.0% as old on average as *Alu*Sc elements (95% credible region: 97.8%-100.0%) and differs from *Alu*Sc at 6 positions and *Alu*Sc8 at 9 positions. The closeness in ages among descendants of these three major replicative sequences suggests that a distinct *Alu* locus was responsible for each sequence, that these loci were replicative at approximately the same time, and that none were descendant from each other. Intermediates are also largely similar in age to the major replicators, except for intermediates with *Alu*Y variants not present in the Region 1 major replicators. The mean age of intermediates with no *Alu*Y-specific variants is 1.39, as is the mean age of *Alu*Y major replicators. The mean age of intermediates with at least one *Alu*Y-specific variant is 1.31, as expected if these are conversion products with the younger *Alu*Y. Overall, the pattern of ages in Group 1 is not consistent with an evolutionary pathway between major replicators, but is consistent with extensive gene conversion between copies of concurrently active major replicators.

There is a temporal gradient between *Alu*Y and major replicators *Alu*Sc8 in Group 1, *Alu*Yb3a2 in Group 2, and *Alu*Ya5 in Group 3 (Figure 8). Such a gradient is consistent with an evolutionary trajectory between these master replicators and *Alu*Y, in which each step in the trajectory is associated with the production of copies

of the sequence at that step. However, we would also expect a temporal gradient if copies of two replicators that were active at different times experienced substantial gene conversion with each other. In such a scenario, conversion products would be a mix of segments of different age, with the average age of the converted element equal to the average of the segments weighted by segment length. Thus, we cannot determine whether there was an evolutionary pathway of replicative sequences between *Alu*Y and these major replicators or whether these nodes were created by gene conversion after the major replicator was already active. If there were such a pathway, it would involve 5 of 49 intermediates in Group 1, 3 of 25 in Group 2, and 2 of 7 in Group 1.

Evolutionary pathways link the major replicators, and sequences representing steps along those pathways may or may not have been replicative. Another potential source of replicative sequences, that could also potentially produce intermediate nodes, is branches off these pathways leading to minor replicators. A new replicative sequence can be created by mutation of a replicative locus that does not eliminate its transposition activity. Mutation to a new replicative sequence will generate a sequence associated with an intermediate node if the sequence after mutation contains only variants present in the region's major replicators or *Alu*Y.

I attempted to estimate an approximate upper bound on the number of intermediate node sequences that would be generated by mutation of an older replicative sequence. I first estimated an upper bound on the probability that a

random mutation to a new replicative sequence generates an intermediate sequence. I consider only mutations to variants present in at least one sequence in the network, as other variants may disable replicability. If the starting sequence is already an intermediate or major replicator, a single mutation will generate an intermediate or major replicative sequence if the mutation is to a site variant present only in the region major replicators or in *Alu*Y. I calculated the maximum possibility of such a mutation among all intermediates and major replicators in the region, p_{max}**.** If the starting sequence is not a major replicator or intermediate, then a single nucleotide mutation can only generate an intermediate if the starting sequence is one position away from an intermediate, and then only if the particular variable site experiences mutation. As this probability is always lower than the probability of an intermediate sequence remaining intermediate after mutation, we can ignore this case for the purpose of estimating an upper bound.

Given that the highest probability that mutation to a new replicative sequence results in an intermediate is p_{max}, we expect the number of intermediate sequences produced through this process to be lower than $p_{max} * N$, where N is the total number of new replicative sequences generated by mutation. The number of non-intermediates produced is greater than $(1 - p_{max}) * N$. The ratio between intermediates and non-intermediate replicative sequences generated by mutation will thus be less than $\frac{p_{max}}{1-p_{max}}$. Using this ratio, we can estimate an upper bound on the number of intermediates generated from mutation by the number of non-intermediates so generated. An estimate of the number of non-intermediate

replicative sequences generated by mutation is simply the count of minor replicators in each region. Using this reasoning, I estimated an upper bound on the number of intermediate replicative sequences generated from mutation from other replicators in each region. I estimated that fewer than 5.7 of 53 intermediates were generated by mutation in Group 1, fewer than 1.2 of 25 in Group 2, and fewer than 0.16 in Group 3. Thus, it appears that the process responsible for minor nodes is only a small contribution to the intermediate nodes.

Subtracting out the number of intermediate nodes that may be explainable as steps along an evolutionary trajectory between major nodes, or as mutations to new replicators not along such a trajectory, 42 intermediates in Group 1, 20 in Group 2, and 4 in Group 3, still require explanation. After considering the mechanisms that can generate intermediate node sequence without gene conversion, we are left with a majority of intermediate nodes to explain by gene conversion. There are at least two distinct explanations for intermediate nodes in which gene conversion plays a central role. The elements associated with an intermediate node could all be gene conversion products between direct copies of the major replicators. Alternatively, some conversion products could themselves be replicative and produce their own copies through transposition. We are unable to distinguish between these possibilities.

Explaining Frequencies of Intermediate Nodes

Though AnTE identifies many intermediate nodes, identified intermediates are only a small fraction of the possible intermediates between major replicators. Overall, there are 49,152 possible intermediates in groups 1-3, of which 91 are identified as ancestral sequences by AnTE. Some possible but unidentified intermediates likely represent ancestral sequences with frequencies too low to distinguish from mutation. Even among intermediates that are identified, there is great frequency variation: within Group 1, for example, there is 190-fold frequency difference between the most and least frequent intermediate subfamily. Frequency variation between intermediates can reflect either differences in the rate of conversion to each class or differences in transposition activity between intermediate replicative sequences.

If gene conversion is the main mechanism governing intermediate frequencies, then two predictions follow. First, I predict that intermediates that could be created as a result of a single conversion event between copies of major replicators will be more commonly identified as source sequences by AnTE, and be present at higher frequency, if identified, than intermediates requiring two or more conversion events. Second, I predict that, among intermediates that could be produced by a single conversion event, the number of possible conversion tracts generating that intermediate is positively associated with frequency.

There are 426 possible intermediates in groups 1-3 that would result from a single conversion event, of which 41 were identified, a rate of 9.6%. This is greatly enriched relative to the overall rate of 0.18% of possible intermediates identified, confirming the first prediction. The intermediates that could result from a single event are also much greater in frequency, having 2.3 times the frequency of other intermediates on average.

The number of possible conversion tracts between major replicator descendants that could create a particular sequence is also strongly related to its frequency. Among conversion products that could be created by a single event, identified source sequences have an average of 4474 tracts that could produce them, while unidentified sequences have an average of 1103 ($p<0.0001$ by two-sided T-test). Among identified source sequences, the number of possible tracts is correlated with frequency ($r=0.38$, $p=0.015$). There would be no reason to expect these associations, related to spatial patterns of site variants along the element, if intermediates were primarily the result of transpositionally-active sequences that transitioned from other active sequences by mutation.

Though these broad patterns indicate a strong role for gene conversion dynamics in shaping the frequency of intermediate classes, much remains unexplained, and I am unable to come up with a simple set of rules to predict intermediate frequency with precision. Frequencies of intermediate nodes reflect the probability of conversion between pairs of elements, the distributions of outcomes of those conversion events, and transposition activity of elements within

intermediate classes, and it is difficult to distinguish the influence of each process on frequencies.

Assessing RepeatMasker Annotation

Because gene conversion appears to be a major contributor to *Alu* sequence diversity, it is important to determine the extent to which gene conversion may lead to error in *Alu* annotation. I used the AnTE results to reassess RepeatMasker annotation. RepeatMasker assigns elements to subfamilies, each associated with a presumed ancestral subfamily consensus sequence. Ideally, all elements assigned to a subfamily are descended from the subfamily consensus sequence. I based my assessment on how closely this ideal is met.

I identified the most common RepeatMasker annotation among elements associated with each node in the *Alu* Network (Figure 11). Most intermediate nodes and minor replicators identified by AnTE are not distinguished by RepeatMasker; thus, only 24 RepeatMasker subfamilies are required to assign the plurality annotation to each of 256 nodes in the Alu network. The plurality assignments of intermediate nodes to the major replicative sequences generally follows proximity, with the more productive replicative sequences favored: *Alu*Sc and *Alu*Sc8 both capture many nearby nodes, while the less successful *Alu*Sc5 is assigned to only six.

Overall, 60% of elements are assigned to nodes corresponding exactly to RepBase subfamily consensus sequences, 20% to intermediates nodes not in RepBase, and 20% to other nodes, presumably representing other replicative sequences. The existence of unidentified replicative sequences indicates the need for additional subfamily structure in a classification, but the high frequency of intermediates presents a more fundamental problem.

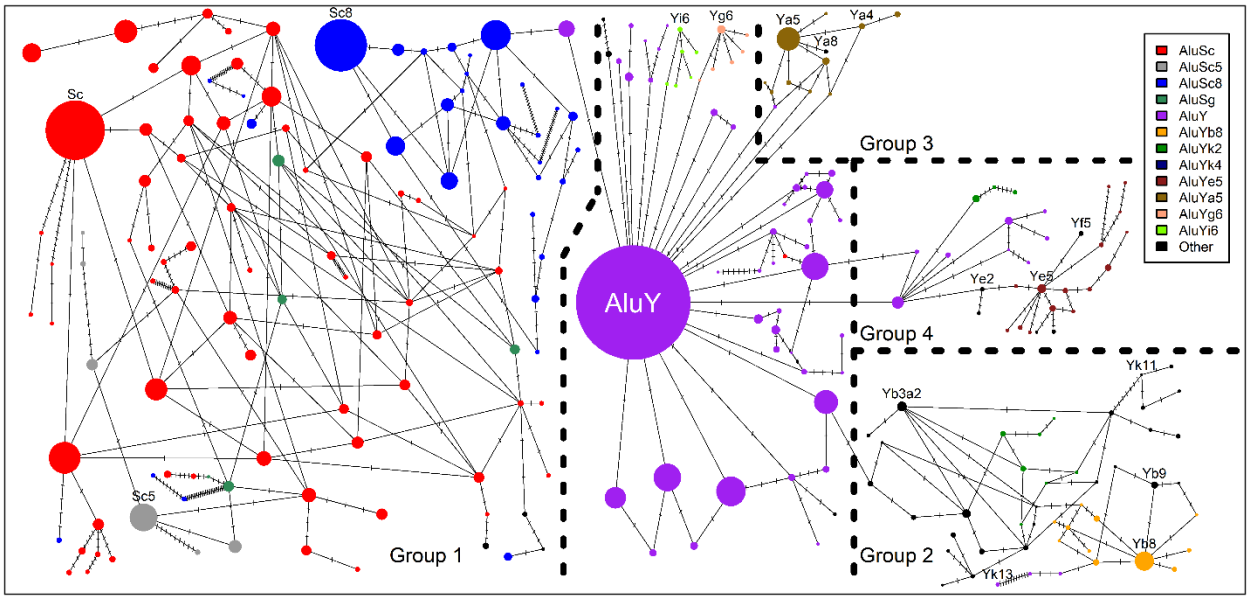

Figure 11: Labeling of Nodes in *Alu* Network by Most Common RepBase Annotation.

Each node is colored by the most common RepeatMasker annotation among elements associated with that node. Only 24 RepeatMasker subfamilies are the plurality winners in at least one node. The network is the same as in Figure 8.

In Group 1, six site variants distinguish *Alu*Sc and *Alu*Sc8, the last one of which is a single nucleotide indel at position 258. Consider the intermediate node representing the sequence that has the *Alu*Sc variant at the first five positions and

the AluSc8 variant at 258; we refer to this node as "*Alu*Sc-258". RepBase assigns 87% of the elements associated with that node to *Alu*Sc and 8.6% to *Alu*Sc8. My analysis indicates that most elements associated with this node are likely either direct gene conversion products between *Alu*Sc and *Alu*Sc8 copies, or were copied from such a product. Are such conversion products "descended from" *Alu*Sc? To answer this question, we must consider two difference concepts of "descent" in a TE family characterized by gene conversion, corresponding to the two different ways such a TE family can copy its sequence, transposition and gene conversion.

The first concept of "descent" is locus-based: an *Alu* locus is descended from *Alu*Sc if the element that was originally inserted at that position was copied from the *Alu*Sc sequence. According to this interpretation of "descent", the RepeatMasker annotation of elements at the AluSc-258 node represents a claim that 87% of those elements were copied from *Alu*Sc. We have little reason to believe this. A locus copied from *Alu*Sc could obtain the position 258 variant by conversion from an AluSc8 copy, but a locus copied from *Alu*Sc8 could also obtain the first five variants from *Alu*Sc through a single conversion event covering all five, located between positions 78 and 152. We cannot confidently estimate the relative probabilities of these scenarios without a much better understanding of the gene conversion process, which RepeatMasker does not model.

The second concept of "descent" is sequence-based: a converted sequence is descended from the sequence from which it converted, such that, if an element originally replicated from *Alu*Sc is converted by an element copied from *Alu*Sc8, the

converted sequence is descended from *Alu*Sc8 while the remainder is descended from *Alu*Sc. In this interpretation, conversion products between descendants of *Alu*Sc and *Alu*Sc8 are hybrids, containing some *Alu*Sc sequence and some *Alu*Sc8 sequence. It is not accurate to claim that a hybrid *Alu*Sc-*Alu*Sc8 element as a whole is descended from *Alu*Sc, because the *Alu*Sc8-descended segment of the element is not. Considering *Alu*Sc-258 again, an *Alu*Sc derived locus experiencing gene conversion from position 153 to 288, or an *Alu*Sc8-derived sequence experiencing gene conversion at positions 78 to 152, would each be assigned to the *Alu*Sc-258 node by AnTE and likely be classified as *Alu*Sc by RepBase, despite containing a large fraction (a large majority, in the latter case) of *Alu*Sc-8 derived sequence. Thus, we cannot say with confidence under either concept of "descent" that elements at intermediate nodes were descended from the RepBase subfamily consensus sequence of their assigned subfamily.

Incomplete Classification Can Bias Age and Mutation Rate Estimates

Due to their high frequency in the genome and low likelihood of being under selection[24], transposable elements are useful models for the study of neutral evolution[21,75]. Substitutions are often inferred by comparing each element to the consensus sequence of its subfamily, under the assumption that this sequence is ancestral to the element, which subsequently evolved according to a typical neutral

substitution process. For hybrid elements, however, this assumption is violated, because the subfamily consensus is only ancestral to part of the element. Differences between the true ancestral sequence of a converted segment and the subfamily consensus will be interpreted as mutations, leading to an upward bias in estimated mutation rates at sites that differ between the ancestor and subfamily consensus (Figure 12). An overestimation of the number of mutations will also upwardly bias age estimates based on molecular clocks.

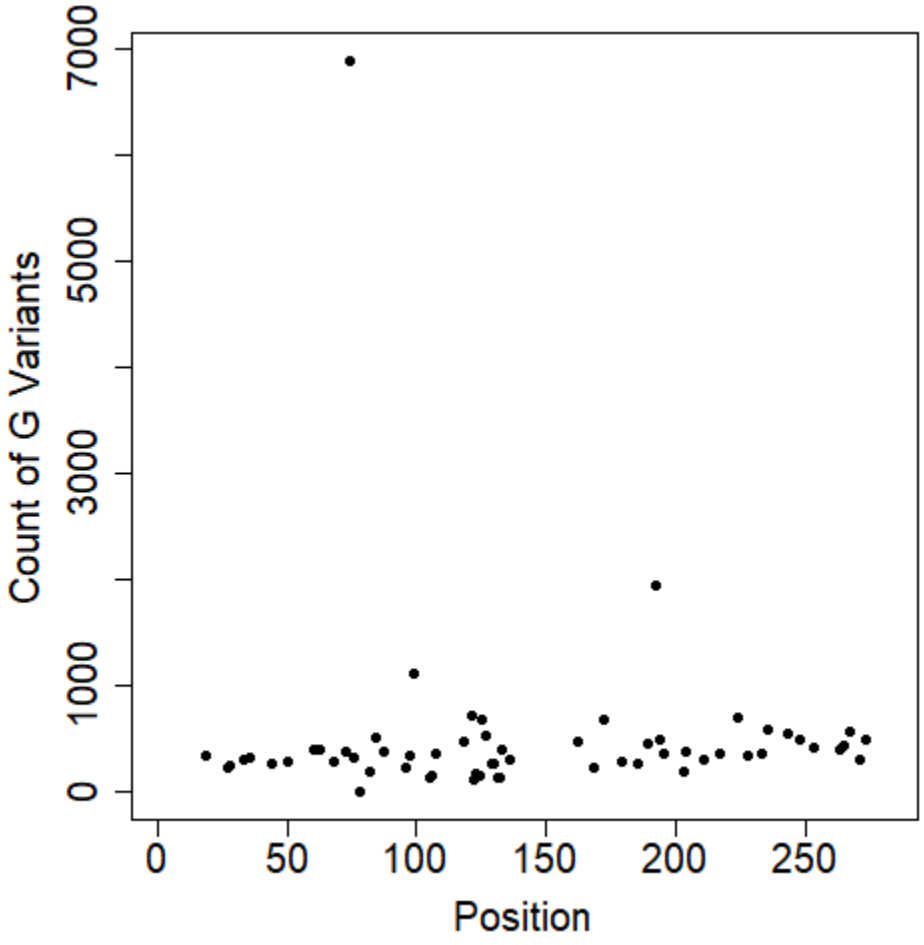

Figure 12: Count of G Variants at each Position that is A in the *Alu*Sc Consensus Among Elements Assigned to *Alu*Sc by RepeatMasker

The number of elements assigned to the *Alu*Sc subfamily by RepeatMasker with a G at each position is plotted for all positions that have an A at the *Alu*Sc consensus. Position 74, which is G in all other major replicators, is an outlier, suggestive of gene conversion. This would lead to inflated mutation rate estimates at position 74 if all differences from the subfamily consensus were assumed to be due to mutation.

One partial solution to this problem is to simply exclude sites that vary between ancestral sequences[56]. While this addresses the problem of some sites having extremely high apparent mutation rates, it does not solve the problem of biased ages. If copies of two replicators active at different times interconvert, and are then assigned to one replicator or the other, the estimated average age of both classes will be biased towards each other, because each class will contain segments descended from the other replicator. To determine whether this is a problem for the RepeatMasker *Alu* annotation, we estimated the average age of elements assigned by RepeatMasker to each RepBase subfamily that is identical to AnTE source sequence, comparing it to the age estimated by AnTE (Figure 13). Though the estimated average ages are in general close, in 3 of 11 cases the 95% credible regions for age do not overlap. Distinguishing intermediate elements from major replicators should allow more accurate estimates for the periods in which those replicators were active.

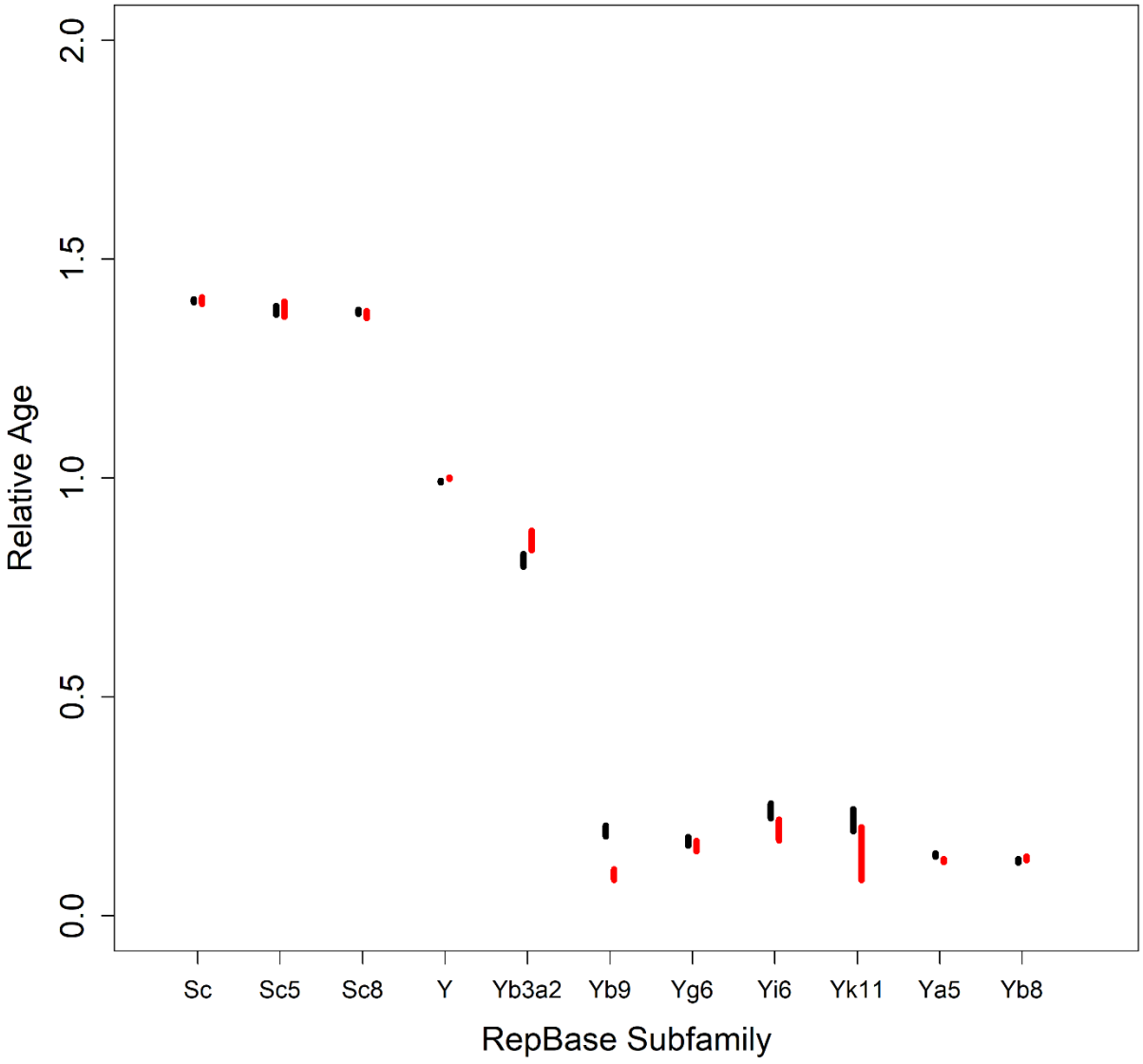

Figure 13: Estimated Average Age of RepBase and AnTE Subfamilies with Identical Consensus Sequences

For every AnTE subfamily with an ancestral sequence exactly matching a consensus sequence for a RepBase subfamily, the 95% credible region for the average age of elements assigned to the RepBase subfamily (in black) and AnTE subfamily (in red) is plotted.

Conclusion

I investigated patterns of sequence diversity among a subset of *Alu* elements in the human genome, attempting to better understand the role of gene conversion in *Alu* evolution. I find strong evidence that gene conversion is a major force influencing the sequence architecture of *Alu* elements. By replacing a segment of one *Alu* element with the homologous sequence from another, the conversion process generates elements with a combination of variants present in the donor and recipient element. Thus, "intermediate sequences", with an alternative combination of site variants present in two or more replicators, are a signature of the gene conversion process. I identified four large groups of related *Alu* sequences, three of which contain both apparent major replicators and high frequencies of intermediate sequences. Though gene conversion is not the only possible explanation for intermediate sequences, I determined that alternative mechanisms are only a minor contribution to the frequency of intermediates. We can therefore conclude that most intermediate sequences between major replicators, which make up around 20% of our dataset, are either conversion products or direct copies of transpositionally-active conversion products.

Prior to this work, there was considerable evidence that gene conversion played a major role in *Alu* evolution[70,72,76]. However, this work is the first systematic attempt, across a large set of *Alu* elements, to identify and quantify the *Alu* sequence types expected to be produced by gene conversion, and explicitly

account for alternative mechanisms by which these sequences could be generated. My findings confirm that gene conversion is not a peripheral behavior of *Alu* elements but a frequent occurrence involving a substantial proportion of elements. Among the oldest group of *Alu* elements in the study dataset, approximately 40% are associated with major replicators and 40% with intermediates, most of which are conversion products or copies of conversion products. Our analysis likely understates the extent of conversion, as complete conversion events, or conversion events not covering variants distinguishing the replicators ancestral to the elements involved, do not produce detectable intermediate sequences.

The consequence of gene conversion on sequence architecture is the accumulation of mosaic conversion products with sequence derived from multiple ancestral replicative sequences. Consistent with previous studies, conversion appears to be much more frequent between more similar sequences[72]; as a result, high-frequency conversion products tend to be hybrids of closely-related replicators. Many of these conversion products appear to have been replicative at some point, often much later than the ancestral sequences from which they derived, generating new classes of *Alu*.

I observe an increasing frequency of intermediates between replicators with increasing age of the replicators. As recent conversion events have been identified even among some of the oldest *Alu* subfamilies[76], it is not surprising that conversion products would accumulate over long time periods. Thus, gene conversion likely

plays an even larger role in the sequence architecture of older classes of *Alu* than those in the study dataset, which primarily contains the youngest classes of *Alu*.

The finding that *Alu* engages in high rates of gene conversion presents major challenges to subfamily classification schemes that largely ignore gene conversion, such as RepeatMasker. In assigning apparent hybrid elements to subfamilies in which each member is supposed to be descended from a replicative sequence with the subfamily consensus, these classifications are misleading, as at most a portion of a hybrid element is derived from any single replicative sequence. The classification presented here demonstrates the feasibility of an alternative approach in which hybrid elements are assigned their own subfamilies.

CHAPTER VI

HIGH RATES OF INTERLOCUS GENE CONVERSION AMONG *ALU* ELEMENTS IN THE GORILLA LINEAGE

Gene conversion is a recombination-associated process in which a donor strand is used as a template to repair a double-strand break in a homologous acceptor strand, resulting in a unidirectional transfer of genetic information.[77] Gene conversion can occur between allelic pairs at the same locus or between nonallelic pairs at paralogous loci.[78,79] Both allelic and interlocus gene conversion have greatly influenced the evolution of eukaryotic genomes.[80] Allelic gene conversion is an important contributor to substitution rate variation and GC-content heterogeneity in a variety of taxa[80,81], and interlocus gene conversion drives concerted evolution among gene families[82]. Interlocus gene conversion events have also been implicated in many human inherited diseases.[83]

Alu transposable elements in primates appear to undergo frequent interlocus gene conversion due to their high similarity and large copy number[72,84]. There are approximately 1 million *Alu* elements in the human genome, comprising around 10% of the human genome overall, most of which shared a common ancestor 45-60

MYA[84]. Thus, there are many more potential *Alu* gene conversion pairs than among most human gene families that have much lower copy number, but still occasionally experience interlocus gene conversion[2,10]. Evidence for high rates of *Alu* gene conversion[70–72,76,85] include identification of numerous "mosaic" *Alu* elements, appearing to result from conversion between elements of different subfamilies[70]. In Chapter V, I provided evidence that approximately 20% of *Alu*Y and younger *Alu*S elements appear to be mosaic. Aleshin and Zhi[72] also found that neighboring *Alu* elements along the genome are substantially more similar than random pairs, suggesting high rates of *Alu* gene conversion among elements in close proximity.

Given its propensity for gene conversion, *Alu* is a potentially useful system for understanding the conversion process as it has occurred in primate genomes, but existing methods for studying it have several limitations. The method of Aleshin and Zhi[72], based on identifying an excess number of shared mutations among nearby elements, can identify signatures of gene conversion but cannot identify individual conversion events. Methods such as GENCONV[86], which compare pairs of sequences to identify regions with higher than expected similarity, can determine that a conversion event occurred but can neither distinguish between donor and recipient loci, nor infer what the sequence was at either locus prior to conversion.

To better understand gene conversion in *Alu* elements, I developed a Bayesian phylogenetic approach designed to identify and characterize gene conversion events. I applied this approach to sequence data from orthologous *Alu* loci among four Great Apes, identifying gene conversion events between pairs of

nearby sequences on each branch of the phylogenetic tree relating these four species. For each conversion event, my method infers the branch on which the event occurred, identifies both the donor and receptor loci, and probabilistically infers the sequence at each locus before and after the conversion. Due to a fortuitously high rate of gene conversion along the lineage leading to gorillas, I obtain sufficient information for a more complete picture of individual conversion events than previous analyses of *Alu* gene conversion.

Obtaining *Alu* Ortholog Alignments

A dataset of *Alu* elements was obtained from the RepeatMasker[87] annotation of the hg38 assembly of the human genome. Only full-length sequences (275-325 bp) were included, producing a dataset of 779,310 elements. Human elements were aligned to a consensus of the *Alu* sequence using the probabilistic version of the Needleman-Wunch algorithm described by Zhu et al.[60] as described in Chapter III.

The 6 primate EPO[88] whole-genome alignment was acquired from Ensembl release 71[89], which is based on the GRCh37 assembly of the human genome. We used five of these genomes: human, chimpanzee, gorilla, orangutan and macaque. The positions of elements in the *Alu* dataset were used to identify *Alu* positions in the whole-genome alignment. Because EPO contains an alignment of orthologs to the human, we used these alignments, together with our alignments of human *Alu* to the consensus, to obtain alignments of each ortholog to the consensus.

All *Alu* loci that did not include human, chimpanzee, gorilla, and orangutan orthologs were filtered out. As macaque is used only for substitution rate estimation, I did not exclude loci missing only macaque. To protect against misalignment of *Alu*, I also filtered out loci in which chimpanzee, gorilla, or orangutan in either 100 bp flanking region around the element differed by more than 10% from the human flanking region.

Testing Potential Conversion Pairs

Each pair of elements were tested for conversion between each element within 100 kb, considering each in turn as potential donor and acceptor, and considering separately the possibility of conversion on each branch in the Great Ape phylogeny (i.e., the phylogeny including human, chimpanzee, gorilla, and orangutan as leaves).

Given a possible donor element, a possible acceptor element, and a possible branch on which a conversion event occurred, I consider the relative likelihood between two scenarios: first, that the two elements evolved independently, neither experiencing gene conversion; second, that there was a conversion event in which a segment of the recipient element was completely converted to that of the donor element.

The marginal likelihood of a conversion event involving a specified potential acceptor, donor, and branch is the sum of the likelihoods of each possible conversion

tract. I thus estimate the likelihood of every possible contiguous tract in the alignment, relative to the scenario of no conversion. For each position in the tract, I consider two possible pairs of trees, the first indicating that the two homologous nucleotides at the possible donor and recipient evolved independently according to the normal Great Ape phylogeny, and the second indicating that the recipient branched off the donor at the point of conversion (Figure 14). The relative likelihood of a tract position is the relative likelihood between the trees indicating non-independent evolution of the homologous nucleotides at that position and the trees indicating independent evolution. Outside of the tract, positions evolve along the same trees in the conversion and non-conversion scenario, so the relative likelihood of each position is one. The relatively likelihood of a tract is the product of the relative likelihood of each individual position in the tract. Using this approach, estimating the conversion likelihood reduces to the problem of estimating the relative likelihood of two sets of trees for every site in the tract.

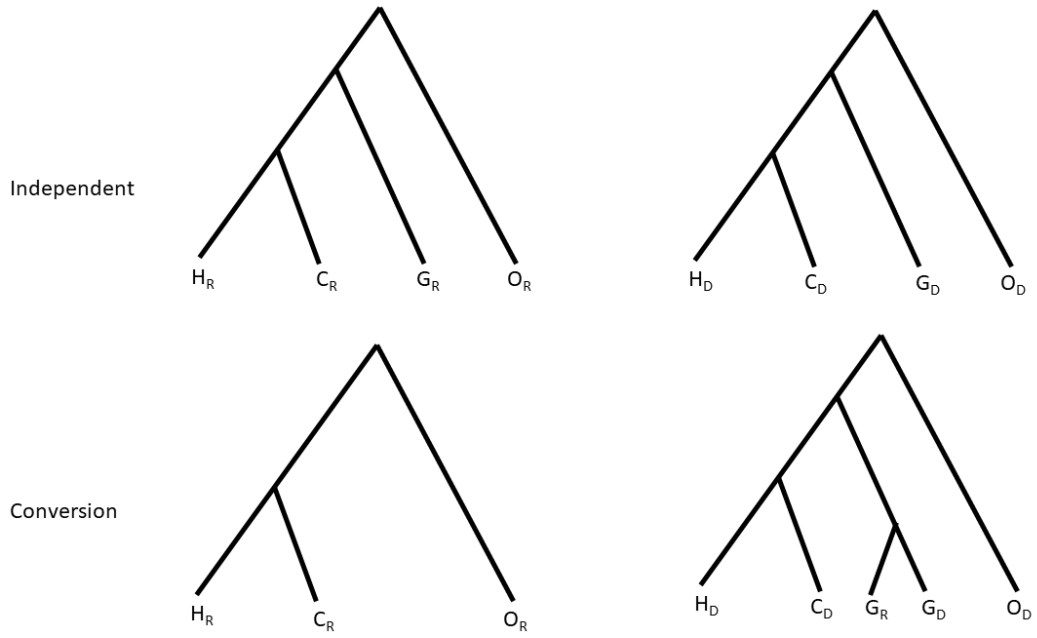

Figure 14: Trees Representing Scenarios of Conversion and Independent Evolution for a Homologous site at a Potential Acceptor and Donor Locus.

Each leaf represents one of eight homologous nucleotides among four orthologues and two loci. Leaf labels indicate species (human (H), chimpanzee (C), gorilla (G), and orangutan (O)) and presence at either the donor (D) or recipient (R) locus. Each set of trees represents one possible scenario. If the site was not converted, both sites evolved independently according to the typical Great Ape phylogeny. If the site was converted, the recipient branches off from the donor lineage after the conversion event. The tree shown corresponds to conversion in the gorilla lineage.

Given a tree, substitution probabilities across each branch, and a set of leaves, the tree likelihood can be calculated using Felsenstein's algorithm.[90] The leaves of each tree are obtained from the homologous site in the element across the potential donor and recipient element in human, chimpanzee, gorilla, and orangutan. Substitution probabilities across the branches of the Great Ape phylogeny are straightforward to obtain, as described below. Estimating substitution probabilities on the branches surrounding the conversion event itself is

more complicated, as the time of conversion is unknown. To address the unknown conversion timing, we average tract likelihoods across 100 possible conversion points spaced equally across the branch of the Great Ape phylogeny under consideration.

The conversion likelihood for an acceptor-donor pair reflects the strength of evidence for conversion of recipient by donor relative to the possibility of independent evolution of each element without conversion. However, these are not the only possible scenarios. In particular, if the element under consideration as recipient was converted, but by a different donor, the conversion scenario may be strongly favored over the independence scenario for many considered donors that are closer to the actual donor than the recipient. To guard against such false positives, for every potential conversion event passing a likelihood threshold, the donor-recipient pair under consideration is compared to 10,000 pairings of the recipient with random elements in our dataset. Then, the proportion of donors with higher likelihoods than the donor under consideration is determined. Ever donor-acceptor pair, then, is associated with two values: the relative likelihood and donor percentile. For a potential conversion pair to go into the conversion set, I required that the pair have relative likelihoods larger than e^{30}, and the potential donor have higher likelihoods with the recipient than 99.95% of random potential donors. These thresholds were chosen to obtain a false positive rate below 5%, as described below.

If a donor-acceptor pair is in the conversion set, it is of interest to infer properties of the conversion event, as well as information about the donor and

acceptor before and after the event. I draw events (given donor, recipient, and branch) from the posterior distribution using the following procedure. I first draw a tract with probability proportional to its relative likelihood. I then draw a point on the branch at which the event occurred, conditional on the selected tract. Conditional on the above, I then draw trees from the posterior using Felsenstein's algorithm, with each tree indicating the nucleotide at each node in each species in the donor and recipient. For analyses, I draw one event from the posterior of each pair in the conversion set.

False Positive Rate Estimation

To estimate false positive rates, for every element in the *Alu* alignment dataset, I randomly placed it somewhere else in the genome, at least 1 Mb away from its true position. Then, I tested it as a potential conversion recipient for all elements within 100 kb. I identify all events passing the same thresholds for admission to the conversion set, putting them in the false-positive set. As large-distance conversion events are unlikely, these events should consist primarily of false positives. I estimate the false positive rate as the number of events in the false positive set divided by the number of events in the conversion set.

Substitution Probability Estimation

Substitution probabilities were estimated from the EPO alignment, which contains inferred ancestral sequences. For each branch of the Great Ape phylogeny, the probability of being in state *Y* given state *X* at the start of the branch was estimated by dividing the number of *X* to *Y* changes by the number of positions in state *X* across all elements at the start of the branch. Substitution probabilities for each possible *X* to *Y* substitution were estimated separately, and C to A and G to T substitutions at CpG sites are distinguished from other such substitutions.

For most purposes, these substitution probabilities across the Great Ape phylogeny can be used directly. Given a conversion event along a branch, however, we must estimate substitution probabilities up to that event, and from that event to the present. To do this, substitution probabilities were first converted to rates. As substitution probabilities are low, I assume no more than one event per site, in which case:

$$\lambda_{X\to Y,b} = \frac{-\log(1 - p_{x\to y,b})}{t}$$

where $\lambda_{x\to y,b}$ is the rate of substitution from X to Y along branch b, $p_{x\to y}$is the probability of substituting from X to Y across that branch, and t is the branch length, defined to be 1 for the entire branch. After estimating the rate across the

entire branch, I use this equation to estimate substitution probabilities along sections of the branch.

Alu Gene Conversion Across the Great Apes

I scanned the Ensembl EPO 5-primate alignment[91] for *Alu* elements identified by RepeatMasker[73] and present in the human, chimpanzee, gorilla and orangutan genomes. I then applied my gene conversion detection algorithm, TEConv, to identify a set of high-confidence gene conversion events between elements in this set. The algorithm evaluated pairs of loci for possible gene conversion, comparing the likelihood of independent evolution of the loci since the origin of the Great Apes versus the likelihood a conversion event occurred on one terminal branch of the Great Ape tree.

Previous research indicated that gene conversion occurs much more often between close paralogs on the same chromosome than between distant paralogs or paralogs on different chromosomes[72,78]. Because of this, I evaluated conversion only between pairs of elements within 100 kb of each other. To estimate the false positive rate, each element was successively placed at a random position in the genome at least 1 Mb from its true position and the algorithm run as normal; this gives the detection rate of the algorithm when there are likely to be essentially no

real events. I set thresholds for relative likelihood of conversion to obtain a false positive rate of 5% in the high-confidence set.

I identified 2,537 gene conversion events across the terminal branches of the orangutan, gorilla, chimpanzee, and human lineages. Surprisingly, nearly all these events, 2,514, occurred along the gorilla lineage, while only 15 occurred along the human lineage, 8 along the chimpanzee and 14 in orangutan. To confirm this unexpected result, I consider another indicator of gene conversion, the substitution rate at highly variable sites within *Alu*. The most variable sites in *Alu* distinguish *Alu* subfamilies; i.e., they differ between major replicative sequences and thus differ between their descendants. A rate of substitution at these sites higher than expected from mutation alone suggests a high rate of conversion between elements in different subfamilies. Considering, for example, the sites that are C in the *Alu* consensus, two positions, 153 and 197 in the alignment, have a G in *Alu*Y, the youngest major division of *Alu* elements, while a third, 94, has a G in the older division *Alu*J. The rates of substitution from C to G at these sites in most branches in the Great Ape phylogeny are well within normal site variation, suggesting most substitutions at these sites are from mutation rather than interlocus gene conversion (Figure 15). In contrast, these sites are outliers in substitution rate in gorilla (Figure 15C), consistent with high rate of gene conversion between subfamilies. Thus, it appears that the terminal gorilla branch experienced vastly higher rates of gene conversion among *Alu* elements than other branches of the

Great Apes. As rates of gene conversion are so low outside the terminal gorilla branch, further analyses are restricted to conversion events on this branch.

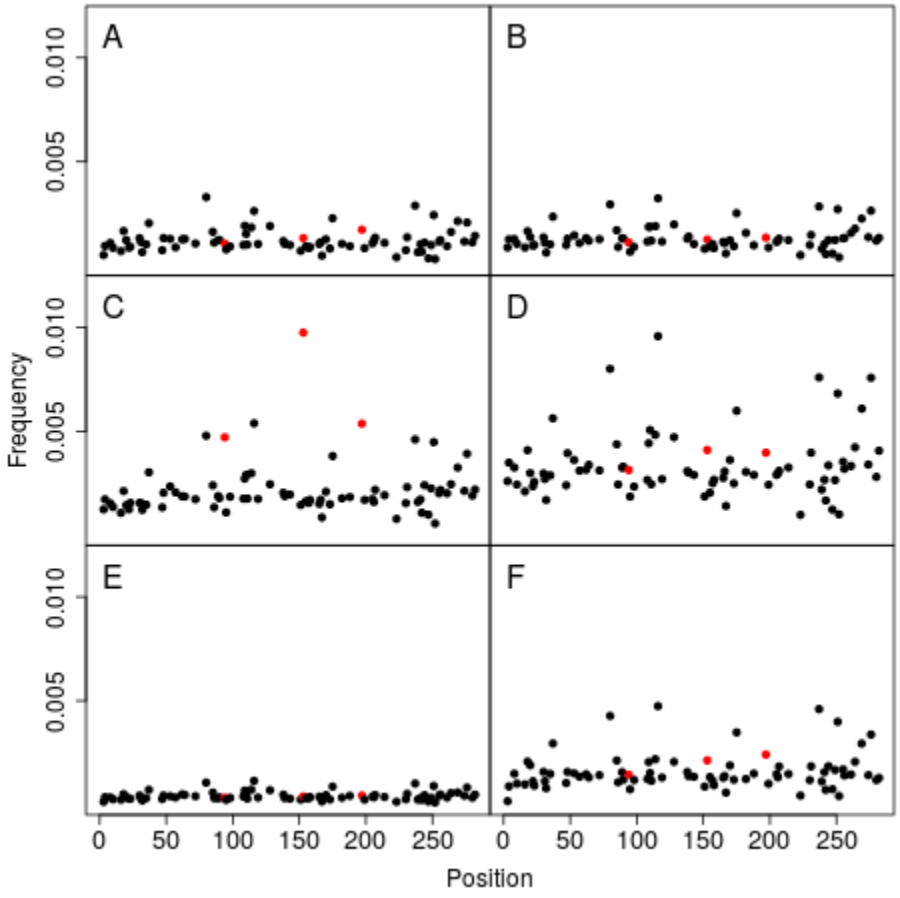

Figure 15: Rate of C to G Substitution Across the *Alu* Sequence, along Six Branches of the Great Ape Phylogeny.

Each point gives the estimated substitution rate for sites that are C at the start of the branch. Only sites that are C in the *Alu* consensus are shown. Sites that are G in major subfamily consensus (*Alu*Y, *Alu*S, or *Alu*J) are shown in red. A) Human terminal branch. B) Chimpanzee terminal branch. C) gorilla terminal branch. D) orangutan terminal branch. E) Branch between gorilla and chimpanzee divergence. F) Branch between orangutan and gorilla divergence.

Why are Rates of *Alu* Gene Conversion Abnormally High in the Gorilla Lineage?

A possible explanation for the two orders of magnitude higher rate of interallelic gene conversion along the gorilla lineage might be found in the PRDM9 recombination targeting system. During meiosis, double strand breaks are induced in chromosomes to initialize recombination.[92] In mammals, specific sites are targeted by DNA binding of the protein PRDM9, resulting in hotspots for recombination and double strand breaks.[92,93] As alleles containing PRDM9 target sites are preferentially converted by alleles which are not targeted[94,95], the action of PRDM9 binding tends to eliminate target binding sites over time, and the PRDRM9 protein evolves extremely rapidly, especially at DNA contact sites[96], to find new binding sites and allow recombination to continue in future generations. The end result is that there is essentially no overlap of recombination hotspots among different Great Ape species[97] and considerable variation even among different human populations[98].

Thus, there is a known system operating in the primate genome by which we expect high frequency of double strand breaks to be induced at particular targets over short time periods on an evolutionary scale. If part of the *Alu* sequence was targeted by PRDM9, we would expect a burst of both allelic and interlocus gene conversion among *Alu* elements over a short time period, likely on a single branch of the Great Ape tree, as we observe.

A PRDM9 mechanism for excess interlocus gene conversion in gorilla *Alu* elements predicts that there should be a PRDM9 binding motif common among *Alu* elements prior to the period of conversion. Currently PRDM9 motifs have been identified by analyzing recombination hotspots[93], but predicted PRDM9 motifs based on analysis of the gorilla PRDM9 structure[96] do not match any sequence close to the *Alu* consensus. This suggests that the putative ancestral gorilla PRDM9-*Alu* binding motif is no longer active, a result that is not surprising given the transiency of recombination hotspots.

If there is a motif that predicts gene conversion acceptor probability, then differences from this motif prior to gorilla divergence should be associated with lower levels of conversion. The latest point prior to gorilla divergence at which sequence can be inferred for both converted and unconverted elements is the root of the Great Ape tree. To search for a motif, I compare the frequency of variants at each site among both elements that underwent gene conversion in gorilla and elements that did not. If a variant is part of the motif, there should be a higher frequency of that variant among converted elements at the root than among unconverted elements. To avoid identifying variants that merely indicate subfamily membership, I conduct this analysis separately for *Alu*Y and *Alu*S elements.

The ratio of non-consensus variants among elements that did not convert relative to those that did is fairly even along the *Alu* sequence, with the notable exception of a significantly elevated region at positions 242-256 (Figure 16, Figure 17). This result is similar in *Alu*Y and *Alu*S sequence. The number of non-consensus

variants in the 242-256 region at the root of the Great Apes strongly predicts subsequent conversion probability (Figure 18). Among 89,959 elements that contained the *Alu* ancestral motif in this region at the beginning of the gorilla lineage, 1.8% were converted along the gorilla lineage, while only 0.68% of the 97,408 elements with a single difference from the ancestral motif were converted along the gorilla lineage ($p<.0001$, Fisher exact test). In contrast to conversion results of acceptor elements, no region of *Alu* appears to strongly relate to the probability that a sequence will act as a donor element (Figure 18).

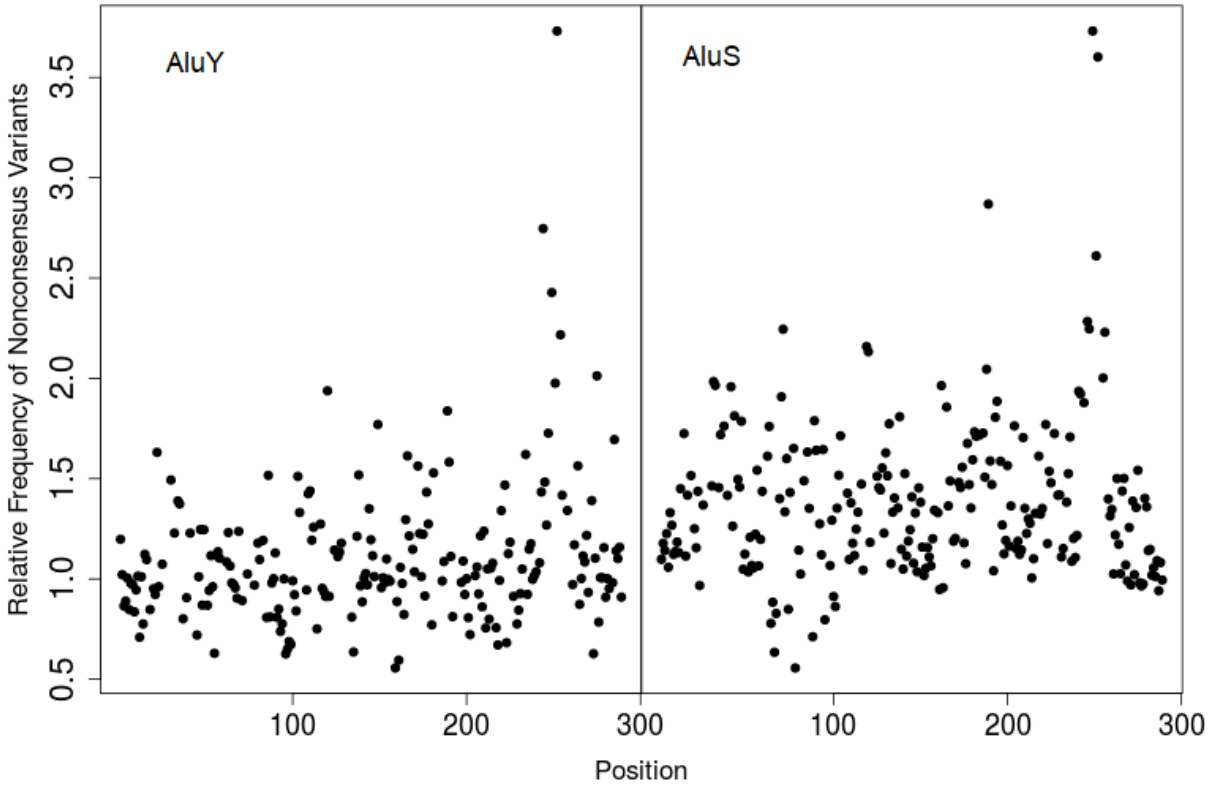

Figure 16: Ratio of the Frequencies of Non-Consensus Variants for Unconverted Elements Versus those for Converted Elements, for each Position in *Alu*.

The proportion of *Alu* loci with non-consensus variants at the root of the Great Apes was calculated for each site, and the ratio of unconverted relative to converted is plotted for

each site. Positions differing between *Alu*S, *Alu*Y, and *Alu*J were excluded. Analysis was conducted separately for AluY elements and AluS elements.

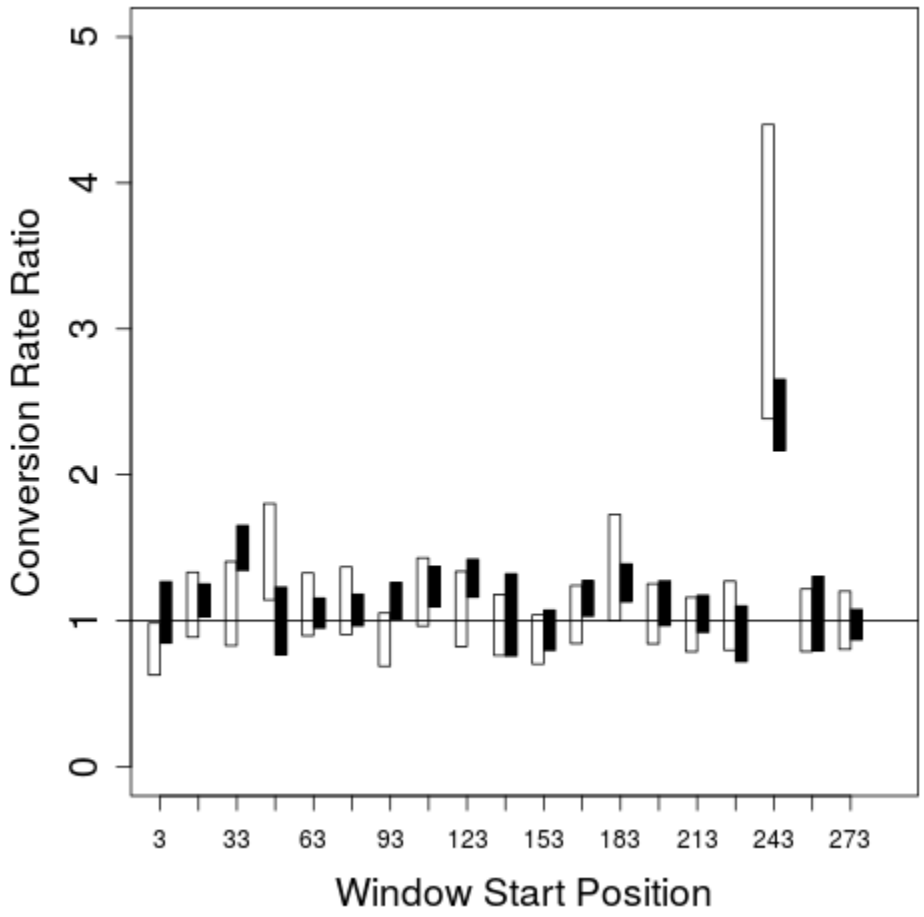

Figure 17: Conversion Acceptor Ratio Between Elements with and Without Perfect Match to the Consensus in 15 bp Windows Across *Alu*.

For 15 bp windows across *Alu*, the proportion of elements that are acceptors in the conversion set was calculated for loci that had perfect match to the consensus at the root of the Great Apes in that window (excluding positions differing between *Alu*S, *Alu*Y, and *Alu*J) and all other elements. The white bars show 95% credible regions for *Alu*Y elements, while the black bars show 95% credible regions for *Alu*S elements.

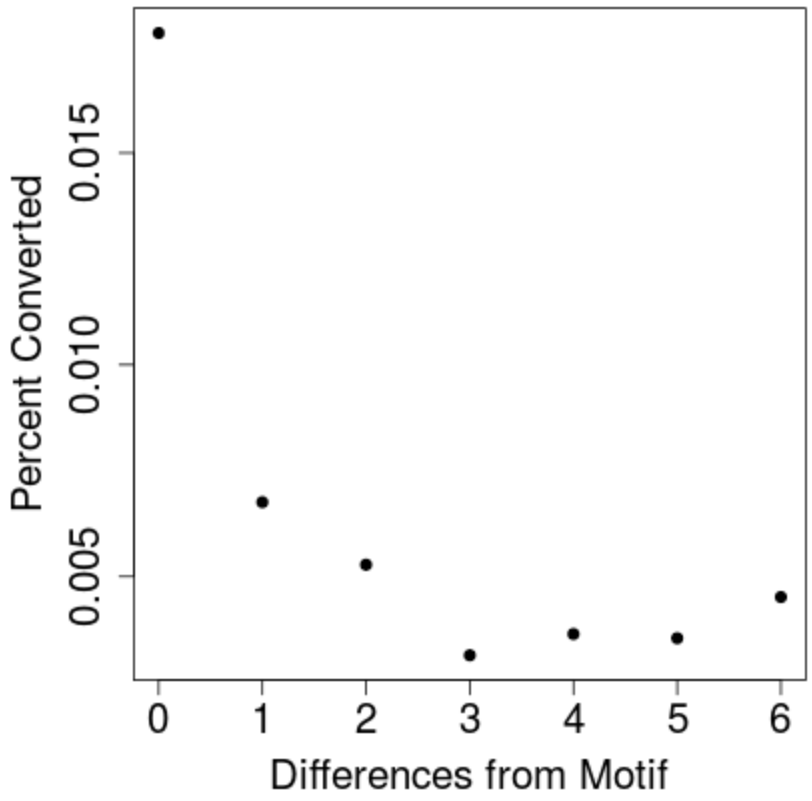

Figure 18: Conversion Percent by Motif Differences.

For each count of differences from putative motif, the percent of elements with that many differences that were gene conversion acceptors is plotted.

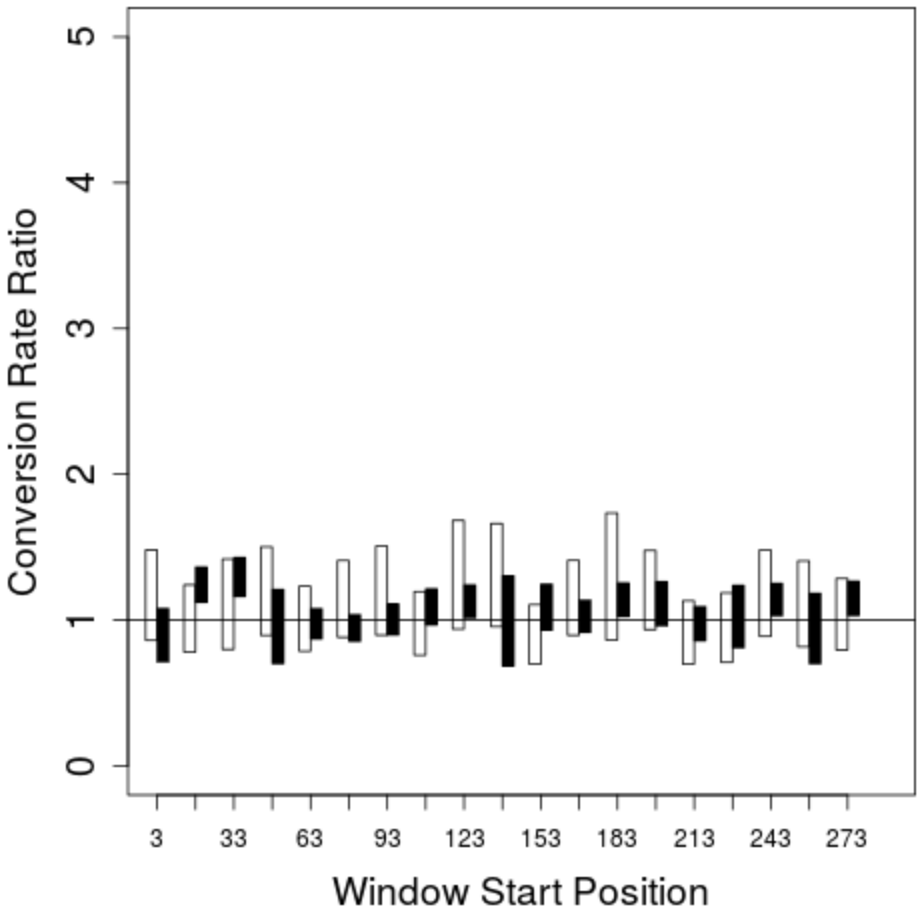

Figure 19: Conversion Donor Ratio Between Elements with and Without Perfect Match to the Consensus in 15 bp Windows Across *Alu*.

For 15 bp windows across *Alu*, the proportion of elements that are donors in the conversion set was calculated for loci that had perfect match to the consensus at the root of the Great Apes in that window (excluding positions differing between *Alu*S, *Alu*Y, and *Alu*J) and all other elements. The white bars show 95% credible regions for *Alu*Y elements, while the black bars show 95% credible regions for *Alu*S elements.

Known PRDM9 target motifs show depletion in the genome[99] because substitutions are accelerated by meiotic drive, a tendency for new alleles that disrupt PRDM9 binding to be preferentially transmitted to offspring; this can be explained if PRDM9 binding induces *cis* double-strand breaks that then tend to be repaired using the new allele on the sister chromatid[94]. The hypothesis that the

ancestral 242-256 *Alu* motif was bound by PRDM9 for some period of time on the gorilla lineage is supported by the observation that the motif is depleted among *Alu* elements in gorillas relative to other Great Apes, while *Alu* loci with 242-256 region motifs differing by 1 or 2 nucleotides from the ancestral motif are more frequent in gorillas (Figure 19). The gorilla genome has 62,138 *Alu* copies that match the motif while other Great Apes have between 68,245 and 69,456 copies with intact motifs.

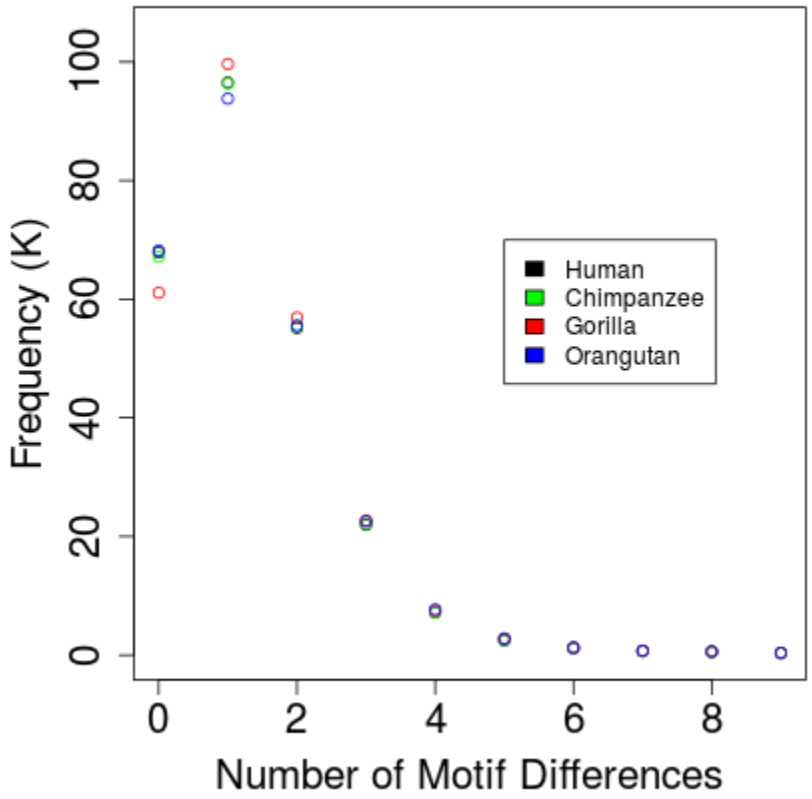

Figure 20: Count of Elements by Number of Differences from Putative PRDM9 Binding Motif in Four Great Apes.

In each of the human, chimpanzee, gorilla, and orangutan genomes, the number of *Alu* elements with each possible number of differences from the putative PRDM9 binding motif is plotted.

Identified gene converted loci do not match modern gorilla hotspots identified by Stevison et al.[97]; 12.4% of gene-converted *Alu* loci are within 10 kb of a Stevison hotspot (318 out of 2571), compared to 13.5% of non-converted loci (34021 out of 251,496). Average gorilla recombination rates within 100 kb of *Alu* loci, also from Stevison et al.[97], do not differ significantly between gene-converted loci and non-converted loci (p=0.29, t-test). However, average human recombination rates, as estimated by Kong et al.[98], are 5.4% higher on average within 100 kb of *Alu* loci that were identified as gene-converted on the gorilla lineage, than within 100 kb of loci that were not gene-converted on the gorilla lineage (p=0.0003); similarly, human recombination rates are 5.1% higher among *Alu* loci identified as gene-conversion donors on the gorilla lineage (p=0.0006).

Conversion Probability Declines with Interlocus Distance

To evaluate the relationship between interlocus gorilla lineage gene conversion events and genomic distance, the distance between the midpoints of donor and acceptor elements was determined. As expected from prior research[72], the rate of interallelic gene conversion along the gorilla lineage declined with interallelic distance (Figure 21). Half of identified conversion events occur within 12 kb and 90% within 50 kb, and the decline appears to be approximately exponential.

To quantify this relationship more precisely, we fit an exponential model estimating the gene conversion probability along the gorilla lineage by distance:

$$p(C|d) = \alpha e^{-\beta d}$$

where $p(C|d)$ is the gene conversion probability given a distance, d, between donor and acceptor loci. The value of α, the gene conversion probability for adjacent loci, was estimated as 0.017 (95% credible region: 0.016 - 0.018) and the exponential decrease parameter, β, was 0.000048 (the 95% credible region was 0.000046 - 0.00005). Extrapolating from the model, fewer than 1% of conversion events are predicted to occur at distances greater than the 100 kb distance cutoff used in this analysis.

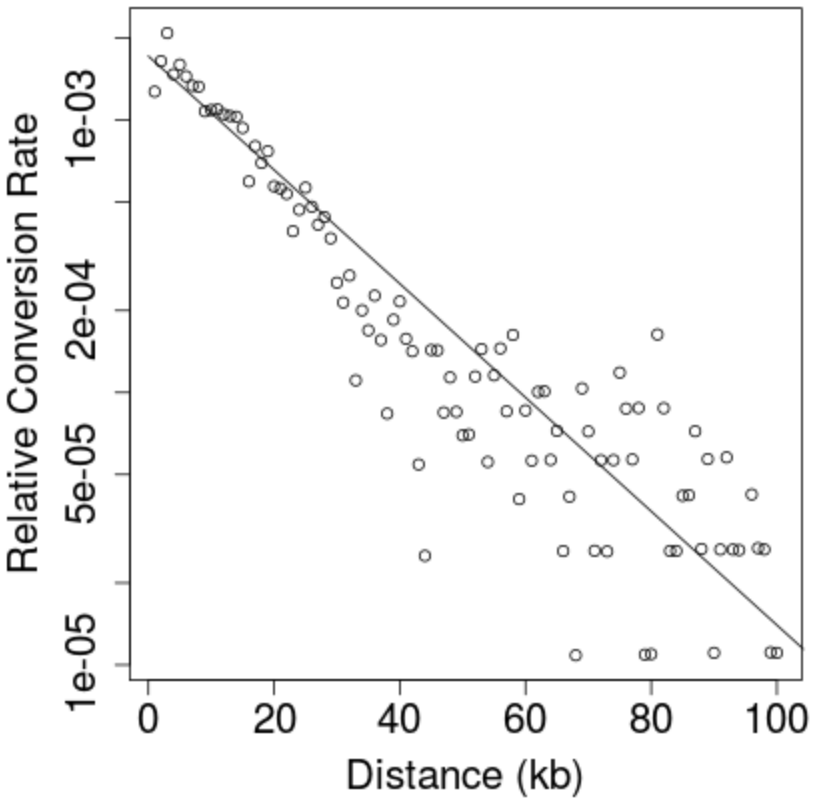

Figure 21: Conversion Rate by Distance.

All possible *Alu* pairs were placed in bins of 1000 bp based on the distance between donor and recipient, measured at midpoint. The proportion of possible pairs in each bin that are in the conversion set is plotted. The line shows the predicted rate at each distance from an exponential model.

This analysis suggests *Alu* conversion pairs are generally quite close, but the high density of *Alu* loci would mean that for any given locus there was typically a large possible selection of partners. The average element has 3.6 potential partners in our dataset within 12 kb and 14 within 50 kb. Among identified gorilla conversion events, 34% involved closest neighbors.

While 51.3% of potential matchups within 50 kb are in the parallel orientation, only 48.4% of inferred conversion events within 50 kb are between elements in parallel, a significant difference (p=0.004, Fisher's exact test). The slight preference for anti-parallel orientation among conversion events in this dataset differs from the result of Aleshin and Zhi[72], who found a stronger signature of gene conversion between neighboring *Alu* elements in parallel. This may indicate differences in gene conversion dynamics over time; the Aleshin and Zhi result reflects long-term patterns of gene conversion, while these results involve events only along the terminal gorilla branch.

Conversion Tract Sizes and Positions

I used inferred intergenic conversion events to estimate a posterior probability distribution of conversion tract sizes. The mean tract size is 118 bp (standard deviation 54 bp), with a median of 109 bp. Tract size frequency among identified conversion events rises rapidly from 25-75 bp, then steadily but more slowly declines in the range from 100-200 bp (Figure 22). To correct for bias against detecting shorter tracts, I simulated conversion events to estimate a false negative rate for each tract size. These simulations were then used this to infer the corrected distribution including missing events, but this does not greatly change the general pattern (Figure 22).

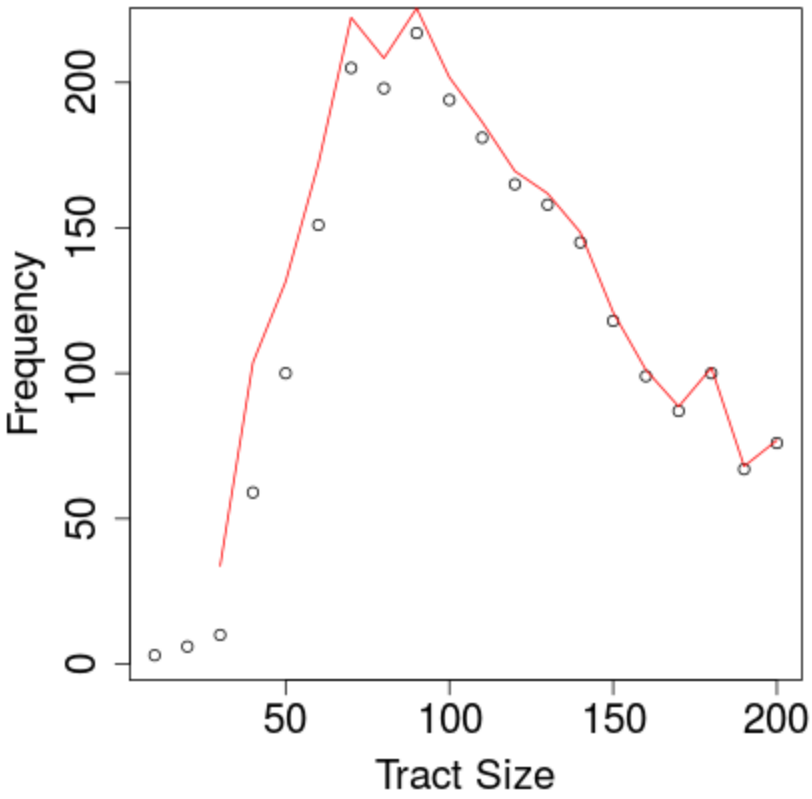

Figure 22: Tract Size Frequencies.

All pairs in the conversion set were placed in 10 bp bins based on tract size, and the number of pairs in each bin was plotted. The red line gives the estimated frequency in each bin after adding in inferred missed conversion events.

I also considered the number of times each position in the *Alu* element was covered by a conversion tract (Figure 23). The most covered region of the element is in the middle of the 289 bp *Alu* alignment, around positions 136-166, and the most covered site, position 149, is included in 65% of tracts. In contrast, the ends are covered much less: for example, position 20 is in only 5% of tracts, and position 269

in 17% of tracts. The putative PDRM9 motif at *Alu* alignment positions 242-256 does not appear as an outlier in coverage (only 36% of tracts cover the motif).

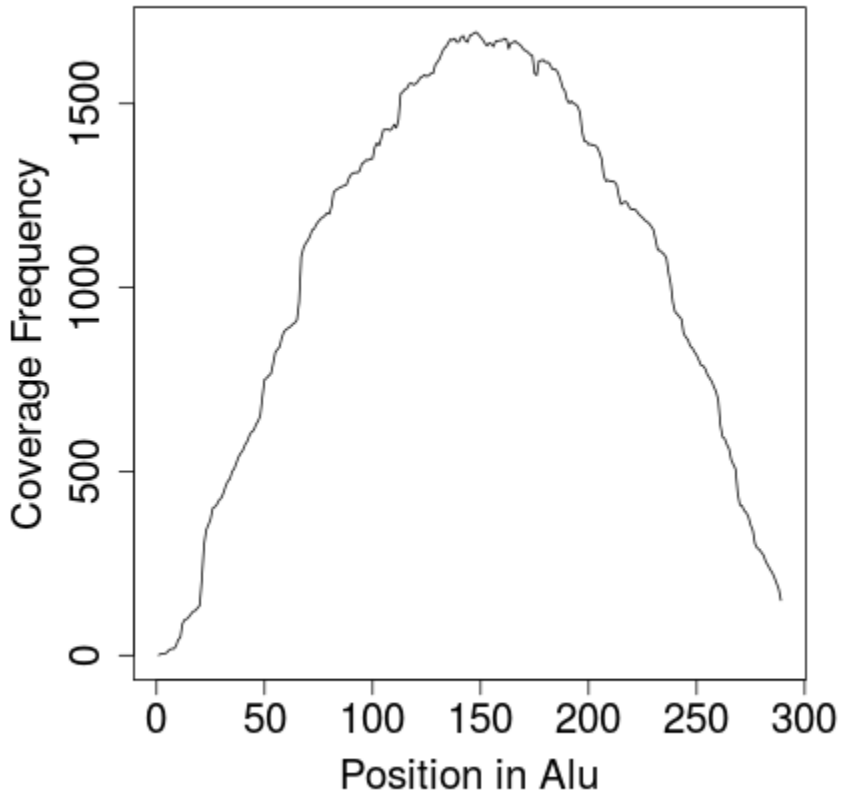

Figure 23: Coverage of the *Alu* Sequence by Conversion Events.
For each site in Alu, the number of times it was included in a conversion tract was plotted.

Gene Conversion Donors and Acceptors

Our ability to detect conversion events is not independent of the identity of donor and recipient elements; conversion between more distantly-related elements

is easier to detect than between similar elements because conversion events between distant homologs introduce more base changes. Conversion from younger to older elements is also easier to detect than the reverse, because older donors include more distinct locus-specific variants that can be used to distinguish them from other possible donors. The conversion set, therefore, is not representative of the conversion events that occurred on the gorilla lineage. The above analysis suggests, however, that the rate of false negatives is low (<5%) for conversion tracts larger than 100 bp. To address sampling bias, I construct a large-tract subset consisting only of events with tracts 100 bp are larger; this subset contains 1428 events.

Elements from the youngest major *Alu* division, *Alu*Y, are disproportionately likely to be both gene conversion recipients and donors in either the full conversion set or the large-tract subset (Table 3). The middle division, *Alu*S, is also disproportionately present as both recipients and donors, though to a smaller extent. Elements from *Alu*J, the oldest major division, are greatly underrepresented as donors and recipients relative to their frequency in the genome. Note that this pattern runs in the opposite direction of the expected bias in identifiability, because older elements are more dissimilar to their potential conversion partners. Thus, the dataset likely understates the age effect on conversion probability, and we can infer that elements that have been in the genome for longer are less likely to be involved in gene conversion as either donors or recipients. This result is expected, as previous research indicates a strong relationship between sequence similarity and

conversion probability[72]. Consistent with this relationship, identified conversion pairs are more likely to belong to the same division than expected from the donor and recipient frequencies alone (Table 4).

	Recipients	Donors	Large-Tract Recipients	Large-Tract Donors	All gorilla Elements	Conversion Probability
*Alu*Y	18.0% (462)	13.0% (334)	21.6% (315)	14% (205)	8.6% (21,813)	2.1%
*Alu*S	77.9% (2002)	78.3% (2014)	75.2% (1095)	76% (1109)	67.3% (171,041)	1.2%
*Alu*J	4.1% (105)	8.5% (219)	3.2% (46)	9.5% (138)	24% (61,010)	0.2%

Table 3: Participation of Major Divisions of *Alu* in Conversion Events

Recipient: Donor	Expected	Observed
Y:Y	2.3%	3.7%
Y:S	14.1%	13.2%
Y:J	1.5%	1.1%
S:Y	10.1%	9.1%
S:S	61.0%	62.3%
S:J	6.6%	6.4%
J:Y	0.5%	0.2%
J:S	3.2%	2.8%
J:J	0.3%	1.0%

Table 4: Percent of Conversion Events Involving Each Possible Pair from the Major Divisions of *Alu*

While younger subfamilies are overrepresented among both donors and acceptors, they are more strongly so for recipients than donors (Fisher exact test, $p<.0001$). The donor-acceptor difference is similar in the overall dataset and the large-tract subset, so this does not appear to be a result of sampling bias. This difference may be partly explained by presence of an intact putative PRDM9 binding motif, which strongly affects an element's probability of being an acceptor but not its probability of being a donor. Only 28% of *Alu*J elements had a fully

intact motif at the root of the Great Apes, compared to 34% of *Alu*S elements and 68% of *Alu*Y elements. If we consider only the 742 pairs in which both elements had fully intact motifs, 22.2% of donors (165) are *Alu*Y compared to 25.7% (191) of recipients, a smaller difference, though still significant (Fisher exact test, p=0.0013). It is likely that there are other sites on the *Alu* sequence than the motif itself that are important for double-strand break targeting. The first-identified human PRDM9 motif appears to produce stronger hotspots in the background of THE1 elements than in other repeats or non-repetitive sequence.[99]

On average, pairs of loci involved in conversion along the gorilla lineage differed at 53.9 sites (standard deviation 11.5 sites) in their ancestral state at the Great Ape root; this is 18.7% of the length of the *Alu* alignment. In contrast, *Alu* elements overall differed by an average 67.3 positions at the Great Ape root, and only 20% differed less 54. Thus, conversion pairs are substantially more similar than *Alu* elements overall, as expected. Surprisingly, however, if we consider only elements involved in conversion events, actual conversion pairs were only slightly more similar to each other (at the root) than pairs of elements chosen at random from the recipient and donor pools, with an average difference of 55.2 sites. Though highly significant (p=0.001, t-test), a difference of only around a single base pair (9% of a standard deviation) suggests that sequence agreement is of relatively small importance to conversion probability among pairs of elements that are individually well-suited to be donors and recipients. This result does not appear driven by

sampling biases; in the long-tract subset, the average true pair differed by 53.9 sites at the root, while random pairs differed by 55.1, both very similar to the full set.

This degree of sequence similarity between identified conversion pairs in this dataset is considerably lower than reported for other identified conversion events in primate genomes, which are generally above 95%[83]. The high density of *Alu* elements offers many more opportunities for pairwise interaction than the typical gene family, so conversion events between low-similarity elements may sometimes occur even if such events are individually low likelihood.

Gene Conversion Outcomes

To identify the substitutions involved in each conversion event, the inferred sequence of each tract immediately after conversion (post-conversion sequence) was compared to the same region in the acceptor at the root of the Great Apes (pre-conversion sequence). As we cannot identify variants acquired in the converted region of the acceptor between orangutan divergence and the gene conversion event, we cannot compare the sequence in the acceptor immediately before and after conversion, and some base changes induced by the conversion event will be missed.

The average conversion event involved 21.4 identifiable substitutions (standard deviation 12) from the donor to the recipient element, around 18% of the average tract size. The average non-converting element experienced 3.7

substitutions along the terminal gorilla lineage. With an overall conversion rate of 0.96%, this implies that identified conversion events were responsible for approximately 5.3% of the substitutions in *Alu* in the gorilla lineage, though the actual number is likely somewhat higher due to unidentified events.

The different substitution types involved in conversion events are given in Table 5. Allelic gene conversion is known to be biased towards GC variants in a wide variety of species[100,101], and Aleshin and Zin[72] found a GC bias in *Alu* interlocus conversion as well. However, I do not observe such a bias overall. While A:T→C:G substitutions (i.e., A→C and its reverse complement) are more frequent than the reverse, as expected, A:T→G:C substitutions are less frequent than the reverse. This pattern appears largely, but not entirely, driven by CpG sites. If we separate out GpC:CpG→ApT:TpA substitutions from all other substitutions, I observe only a slight excess of other G:C→A:T substitutions relative to the reverse. The patterns in the long-tract subset are similar to the full set, indicating that they are likely not driven by sampling biases.

Substitution Type	Full Set Count	Full Set Reverse Count	Long-tract Subset Count	Long-Tract Subset Reverse Count
C:G→A:T	3409	3930	2554	2899
C:G→T:A	20180	17205	14736	12494
C:G→G:C	4152		3086	
A:T→T:A	3842		2859	
CpG:GpC→TpA:ApT	6561	3752	4897	2745
C:G→T:A (non-CpG)	13619	13453	9839	9749

Table 5: Counts of Substitution Types in Gene Conversion Events

Why do CpG sites show a strong AT bias, in contrast to the GC bias in gene conversion found in previous research[101]? Deamination of methylated CpG sites, which results in a GpC:CpG→ApT:TpA substitution, occurs at a much higher rate than other mutation types[102]. The *Alu* consensus has 23 of these hypermutable CpG sites, which tend to deplete rapidly with age. Recipient elements are more enriched in younger subfamilies than donor elements (Table 3) and had 9.6 intact CpG sites on average compared to 8.1 for donors at the root of the Great Apes. On the conversion tracts themselves, counting at the root, donors had 2.7 intact CpGs compared to 3.9 for recipients. Thus, the bias towards AT could be caused by the tendency for recipient elements in our dataset to be younger (and therefore more CpG-rich) than donors, perhaps because younger sequences are better binding targets for PRDM9, counteracting the general GC bias in conversion.

Conclusion

Using a novel algorithm for gene conversion detection among transposable elements, I identified 2514 interlocus gene conversion events among *Alu* elements in the gorilla genome, affecting around 1% of aligned elements among the Great Apes. The rate of conversion in the terminal gorilla branch vastly exceeded the conversion rate in the chimpanzee, human, or orangutan lineage. Conversion probability in the gorilla genome was strongly associated with a 15 bp motif within the *Alu* consensus sequence. These observations are consistent with a known

mechanism for gene conversion: recognition of a target motif by PRDM9[99,103], resulting in an induced double strand break that is repaired by a nearby homologous locus[79]. Due to the rapid rate of PRDM9 evolution[96], it is unsurprising that a high rate of gene conversion in *Alu* would be restricted to a single branch; targeting of the *Alu* consensus motif, as with any PRDM9 motif, would be a transient phenomenon.

It is surprising that PRDM9 would target the *Alu* consensus at all. It would seem a great risk to the host to target the *Alu* consensus sequence for double strand breaks, given the high frequency of *Alu* in the genome. *Alu* interlocus gene conversion itself is likely relatively harmless to the host, as non-replicative *Alu* elements appear to be largely neutral residents in the genome[24]; the fitness of the host is not expected to be strongly related to variation within *Alu* elements. However, double-strand breaks can also be repaired by nonallelic homologous recombination (NAHR)[104], which can lead to harmful genetic deletions and duplications.[36,105] A number of human diseases are associated with such *Alu*-induced genomic rearrangements[16]. Nevertheless, other PRDM9 targets have been identified in less common repeat elements; in particular, a modern human recombination hotspot appears to be targeted to a motif located in the inactive THE1 retrotransposon family and is strongest in this backgrou3nd.[99,106] McVean[107] suggested that the PRDM9 may tend to target motifs found in repetitive sequence to avoid targeting functional regions for double strand breaks, and this may overcome the cost of increased NAHR risk.

Gene conversion is in important force influencing the evolution of *Alu* elements.[54,85] The great disparity in conversion rate between the gorilla lineage and other branches suggests that much conversion in *Alu* may occur primarily in bursts, whenever motifs common to *Alu* elements are targeted for double-strand breaks, though it is unclear how often such targeting has occurred.

CHAPTER VII

CONCLUSION

With the advent of widespread genome sequencing, considerable scientific effort has been dedicated towards using this sequence data to better understand the structure, function and evolution of genomes. As transposable elements are a large component of many eukaryotic genomes, and appear to be an active player in many genomic processes[7], this project necessarily includes using TE sequence data to inform our understanding of the biology and evolution of TE families. The first step towards such an understanding is to identify the TEs within genomes and classify them into families, and, though not complete, considerable progress has been made towards this goal. TE identification and annotation methods[64,87,108–110] have revealed the ubiquity and diversity of TEs within the genomes of a wide variety of taxa. Having acquired large databases of TE sequences, the next task is to derive the evolutionary histories of these TE families and characterize the mechanisms by which they have evolved.

Despite the large quantity of sequence data available to infer TE evolutionary relationships, it is a difficult problem. Much variation among extant TEs comes

from post-insertion mutation, which are uninformative as to the evolutionary relationships among elements. Aside from these uninformative variants, many elements are identical or highly similar, making it difficult or impossible to work out precise evolutionary relationships among them. Addressing this problem is challenging on its own, but becomes much more so for a TE family such as *Alu* that experiences high rates of gene conversion. As gene conversion and transposition both involve replication of TE sequence, these processes are difficult to distinguish from one another: consider a locus that experiences conversion of the complete element, eliminating all information about the source it originally replicated from.

The major aim of this dissertation was to develop a framework for precise evolutionary inference in transposable elements, intending to address the particular difficulties with these sequences. To start, I developed, in Chapters II and III, a new interpretation of the "subfamily" concept and a method for subfamily classification based on this interpretation. To be usable for precise evolutionary inference, assignment of an element to a subfamily should represent a specific evolutionary claim that can be evaluated with evidence. Yet, despite widespread usage in nearly all TE analysis, usage of the subfamily concept has often been ambiguous about what exactly it means for an element to belong to a subfamily; in particular, in previous classification schemes it is unspecified how hybrid elements resulting from gene conversion between elements from different subfamilies should be classified. My definition of subfamilies as containing all elements with identical mutation-reversed sequences eliminates this ambiguity and allows for classification of hybrid

elements in a natural way, and the AnTE algorithm I developed provides the means for carrying out such a classification. Another important advance of AnTE is probabilistic classification, carried out under a Bayesian framework. Due to the similarity of many TE sequences, there is often considerable uncertainty in ancestry; any deterministic classification expresses too high a degree of confidence, which carries through to downstream analyses based on such a classification. Probabilistic subfamily assignment is therefore of great importance to precise TE evolutionary inference.

The result of an AnTE analysis is an estimate of the probability each element in a TE family sequence database belongs to each possible subfamily, from which an an expected frequency for each subfamily can be calculated. Importantly, the AnTE results themselves imply no claim about the origin of subfamilies, which could be from transposition, gene conversion, or a mix; we do not know enough about the dynamics of either process to make such claims with confidence for individual subfamilies. However, the overall pattern of subfamily frequencies can be used to test hypotheses about the evolutionary processes that may have generated such a pattern. In Chapter V, I used this reasoning to assess the role of gene conversion in producing observed patterns of *Alu* sequence diversity, finding that gene conversion appeared responsible for a "network" structure of *Alu* diversity, in which a large proportion of *Alu* elements appear to have sequences intermediate between that of major replicators. Though previous analyses[70,71,85] have reached qualitatively similar conclusions using ad hoc approaches, AnTE is a major advance by allowing

for robust, systematic and precise quantitative analyses of the role of gene conversion, replication and mutation in generating observed patterns of sequence diversity. We can use AnTE to develop a clearer picture, with higher confidence, and on a larger scale than previous approaches.

Ultimately, the purpose of improved TE evolutionary inference is to better understand the biology and evolutionary dynamics of TE families. Despite comprising a large portion of many eukaryotic genomes, including a majority of the human genome[4], fundamental aspects of TE evolution remain poorly understood. For example, even for well-studied TE families like *Alu* we lack a clear picture of the distribution of replicative activity across elements[29] and of the causes of succession between different subfamilies[111]. I believe that the methods presented here provide a useful advance towards answering these questions. Given accurate age and frequency estimates for TE subfamilies, models of TE replication dynamics can be constructed and evaluated based on how well they match these results. Precision in subfamily assignment and age estimation is important for accurate model comparison.

Perhaps the most important implication of the large TE content of many genomes comes from the tendency of homologous sequences to recombine, which can lead to major duplications and deletions of genomic sequence, including of functional sequence[36,105]; this is of great consequence both from an evolutionary[36] and medical[16] perspective. The methodology developed in Chapter VI for detection of gene conversion events, one consequence of homologous recombination, adds to

our understanding to recombination dynamics. As gene conversion between elements is a much less drastic effect than deletion or duplication, previous methods[72] have identified only signatures of the conversion process but not individual events. Exploiting the power of Bayesian phylogenetics, the TEConv methodology can identify not only individual gene conversion events on a particular branch of a phylogenetic tree, but also provide probabilistic estimates for nearly every aspect of that event. The burst of gene conversion events we identify among *Alu* elements on the gorilla lineage thus provides a powerful dataset for understanding the gene conversion process in primates.

A similar Bayesian phylogenetic approach could likely be applied to analyze TE-mediated deletion. Though previous analyses have already identified many apparent *Alu*-mediated deletions in the human genome[36], the Bayesian phylogenetic approach would allow for greater detail as to the causes and consequences of these events; i.e., by using phylogenetic information to infer the before and after state and the position at which the crossover took place. It would be of particular interest to determine whether the burst of gene conversion observed in gorilla was accompanied by a burst of *Alu*-mediated deletions. If so, the PRDM9 mechanism I identify may be an important factor in genomic instability through its interactions with transposable element families. As PRDM9 appears to only rarely target TE families such as *Alu* which are both young and large, and therefore highly susceptible to interlocus recombination, this could partly explain variation in the rate of structural changes[112] between lineages.

Throughout my dissertation, I have focused primarily on *Alu* elements, with the exception of the AnTE analysis of the gibbon LAVA family in Chapter IV. It is thus worth considering the generality of the methods presented to other TE families. Most of the ideas presented here are of broad applicability: the need for probabilistic evolutionary inference, for precise definition of the "subfamily" concept, for careful and systematic assessment of the role of gene conversion and replication in generating patterns of sequence diversity. However, different TE families present distinct challenges, which may require modification of the methods to address. For example, there are large numbers of full-length *Alu* elements in each Great Ape genome, so it was unnecessary in my analyses to consider analysis of fragmented elements, though this may be important for other families. Optimization of the AnTE algorithm to much larger elements, such as the LINEs, is an important avenue for future research.

REFERENCES

1. Tenaillon, M. I., Hufford, M. B., Gaut, B. S. & Ross-Ibarra, J. Genome Size and Transposable Element Content as Determined by High-Throughput Sequencing in Maize and Zea luxurians. *Genome Biol. Evol.* **3,** 219–229 (2011).
2. Kidwell, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115,** 49–63 (2002).
3. Chalopin, D., Naville, M., Plard, F., Galiana, D. & Volff, J.-N. Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. *Genome Biol. Evol.* **7,** 567–580 (2015).
4. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genet* **7,** e1002384 (2011).
5. Kidwell, M. G. & Lisch, D. R. Transposable elements and host genome evolution. *Trends Ecol. Evol.* **15,** 95–99 (2000).
6. Kazazian, H. H. Mobile Elements: Drivers of Genome Evolution. *Science* **303,** 1626–1632 (2004).
7. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18,** 71–86 (2017).
8. Volff, J.-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays* **28,** 913–922 (2006).
9. Nekrutenko, A. & Li, W.-H. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.* **17,** 619–621 (2001).

10. van de Lagemaat, L. N., Landry, J.-R., Mager, D. L. & Medstrand, P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* **19,** 530–536 (2003).
11. Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene expression. *Science* **351,** aac7247 (2016).
12. Jordan, I. K., Rogozin, I. B., Glazko, G. V. & Koonin, E. V. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* **19,** 68–72 (2003).
13. Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. *Mob. DNA* **7,** 9 (2016).
14. Franke, G. *et al.* Alu-Alu recombination underlies the vast majority of large VHL germline deletions: Molecular characterization and genotype–phenotype correlations in VHL patients. *Hum. Mutat.* **30,** 776–786 (2009).
15. Deininger, P. L. & Batzer, M. A. Alu Repeats and Human Disease. *Mol. Genet. Metab.* **67,** 183–193 (1999).
16. Belancio, V. P., Roy-Engel, A. M. & Deininger, P. L. All y'all need to know 'bout retroelements in cancer. *Semin. Cancer Biol.* **20,** 200–210 (2010).
17. Reilly, M. T., Faulkner, G. J., Dubnau, J., Ponomarev, I. & Gage, F. H. The Role of Transposable Elements in Health and Diseases of the Central Nervous System. *J. Neurosci.* **33,** 17577–17586 (2013).
18. Kriegs, J. O. *et al.* Retroposed Elements as Archives for the Evolutionary History of Placental Mammals. *PLOS Biol.* **4,** e91 (2006).

19. Antunez-de-Mayolo, G. *et al.* Phylogenetics of worldwide human populations as determined by polymorphic Alu insertions. *ELECTROPHORESIS* **23,** 3346–3356 (2002).

20. Salem, A.-H. *et al.* Alu elements and hominid phylogenetics. *Proc. Natl. Acad. Sci.* **100,** 12787–12791 (2003).

21. Arndt, P. F., Hwa, T. & Petrov, D. A. Substantial Regional Variation in Substitution Rates in the Human Genome: Importance of GC Content, Gene Density, and Telomere-Specific Effects. *J. Mol. Evol.* **60,** 748–763 (2005).

22. Arndt, P. F., Petrov, D. A. & Hwa, T. Distinct Changes of Genomic Biases in Nucleotide Substitution at the Time of Mammalian Radiation. *Mol. Biol. Evol.* **20,** 1887–1896 (2003).

23. Webster, M. T., Smith, N. G. C., Hultin-Rosenberg, L., Arndt, P. F. & Ellegren, H. Male-Driven Biased Gene Conversion Governs the Evolution of Base Composition in Human Alu Repeats. *Mol. Biol. Evol.* **22,** 1468–1474 (2005).

24. Cordaux, R., Lee, J., Dinoso, L. & Batzer, M. A. Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. *Gene* **373,** 138–144 (2006).

25. Deininger, P. L., Batzer, M. A., Hutchison III, C. A. & Edgell, M. H. Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8,** 307–311 (1992).

26. Shen, M. R., Batzer, M. A. & Deininger, P. L. Evolution of the master Alu gene(s). *J. Mol. Evol.* **33,** 311–320 (1991).

27. Price, A. L., Eskin, E. & Pevzner, P. A. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Res.* **14,** 2245–2252 (2004).

28. Bennett, E. A. *et al.* Active Alu retrotransposons in the human genome. *Genome Res.* **18,** 1875–1883 (2008).

29. Cordaux, R., Hedges, D. J. & Batzer, M. A. Retrotransposition of Alu elements: how many sources? *Trends Genet.* **20,** 464–467 (2004).

30. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3,** 370–379 (2002).

31. Ullu, E. & Tschudi, C. Alu sequences are processed 7SL RNA genes. *Nature* **312,** 171–172 (1984).

32. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* **35,** 41–48 (2003).

33. Wallace, N., Wagstaff, B. J., Deininger, P. L. & Roy-Engel, A. M. LINE-1 ORF1 protein enhances Alu SINE retrotransposition. *Gene* **419,** 1–6 (2008).

34. Aleshin, A. & Zhi, D. Recombination-Associated Sequence Homogenization of Neighboring Alu Elements: Signature of Nonallelic Gene Conversion. *Mol. Biol. Evol.* **27,** 2300–2311 (2010).

35. Roy, A. M. *et al.* Potential Gene Conversion and Source Genes for Recently Integrated Alu Elements. *Genome Res.* **10,** 1485–1495 (2000).

36. Sen, S. K. *et al.* Human Genomic Deletions Mediated by Recombination between Alu Elements. *Am. J. Hum. Genet.* **79,** 41–53 (2006).

37. Hoen, D. R. *et al.* A call for benchmarking transposable element annotation methods. *Mob. DNA* **6,** 13 (2015).

38. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6,** 11 (2015).

39. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0.* (2013).

40. Speir, M. L. *et al.* The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* **44,** D717–D725 (2016).

41. Willard, C., Nguyen, H. T. & Schmid, C. W. Existence of at least three distinct Alu subfamilies. *J. Mol. Evol.* **26,** 180–186 (1987).

42. Britten, R. J., Baron, W. F., Stout, D. B. & Davidson, E. H. Sources and evolution of human Alu repeated sequences. *Proc. Natl. Acad. Sci.* **85,** 4770–4774 (1988).

43. Jurka, J. & Smith, T. A fundamental division in the Alu family of repeated sequences. *Proc. Natl. Acad. Sci.* **85,** 4775–4778 (1988).

44. Jurka, J. & Milosavljevic, A. Reconstruction and analysis of human alu genes. *J. Mol. Evol.* **32,** 105–121 (1991).

45. Kapitonov, V. & Jurkal, J. The age of Alu subfamilies. *J. Mol. Evol.* **42,** 59–65 (1996).

46. Queiroz, K. de & Gauthier, J. Phylogenetic Taxonomy. *Annu. Rev. Ecol. Syst.* **23,** 449–480 (1992).

47. Wiley, E. O. The Evolutionary Species Concept Reconsidered. *Syst. Zool.* **27,** 17–26 (1978).

48. Zmasek, C. M. & Eddy, S. R. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* **17,** 821–828 (2001).

49. Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science* **294,** 2310–2314 (2001).

50. Kass, D. H., Batzer, M. A. & Deininger, P. L. Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. *Mol. Cell. Biol.* **15,** 19–25 (1995).

51. Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C. & Patrinos, G. P. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet. Lond.* **8,** 762–75 (2007).

52. Hastings, P. J. Mechanisms of Ectopic Gene Conversion. *Genes* **1,** 427–439 (2010).

53. Benovoy, D. & Drouin, G. Ectopic gene conversions in the human genome. *Genomics* **93,** 27–32 (2009).

54. Roy-Engel, A. M. *et al.* Non-traditional Alu evolution and primate genomic diversity1. *J. Mol. Biol.* **316,** 1033–1040 (2002).

55. Wacholder, A. C. *et al.* Inference of Transposable Element Ancestry. *PLoS Genet* **10,** e1004482 (2014).

56. VEMULAPALLI, V. DELINEATING THE EVOLUTIONARY DYNAMICS OF MUTATION AND SELECTION. (2012).

57. Gu, W. *et al.* SINEs, evolution and genome structure in the opossum. *Gene* **396,** 46–58 (2007).

58. Gilks, W. R., Richardson, S. & Spiegelhalter, D. *Markov Chain Monte Carlo in Practice*. (CRC Press, 1995).

59. Chib, S. & Greenberg, E. Understanding the Metropolis-Hastings Algorithm. *Am. Stat.* **49,** 327–335 (1995).

60. Zhu, J., Liu, J. S. & Lawrence, C. E. Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14,** 25–39 (1998).

61. Lunter, G. *et al.* Uncertainty in homology inferences: Assessing and improving genomic sequence alignment. *Genome Res.* **18,** 298–309 (2008).

62. Carbone, L. *et al.* Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513,** 195–201 (2014).

63. Carbone, L. *et al.* Centromere Remodeling in Hoolock leuconedys (Hylobatidae) by a New Transposable Element Unique to the Gibbons. *Genome Biol. Evol.* **4,** 760–770 (2012).

64. Gu, W., Castoe, T. A., Hedges, D. J., Batzer, M. A. & Pollock, D. D. Identification of repeat structure in large genomes using repeat probability clouds. *Anal. Biochem.* **380,** 77–83 (2008).

65. Britten, R. J. Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 6148–6150 (1994).

66. Liu, G. E., Alkan, C., Jiang, L., Zhao, S. & Eichler, E. E. Comparative analysis of Alu repeats in primate genomes. *Genome Res.* **19,** 876–885 (2009).

67. Kapitonov, V. & Jurkal, J. The age of Alu subfamilies. *J. Mol. Evol.* **42,** 59–65 (1996).

68. Marchani, E. E., Xing, J., Witherspoon, D. J., Jorde, L. B. & Rogers, A. R. Estimating the age of retrotransposon subfamilies using maximum likelihood. *Genomics* **94,** 78–82 (2009).

69. Arndt, P. F., Petrov, D. A. & Hwa, T. Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. *Mol. Biol. Evol.* **20,** 1887–1896 (2003).

70. Roy, A. M. *et al.* Potential Gene Conversion and Source Genes for Recently Integrated Alu Elements. *Genome Res.* **10,** 1485–1495 (2000).

71. Roy-Engel, A. M. *et al.* Non-traditional Alu evolution and primate genomic diversity1. *J. Mol. Biol.* **316,** 1033–1040 (2002).

72. Aleshin, A. & Zhi, D. Recombination-Associated Sequence Homogenization of Neighboring Alu Elements: Signature of Nonallelic Gene Conversion. *Mol. Biol. Evol.* **27,** 2300–2311 (2010).

73. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*. (2013).

74. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*. (2013).

75. Xing, J. *et al.* Alu Element Mutation Spectra: Molecular Clocks and the Effect of DNA Methylation. *J. Mol. Biol.* **344,** 675–682 (2004).

76. Kass, D. H., Batzer, M. A. & Deininger, P. L. Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. *Mol. Cell. Biol.* **15,** 19–25 (1995).

77. Taghian, D. G. & Nickoloff, J. A. Chromosomal double-strand breaks induce gene conversion at high frequency in mammalian cells. *Mol. Cell. Biol.* **17,** 6386–6393 (1997).

78. Benovoy, D. & Drouin, G. Ectopic gene conversions in the human genome. *Genomics* **93,** 27–32 (2009).

79. Hastings, P. J. Mechanisms of Ectopic Gene Conversion. *Genes* **1,** 427–439 (2010).

80. Duret, L. & Galtier, N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu. Rev. Genomics Hum. Genet.* **10,** 285–311 (2009).

81. Duret, L. & Arndt, P. F. The Impact of Recombination on Nucleotide Substitutions in the Human Genome. *PLoS Genet* **4,** e1000071 (2008).

82. Liao, D. Concerted Evolution: Molecular Mechanism and Biological Implications. *Am. J. Hum. Genet.* **64,** 24–30 (1999).

83. Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C. & Patrinos, G. P. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet. Lond.* **8,** 762–75 (2007).

84. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3,** 370–379 (2002).

85. Carroll, M. L. *et al.* Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity11Edited by J. Karn. *J. Mol. Biol.* **311,** 17–40 (2001).

86. SAWYER, S. GENECONV : a computer package for the statistical detection of gene conversion. *http://www.math.wustl.edu/ ~sawyer* (1999).

87. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0.* (2013).

88. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18,** 1814–1828 (2008).

89. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** D749–D755 (2014).

90. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17,** 368–376 (1981).

91. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662–D669 (2015).

92. Lam, I. & Keeney, S. Mechanism and Regulation of Meiotic Recombination Initiation. *Cold Spring Harb. Perspect. Biol.* **7,** a016634 (2015).

93. Parvanov, E. D., Petkov, P. M. & Paigen, K. Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science* **327,** 835–835 (2010).

94. Davies, B. *et al.* Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature* **530,** 171–176 (2016).

95. Úbeda, F. & Wilkins, J. F. The Red Queen theory of recombination hotspots. *J. Evol. Biol.* **24,** 541–553 (2011).

96. Schwartz, J. J., Roach, D. J., Thomas, J. H. & Shendure, J. Primate evolution of the recombination regulator PRDM9. *Nat. Commun.* **5,** 4370 (2014).

97. Stevison, L. S. *et al.* The Time Scale of Recombination Rate Evolution in Great Apes. *Mol. Biol. Evol.* **33,** 928–945 (2016).

98. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467,** 1099–1103 (2010).

99. Myers, S. *et al.* Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. *Science* **327,** 876–879 (2010).

100. Galtier, N. & Duret, L. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* **23,** 273–277 (2007).

101. Williams, A. L. *et al.* Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife* **4,** e04637 (2015).

102. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12,** 756–766 (2011).

103. Baudat, F. *et al.* PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science* **327,** 836–840 (2010).

104. Chen, J.-M., Cooper, D. N., Férec, C., Kehrer-Sawatzki, H. & Patrinos, G. P. Genomic rearrangements in inherited disease and cancer. *Semin. Cancer Biol.* **20,** 222–233 (2010).

105. Ade, C., Roy-Engel, A. M. & Deininger, P. L. Alu elements: an intrinsic source of human genome instability. *Curr. Opin. Virol.* **3,** 639–645 (2013).

106. Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.* **40,** ng.213 (2008).

107. McVean, G. What drives recombination hotspots to repeat DNA in humans? *Philos. Trans. R. Soc. B Biol. Sci.* **365,** 1213–1218 (2010).

108. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21,** i351–i358 (2005).

109. Wheeler, T. J. *et al.* Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.* **41,** D70–D82 (2013).

110. Bao, Z. & Eddy, S. R. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Res.* **12,** 1269–1276 (2002).

111. Han, K. *et al.* Under the genomic radar: The Stealth model of Alu amplification. *Genome Res.* **15,** 655–664 (2005).

112. Ventura, M. *et al.* Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res.* **21,** 1640–1649 (2011).