1-1-2010

Computational methods for the analysis of primate mobile elements

Richard Cordaux

Shurjo K. Sen

Miriam K. Konkel

Mark A. Batzer

Recommended Citation

Cordaux, R., Sen, S., Konkel, M., & Batzer, M. (2010). Computational methods for the analysis of primate mobile elements. *Methods in Molecular Biology*, *628*, 137-151. https://doi.org/10.1007/978-1-60327-367-1_8

Published in final edited form as:
Methods Mol Biol. 2010 ; 628: 137–151. doi:10.1007/978-1-60327-367-1_8.

Computational methods for the analysis of primate mobile elements

Richard Cordaux[1], **Shurjo K. Sen**[2], **Miriam K. Konkel**[2], and **Mark A. Batzer**[2]

[1]Laboratoire Ecologie, Evolution et Symbiose, CNRS UMR 6556, Université de Poitiers, 40 Avenue du Recteur Pineau, 86022 Poitiers, France.

[2]Department of Biological Sciences, Biological Computation and Visualization Center, Center for BioModular Multi-scale Systems, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA.

Abstract

Transposable elements (TE), defined as discrete pieces of DNA that can move from site to another site in genomes, represent significant components of eukaryotic genomes, including primates. Comparative genome-wide analyses have revealed the considerable structural and functional impact of TE families on primate genomes. Insights into these questions have come in part from the development of computational methods that allow detailed and reliable identification, annotation and evolutionary analyses of the many TE families that populate primate genomes. Here, we present an overview of these computational methods, and describe efficient data mining strategies for providing a comprehensive picture of TE biology in newly available genome sequences.

Keywords

Computational methods; transposable element; insertion; identification; classification; consensus sequence; subfamily; phylogenetic reconstruction; transpositional activity; primate; genome evolution

1. Introduction

Transposable elements (TE), defined as discrete pieces of DNA that can move from site to another site in genomes, have long been considered as non-significant components of genomes. This view started to change, however, when whole genome sequences became available. Hence, nearly half of the human genome is now recognized as being of TE origin (1). It is likely that this is an underestimate because some ancient TEs in the genome may have degraded beyond recognition by current methods. Primates constitute an excellent taxonomic group in which to analyze TE diversity and evolution because, in addition to humans, complete genome sequences of the chimpanzee and rhesus macaque are now available (2;3) with more genome sequences on the way. Comparative genome-wide analyses have revealed the considerable structural and functional impact of TE families on primate genomes.

The primary mode of TE-mediated instability is *de novo* integration of new elements, which can have a variety of functional consequences (4). However, additional changes in local sequence architecture arising as a by-product of TE activity include, but are not limited to, insertion-mediated deletions (5;6), recombination-mediated deletions (7;8), segmental

Corresponding author: Dr. Mark A. Batzer mbatzer@lsu.edu Tel: (225) 578-7102 Fax: (225) 578-7113 .

duplications (9;10), inversions (11;12) and inter- or intra-chromosomal transduction of host genomic sequence (13;14). Paradoxically, TE activity is not associated with genomic instability alone; retrotransposon mRNAs can also occasionally serve as molecular bandages for repairing potentially lethal DNA double-strand breaks (15;16). Another interesting aspect of TE biology in primate genomes has been the discovery that functions encoded by TEs originally for their own purposes can be efficiently adapted by host genomes into unrelated beneficial roles (17; 18). This process of so-called molecular domestication illustrates that TEs may on occasion share a mutualistic relationship with their host genomes, and that the “parasite” tag historically attached to TEs may be somewhat unfair in some cases.

In a broader sense, these observations raise the question of the nature of the host-TE relationship throughout evolution. A popular opinion is that within the evolutionary timescale of the primate radiation, most TE families have been slightly deleterious or at best neutral within the genome, and have achieved their high numbers through a finely tuned strategy of parasitism (19;20; 21). However, contrary to this viewpoint, various analyses have proposed different functional roles for some TE families, such as origins of replication, gene expression regulators, agents of DNA repair and X-chromosome inactivation or scaffolds for meiotic replication (22;23; 24). These views need not be reciprocally exclusive, and it may be overly simplistic to treat the interactions between TE families and primate genomes as being a zero-sum game. Indeed, a systems biology approach wherein interactions between host genomes and TEs are seen in the context of an ecosystem may be a suitable way of representing this complex relationship (25;26). In any event, addressing these questions requires exhaustive and reliable identification, annotation and evolutionary analyses of the many TE families that populate primate genomes. A number of computational methods have been developed to this end, which are reviewed in the following protocol.

2. Materials

Computational TE analyses can be performed on a local desktop machine with internet access. However, large-scale studies require a local software installation, typically in a UNIX environment (*see* Note 1) with considerable memory (preferably 4 GB, 16GB, or more RAM, depending on the study size). Common (bio-) computational skills should be sufficient for successful use and implementation of the required software.

3. Methods

3.1. TE identification

In this section, we describe methods to identify: (i) TEs for which prior sequence knowledge exists, (ii) TEs with no prior information available (i.e. *de novo* identification), and (iii) TEs which are differentially inserted among genomes (i.e. polymorphic for presence or absence).

3.1.1. Identification of known TEs

1. TE library: to identify known TEs in a target sequence, we rely on an existing TE library containing the consensus sequences (see section 3.2.2) of multiple TE families. The most comprehensive database of eukaryotic TEs is Repbase (http://girinst.org/) (27;28). Repbase can be searched for consensus sequences directly, or a desired library can be downloaded.

Note 1 While UNIX is typically stated as a requirement, many of these tools also work under the UNIX-based Macintosh OS X operating system, and also under Microsoft Windows with environments like Cygwin or MSYS.

2. Selection of genome sequences: human genomic sequences can be retrieved from UCSC (http://genome.ucsc.edu; select genomes and species of interest) (*see* Note 2).

3. TE annotation: using the selected TE library as reference, TEs in the query sequence are identified by similarity searches and annotated using RepeatMasker (http://repeatmasker.org) (*see* Note 3). Analysis of a relatively small data set can be performed online at http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker. For larger analyses (e.g. whole genomes), we suggest a local installation of RepeatMasker (http://www.repeatmasker.org/RMDownload.html) (*see* Note 4).

4. Submission of query sequences to RepeatMasker: RepeatMasker requires files to be in the FASTA format (*see* Note 5). Submission of several sequences at once is possible. There is no explicit maximum size constraint for query sequence(s). However, lengthy sequences often are slow to process, accompanied by the risk of an error message caused by connection time-out. The query sequence can be uploaded or pasted into the sequence window on the RepeatMasker web site. Select “Cross_match” as the search engine, and “slow” as the speed/sensitivity to ensure a search with the highest level of TE annotation (highest sensitivity; *see* Note 6). A DNA source is then selected, that determines the choice of TE library. We suggest selecting “Repetitive sequences in lower case” from the masking option bar to show the annotated repetitive sequence in the output file in lower case (*see* Notes [7], [8]).

5. Results: the output presents the annotation of repeats in the query sequence. The general output indicates what search options were selected; which (if any) and how many TEs are identified; what percentage of the query sequence contains TEs; and several result files that can be saved or reviewed in the web interface. The HTML version of the results gives detailed information about the identified TEs, including length, orientation, TE-subfamily, and matching region. Another important analysis output is the ID number(s) of the identified TEs. This indicates whether multiple TEs or a single element with interruptions have been identified (*see* Note 9). In addition, an alignment of the identified TE to the TE subfamily consensus sequence for which the sequence was identified as the best match is available.

3.1.2. *De novo* TE identification by genome self-alignment—*De novo* identification of repeats has proven challenging, especially for large and TE-rich genomes. So, a single dominant method for this task is not yet established. Commonly used software packages include

Note [2] The human genome can be in theory replaced by any other genome. If working with a genome for which a library does not exist and no analysis of TEs in a closely related species has been performed, *de novo* identification of TEs needs to be performed first to create a personal library for the species (see section 3.1.2.). Alternately, an analysis on the basis of protein similarities can be performed (e.g. see http://www.repeatmasker.org/cgi-bin/RepeatProteinMaskRequest). However, the latter approach does not detect TEs that lack typical protein structures, e.g. SINEs are not identified.

Note [3] The classic Repbase library is modified for RepeatMasker, in particular to improve the annotation of long TEs.

Note [4] Also required are: (i) a UNIX-based system with perl 5.8.0 or higher, (ii) either Cross_Match (obtained from http://www.phrap.org, select “Phred/Phrap/Consed”) or WU Blast (available from http://blast.wustl.edu/licensing/), and (iii) a TE library downloadable from http://www.girinst.org.

Note [5] FASTA is a text-based file format that represents nucleic acid or protein sequences and is characterized by a text description line beginning with > (no space between > and the text), followed by sequence in the next text line.

Note [6] Cross_match is described as more sensitive in identifying TEs compared to WU Blast.

Note [7] We also suggest that readers familiarize themselves with other options for possible integration within their analysis. These options are largely self-explanatory. In addition, the RepeatMasker documentation provides further detailed information.

Note [8] In principle, the same analysis can be performed with a local installation of RepeatMasker. The corresponding parameters can be selected from the command line.

Note [9] The ID information is important because long elements are particularly disposed to have multiple Ns (i.e. ambiguous or unsequenced bases) within their sequence boundaries (depending on the quality of the genome assembly), and many TEs may also be nested within other TEs. Using ID information, it can often be distinguished if the fragments of the TE belong to one or two separate insertions. While the ID information in most cases is accurate, we recommend checking this information manually if this information is of particular interest for the performed analysis.

PILER (29), ReAS (30), RECON (31), and RepeatScout (32). Below we describe the use of RepeatScout (http://repeatscout.bioprojects.org/) (*see* Note 10>):

1. Prerequisites: preferably, a computer with LINUX or UNIX and at least 4GB (ideally more) of RAM and a C compiler (typically freely available on UNIX machines) are needed.

2. Downloading and installing RepeatScout: RepeatScout_1.0.0 is available from http://repeatscout.bioprojects.org/. The software should be extracted and compiled with a command such as: tar –zxf RepeatScout-1.0.2.tar.gz; cd RepeatScout-1; make This yields two executable files: build_lmer_table and RepeatScout-v1.

3. Genome download: assembled genomes can be obtained from NCBI (ftp://ftp.ncbi.nih.gov/genomes) or UCSC (http://hgdownload.cse.ucsc.edu/downloads.html. For a full-genome analysis, download the chromFa.tar.gz file (*see* Note 11).

4. Repeat identification: first, an "l-mer" table is constructed; "l" (which defaults to 3) represents the length of the l-mer seeds and should be adjusted to meet the specific needs of the analysis. The following setting for l is suggested (*see* Note 12): ceil(log_4 (L)+1)with ceil(x) = smallest integer greater than x; log_4(x) = log base 4 of x; L: length of input sequence. A typical execution sequence to build an l-mer table begins with a command like: build_lmer_table –sequence source.fa –freq source.freq This calculates the frequency of l-mers in the specified source.fa DNA sequence. Next, an output file containing the *de novo* identified repeats is created. RepeatScout-v1 is executed with the built l-mer table (source.freq) and the sequence (source.fa) in the following manner: RepeatScout-v1 –sequence source.fa –freq source.freq –output repeats.fa

5. Filtering out non-TE sequences: repetitive elements include TEs as well as low-complexity elements, segmental duplications, or exons. Non-TE sequences may be filtered out with further processing. Low-complexity repeats may be removed with the perl script "filter-stage-1.prl." Next, RepeatMasker (see section 3.1.1) is run with the filtered RepeatScout-v1 library. "filter-stage-2.prl" excludes all repeats with very low copy numbers (default < 10). Lastly, segmental duplications and exons are identified and may be erased from the library by using the locations identified by RepeatMasker and matching them with gff files containing segmental duplications and exons.

3.1.3. *De novo* identification of polymorphic TEs by genome alignment to another genome

1. Preconditions: two genome sequences are required (*see* Note 13). This approach has been successfully implemented for human *Alu* (33) and LINE-1 (34) elements. A computer with the UNIX or LINUX operating system (or compatible variants) is needed. The user should be comfortable working at the command line. The ability to write programs in Perl, Python, and/or shell scripts is also valuable.

Note 10 RepeatScout requires assembled sequences, or at least scaffolds of a genome for the annotation of repeats. The assembly of new genomes, especially without general knowledge of the repeat composition, is challenging and may result in loss of repetitive sequences. ReAS is one of few programs for the *de novo* identification of TEs that requires whole shotgun reads and not assembled genomes.

Note 11 For many users, an analysis of a single or fractional chromosome per-run may be a present-day limit, given common RAM configurations and the RepeatScout v1 software itself. RepeatScout v1 does not provide intrinsic support for multiple CPUs; and its internal use of signed 4-byte integers limits runs to FASTA files with a maximum of 2 Gbp.

Note 12 A list of modifiable parameters, which usually do not need to be adjusted, can be found in the help file (--h) for RepeatScout.

Note 13 Alternately, sequence traces can be used with some procedural modifications; we highlight these in the notes of the appropriate sections.

2. Local installation of BLAST (Basic Local Alignment Search Tool) (35): BLAST, downloadable from ftp://ftp.ncbi.nih.gov/blast/, exists as a pre-compiled program suitable for many operating systems.

3. Selection and download of genomes: while we will provide a detailed description of this method for two human genomes, obtained from NCBI at ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ (*see* Note 14), in principle any two genomes can be used for this analysis. In our case, the first genome (hereafter genome A) is the human reference genome (ref_genome in NCBI). The second human genome (hereafter genome B) is the publicly available version of the Celera genome (alt_ genome in NCBI). 4) Download of TE consensus sequence: a TE consensus sequence of interest (here *Alu*) is downloaded from Repbase as a query sequence (*see* Note 15).

4. Identification of TEs and extraction of all matching TEs from genome A: genome A is queried with the *Alu* consensus sequence using the local installation of BLAST, and all candidate elements including 300 bp of flanking sequence on either side are extracted from genome A sequence.

5. Querying genome B with extracted loci from genome A: each extracted locus from genome A is used as a query sequence against genome B. If the query sequence matches in length and identity to a level of ≥ 98%, the locus is disqualified as a polymorphic candidate and discharged. In contrast, if either the *Alu* element alone or the flanking sequence is identified as a best match, the locus is a potential polymorphic candidate, and is used for a second, more detailed analysis. For the second analysis we take the *Alu* element out of the sequence and attach the flanking sequences to each other. This can be done with several BioPython commands such as:

   ```
   flankSize = 300 #choose a flanking sequence of 300bp
   seqSize = len(mySeq) #find the length of DNA sequence mySeq
   flankHead = mySeq[0:flankSize] #extract the head flanking portion
   flankTail = mySeq[seqSize-flankSize:seqSize] #extract the tail
   joinedSeq = flankHead + flankTail #assemble the two fragments together
   ```

   The flanking sequence of each locus is again queried against genome B to identify close-to-perfect matches of the flanking sequence. Close-to-perfect matches correspond to loci considered to contain polymorphic *Alu* elements present in genome A and absent in genome B. Other loci are discarded.

6. Identification of TEs from genome B absent in genome A: genome A is swapped with genome B, and steps 5 and 6 are repeated.

7. Comparison of confirmed polymorphic TEs to dbRIP: polymorphic human retrotransposons can be checked for novelty using the dbRIP database, a database of polymorphic human retrotransposons, by submitting the candidate loci to http://falcon.roswellpark.org:9090/searchRIP.html (36).

8. Confirmation of computational results: apart from a detailed manual confirmation of the data set, we recommend performing wet-bench PCR analyses on a panel of individual genomic DNA samples to confirm that the identified TEs are indeed polymorphic for insertion presence or absence (*see* **Ray et al., in this issue**).

Note 14Genomes of other species are also available from ftp://ftp.ncbi.nih.gov/genomes. Different versions of assembled reference genomes can be downloaded from UCSC (http://genome.ucsc.edu). To our knowledge the ref_genome is not available from UCSC.

Note 15Depending on the TE of interest, an approximately 50 bp-long conserved region of the TE may be used as a query sequence.

3.2. TE classification

In this section, we describe methods: (i) to classify TEs into groups of closely related copies (termed subfamilies), and (ii) to construct consensus sequences of TE subfamilies.

3.2.1. TE subfamily classification—

A transpositionally active TE in a genome can produce novel copies of itself, each of which is initially identical in nucleotide sequence to the copy that generated it. Therefore, any sequence feature present in the ancestral TE copy will be shared with its "progeny". TE subfamilies are thus defined as collections of TE copies exclusively sharing diagnostic sequence features. Such features typically include nucleotide substitutions located at homologous sites in all copies within a subfamily, termed "shared sequence variants" (SSV). SSVs are distinguishable from post-insertional random substitutions, which would show no site preference. Efficient SSV identification forms the basis for computational classification of TE copies into discrete subfamilies. A schematic algorithm for this purpose is described below:

1. Generation of a multiple alignment of TE copies of interest: this can be achieved by running the ClustalW alignment program (*see* Note 16), using a fasta file of the TE sequences as input. Visually inspect the alignment and make further refinements using a suitable sequence alignment editor, such as BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) or Megalign (http://www.dnastar.com/products/lasergene.php). The alignment forms the input for the algorithms mentioned in the next step.
2. Automated TE subfamily classification: to the best of our knowledge, only two specialized algorithms exist for this purpose: (a) MASC (Multiple Aligned Sequence Classification) (37) hierarchically and recursively splits the multiple alignment into smaller groups of two, continuing till the absence of multiple SSVs invalidates further subdivision. Although MASC is not currently available as a binary distribution, the original algorithm has been described in detail elsewhere (38) and reasonable competence with bioinformatics programming should enable users to adapt it for their specific analyses. (b) A second approach would be to use a modification of the MULTIPROFILER algorithm (39) to scan the multiple alignment for groups of TEs characterized by overrepresented *n*-tuples of SSVs (where n has an integral value >1), followed by a final step where subfamilies differing from their closest relatives by a single SSV are identified using a probability-based approach. Although this approach has till now been used for the construction of consensus sequences only for the *Alu* family (40), a set of Perl and C programs is available at http://www-cse.ucsd.edu/groups/bioinformatics/software.html#alu-subfam), that should in principle be modifiable for other TE families.

3.2.2. Construction of TE subfamily consensus sequences—

Over time, TE copies of a "source" TE for any particular subfamily each accumulate random substitutions, and for even moderately old subfamilies, individual members may be quite different from the original source TE. However, the same random nature of these substitutions means that, for any particular subfamily, most elements will retain the original nucleotide of the ancestral TE copy at individual positions along the length of the TE. Thus, by using a majority-rule algorithm that also accounts for increased mutation frequencies at CpG dinucleotides (i.e., wherever a C is followed by a G in 5′ to 3′ orientation), it is possible to accurately reconstruct the ancestral sequence that gave rise to the members of a TE subfamily. We describe a schematic algorithm below:

[Note 16] ClustalW is available as a command line interface or as a graphical user interface (ClustalX), downloadable at ftp://ftp.ebi.ac.uk/pub/software/clustalw2/. It is also implemented in biological sequence analyses software, such as BioEdit.

1. Construct a multiple alignment of TE copies grouped together as a subfamily (see section 3.2.1): quality of the multiple alignment will directly influence the accuracy of the reconstructed consensus sequences, and manual curation of the initial computational alignment will almost always result in a better finished product. Higher numbers of copies in the alignment will result in a consensus sequence with greater statistical support.

2. For each position, determine the majority nucleotide. Most multiple alignment software suites allow this to be done in a few clicks (e.g., in BioEdit, click alignment > positional frequency summary, or in MegAlign, click view > alignment report).

3. CpG dinucleotides have sixfold higher mutation rates compared to other dinucleotides, mostly through transitions at one of the two positions leading to either CpA or TpG (41). However, post-insertional substitutions mimicking CpA or TpG dinucleotides present in the ancestral consensus sequence can be sorted out on the basis of the proportion of subfamily members that carry a particular dinucleotide. If the ancestral state was either CpA or TpG, most copies will retain this state and the consensus sequence will tend to be unequivocal. If, however, a CpG in the original consensus sequence mutates to CpA or TpG, the ancestral and derived states will be present in almost equal proportions, and the resulting ambiguity at the dinucleotide position can be used to correct the consensus sequence to CpG.

4. The accuracy of the consensus sequence reconstructed using the above two steps can be tested using the following formula: $S = S_1S_2 + (1 - S_1)(1-S_2)/3$, where S_1 and S_2 represent sequence similarities between TE elements 1 and 2 of a particular family and the reconstructed source element, and S represents the mutual sequence similarity between the two copies (42). Close correspondence between the observed and expected values of S indicates that the consensus sequence is an accurate reconstruction (43) (*see* Note 17).

3.3. Analyses of TE evolution

To decipher the evolutionary history of TE subfamilies and address questions about e.g. their timing of transpositional activity, several approaches can be used. For example, very recently active TEs are expected to exhibit differential distribution among individuals, i.e. individual copies will be polymorphic for presence or absence at orthologous genomic sites among the compared samples. The method described in section 3.1.3 allows identification of such differentially inserted TE loci. TE insertions that are responsible for genetic disorders are examples of active subfamilies for which copies have inserted in the genome within the recent past. At a deeper timescale, TE subfamilies that have been active at different evolutionary periods are also expected to be differentially inserted among species. The timing of subfamily activity can thus be deduced from the timing of divergence of the host genomes that carry or lack copies of the TE subfamily of interest (44). In this section, we describe further computational approaches: (i) to estimate the age (i.e. the timing of transpositional activity) of TE subfamilies independently of the genomic location of the copies, and (ii) to infer TE amplification dynamics by reconstructing phylogenetic relationships among members of TE subfamilies.

3.3.1. Inference of TE subfamily ages—Because a subfamily consensus sequence (as obtained in section 3.2.2) represents the putative sequence of the active TE copy that gave rise to other copies in the subfamily, and because individual copies gradually diverge from the

Note 17For subfamilies with relatively recent periods of activity, individual copies will be similar to the consensus sequence; however, for older repeats individual members are usually far more divergent, and a well-constructed subfamily consensus sequence is the only suitable query for computational data mining.

“source” copy across time, the quantity of sequence divergence accumulated by individual copies relative to their reconstructed consensus sequence can be used to infer the approximate age of the TE subfamily, provided that the substitution rate is known for the lineage being investigated.

Average sequence divergence of individual copies to their consensus sequence can be obtained by creating a multiple alignment containing TE copies from a subfamily together with the subfamily consensus sequence. Pairwise genetic distances between the consensus sequence and each individual copy are calculated, and then averaged. Such calculations can be performed with various software packages for evolutionary and phylogenetic analyses, such as MEGA (45) (*see* Note 18):

1. Open a fasta alignment with the text editor implemented in MEGA and convert the alignment to the MEGA format (containing a .meg extension). The converted file can then be opened with the data analyses module of MEGA.
2. Create a group containing the consensus sequence and another group containing all individual subfamily copies: click Data > Setup/Select taxa & groups
3. Calculate average divergence between the two groups: click Distances > Compute Between Groups Means
4. Subfamily age is calculated as the average divergence to consensus divided by the substitution rate (*see* Note 19).

3.3.2. Phylogenetic analyses—Phylogenetic analyses can be performed to infer the relationships between individual copies within a subfamily and explore subfamily amplification dynamics. Several major methods of tree reconstruction are available, that differ in their underlying philosophy, including distance-, parsimony- and probability-based methodologies. Each method has its own advantages and drawbacks, and no single method is the best for all analyses. A number of software suites are available to conduct phylogenetic analyses, including MEGA. A comprehensive list of phylogenetic packages available for download or usable via a web interface can be found at http://evolution.genetics.washington.edu/phylip/software.html. Phylogenetic reconstruction starts with a multiple alignment of the TE copies of interest, which is achieved as described in section 3.2.2. The alignment is then used for tree reconstruction. For example, in MEGA, multiple phylogeny algorithms are available by clicking Phylogeny > Construct Phylogeny. Alternatively, for datasets with low sequence divergence, higher phylogenetic resolution may be reached by using network phylogenetic approaches (46;47). Several programs for reconstructing networks are available, such as NETWORK (48) (*see* Note 20).

Acknowledgments

Our research is supported by National Science Foundation BCS-0218338 (MAB) and EPS-0346411 (MAB), National Institutes of Health RO1 GM59290 (MAB) and PO1 AG022064, and the State of Louisiana Board of Regents Support Fund (MAB). RC is supported by a Young Investigator ATIP award from the Centre National de la Recherche Scientifique (CNRS).

Note 18 Freely available for download at http://www.megasoftware.net/

Note 19 Alternatively, the age of a subfamily can be estimated without reconstructing a subfamily consensus sequence. This can be achieved by calculating the average divergence between any two copies of the subfamily (in MEGA, open a .meg file containing an alignment of individual TE copies of interest and click Distances > Compute Overall Mean). Assuming that divergence has accumulated at the same rate among copies, approximate subfamily age can be estimated as half the average divergence divided by the substitution rate.

Note 20 Freely available for download at http://www.fluxus-engineering.com/netwinfo.htm. NETWORK requires a specific file format (containing an .rdf extension) that can be created manually using a text editor or automatically by converting a fasta file into .rdf format using a program available for purchase from the NETWORK website.

6. References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921. [PubMed: 11237011]
2. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 2005;437:69–87. [PubMed: 16136131]
3. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, et al. Evolutionary and biomedical insights from the rhesus macaque genome. Science 2007;316:222–34. [PubMed: 17431167]
4. Hedges DJ, Deininger PL. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat Res 2007;616:46–59. [PubMed: 17157332]
5. Callinan PA, Wang J, Herke SW, Garber RK, Liang P, Batzer MA. Alu Retrotransposition-mediated Deletion. J Mol Biol 2005;348:791–800. [PubMed: 15843013]
6. Han K, Sen SK, Wang J, Callinan PA, Lee J, Cordaux R, et al. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. Nucleic Acids Res 2005;33:4040–52. [PubMed: 16034026]
7. Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, et al. Human genomic deletions mediated by recombination between Alu elements. Am J Hum Genet 2006;79:41–53. [PubMed: 16773564]
8. Han K, Lee J, Meyer TJ, Wang J, Sen SK, Srikanta D, et al. Alu recombination-mediated structural deletions in the chimpanzee genome. PLoS Genet 2007;3:1939–49. [PubMed: 17953488]
9. Bailey JA, Liu G, Eichler EE. An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet 2003;73:823–34. [PubMed: 14505274]
10. Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV. Duplication, coclustering, and selection of human Alu retrotransposons. Proc Natl Acad Sci U S A 2004;101:1268–72. [PubMed: 14736919]
11. Lobachev KS, Stenger JE, Kozyreva OG, Jurka J, Gordenin DA, Resnick MA. Inverted Alu repeats unstable in yeast are excluded from the human genome. Embo J 2000;19:3822–30. [PubMed: 10899135]
12. Stenger JE, Lobachev KS, Gordenin D, Darden TA, Jurka J, Resnick MA. Biased distribution of inverted and direct Alus in the human genome: implications for insertion, exclusion, and genome stability. Genome Res 2001;11:12–27. [PubMed: 11156612]
13. Pickeral OK, Makalowski W, Boguski MS, Boeke JD. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res 2000;10:411–5. [PubMed: 10779482]
14. Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA. Emergence of primate genes by retrotransposon-mediated sequence transduction. Proc Natl Acad Sci U S A 2006;103:17608–13. [PubMed: 17101974]
15. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, et al. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet 2002;31:159–65. [PubMed: 12006980]
16. Sen SK, Huang CT, Han K, Batzer MA. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. Nucleic Acids Res 2007;35:3741–51. [PubMed: 17517773]
17. Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 2000;403:785–9. [PubMed: 10693809]
18. Cordaux R, Udit S, Batzer MA, Feschotte C. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci U S A 2006;103:8101–6. [PubMed: 16672366]
19. Boissinot S, Entezam A, Furano AV. Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol 2001;18:926–35. [PubMed: 11371580]
20. Cordaux R, Lee J, Dinoso L, Batzer MA. Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. Gene 2006;373:138–44. [PubMed: 16527433]
21. Schmid CW. Alu: a parasite’s parasite? Nat Genet 2003;35:15–6. [PubMed: 12947404]
22. Brosius J, Gould SJ. On “genomenclature”: a comprehensive (and respectful) taxonomy for pseudogenes and other “junk DNA”. Proc Natl Acad Sci U S A 1992;89:10706–10. [PubMed: 1279691]

23. Liu WM, Chu WM, Choudary PV, Schmid CW. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. Nucleic Acids Res 1995;23:1758–65. [PubMed: 7784180]

24. Schmid CW. Does SINE evolution preclude Alu function? Nucleic Acids Res 1998;26:4541–50. [PubMed: 9753719]

25. Brookfield JF. The ecology of the genome - mobile DNA elements and their hosts. Nat Rev Genet 2005;6:128–36. [PubMed: 15640810]

26. Le Rouzic A, Dupas S, Capy P. Genome ecosystem and transposable elements species. Gene 2007;390:214–20. [PubMed: 17188821]

27. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005;110:462–7. [PubMed: 16093699]

28. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 2006;7:474. [PubMed: 17064419]

29. Edgar RC, Myers EW. PILER: identification and classification of genomic repeats. Bioinformatics 2005;21(Suppl 1):i152–8. [PubMed: 15961452]

30. Li R, Ye J, Li S, Wang J, Han Y, Ye C, et al. ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. PLoS Comput Biol 2005;1:e43. [PubMed: 16184192]

31. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 2002;12:1269–76. [PubMed: 12176934]

32. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics 2005;21(Suppl 1):i351–8. [PubMed: 15961478]

33. Wang J, Song L, Gonder MK, Azrak S, Ray DA, Batzer MA, et al. Whole genome computational comparative genomics: A fruitful approach for ascertaining Alu insertion polymorphisms. Gene 2006;365:11–20. [PubMed: 16376498]

34. Konkel MK, Wang J, Liang P, Batzer MA. Identification and characterization of novel polymorphic LINE-1 insertions through comparison of two human genome sequence assemblies. Gene 2007;390:28–38. [PubMed: 17034961]

35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10. [PubMed: 2231712]

36. Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat 2006;27:323–9. [PubMed: 16511833]

37. Milosavljevic, A.; Haussler, D.; Jurka, J. Informed parsimonious inference of prototypical genetic sequence. In: Rivest, R.; Haussler, D.; Warmuth, MK., editors. Proceedings of the Second Annual Workshop on Computational Learning Theory; San Mateo. Morgan Kaufman; 1989. p. 102-117.

38. Milosavljevic, A. Categorization of Macromolecular Sequences by Minimal Length Encoding. University of California at Santa Cruz; 1990.

39. Keich U, Pevzner PA. Finding motifs in the twilight zone. Bioinformatics 2002;18:1374–81. [PubMed: 12376382]

40. Price AL, Eskin E, Pevzner PA. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Res 2004;14:2245–52. [PubMed: 15520288]

41. Xing J, Hedges DJ, Han K, Wang H, Cordaux R, Batzer MA. Alu element mutation spectra: molecular clocks and the effect of DNA methylation. J Mol Biol 2004;344:675–82. [PubMed: 15533437]

42. Jurka, J. Approaches to identification and analysis of interspersed repetitive DNA sequences. In: Adams, MD.; Fields, C.; Venter, JC., editors. Automated DNA Sequencing and Analysis. Academic Press; London: 1994. p. 294-298.

43. Smit AF, Toth G, Riggs AD, Jurka J. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol 1995;246:401–417. [PubMed: 7877164]

44. Pace JK 2nd, Feschotte C. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res 2007;17:422–32. [PubMed: 17339369]

45. Kumar S, Tamura K, Nei M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Briefings in Bioinformatics 2004;5:150–163. [PubMed: 15260895]

46. Posada D, Crandall KA. Intraspecific gene genealogies: trees grafting into networks. Trends In Ecology And Evolution 2001;16:37–45. [PubMed: 11146143]

47. Cordaux R, Hedges DJ, Batzer MA. Retrotransposition of Alu elements: how many sources? Trends Genet 2004;20:464–7. [PubMed: 15363897]

48. Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 1999;16:37–48. [PubMed: 10331250]