Databases and ontologies Advance Access publication August 31, 2011

AluHunter: a database of potentially polymorphic *Alu* insertions for use in primate phylogeny and population genetics

Christina M. Bergey
Department of Anthropology, New York University, New York, NY 10003, USA

Associate Editor: David Posada

ABSTRACT

Summary: *AluHunter* is a database of taxon-specific primate *Alu* elements for use in phylogeny and population genetics. The software automatically isolates potentially polymorphic *Alu* insertions in sequences submitted to GenBank by screening the elements against reference genomes. The resultant database of variable markers is a valuable resource for researchers interested in characterizing *Alu* elements in their primate taxon of interest.

Availability and Implementation: The *AluHunter* database can be accessed at http://www.aluhunter.com.

Contact: cmb433@nyu.edu

Received on April 15, 2011; revised on August 8, 2011; accepted on August 23, 2011

1 INTRODUCTION

Alu elements are a class of primate-specific short interspersed elements (SINEs) that replicate through the genome via a ‘copy and paste’ mechanism. Inserted *Alu* elements are rarely precisely excised and always ancestrally absent, making them nearly free of homoplasy, of known polarity, and thus ideal markers for inferring phylogeny (Hillis, 1999; Ray *et al*., 2006). If two taxa share a retrotransposon insertion at a certain genomic location, it is most likely due to shared descent. As a result, one can easily infer the phylogeny of a group of primates using presence or absence data for a sufficient quantity of *Alu* elements by grouping via Dollo parsimony those that share a particular *Alu* insertion to the exclusion of the taxa that do not. Their use has allowed the resolution of certain phylogenetic problems within primates that traditional DNA sequence-based or morphology-based methods had failed to decisively resolve, such as the placement of tarsiers within Haplorhines (Schmitz *et al*., 2001).

Determining whether an *Alu* known to be present in one taxon is present in another can be easily accomplished with the standard laboratory techniques. Given amplification primers flanking a known *Alu* element, a researcher need only to amplify the region with PCR and use gel electrophoresis to determine the amplicon’s size. Sequencing can then confirm whether the *Alu* is present in an organism’s genome, since relying on size estimates alone is prone to misinterpretation of rare shifted parallel insertions of *Alu* elements (Schmitz *et al*., 2001). Unfortunately, the process of finding novel, phylogenetically informative *Alu* elements—ones already known to be absent in other taxa—can take weeks of lab work, a process that can benefit from an efficient *in silico* solution.

Because of their prevalence in primate genomes, *Alu* elements are often unintentionally sequenced, meaning sequence repositories such as GenBank are sources of millions of non-human primate *Alu* elements. However, most of these insertions are useless for purposes of phylogenetic or population genetic analysis, because they are not variable at the generic or subgeneric levels. The human genome, for example, contains over 1 million *Alu* elements, but <0.5% are polymorphic within the species (Roy-Engel *et al*., 2001). Isolating useful *Alu* elements for researchers of primate phylogeny or population genetics requires the determination of whether an *Alu* was inserted before the diversification of a taxon of interest and is therefore fixed, or was inserted recently and is therefore potentially polymorphic. One simple bioinformatic solution is to search a sister taxon’s genome for the *Alu* element’s flanks. If the flanks are found in a sister taxon’s genome with no *Alu* element in between, one can assume that the *Alu* was inserted since the split with the sister taxon and is therefore potentially polymorphic within the taxon of interest. This method has only recently become possible with the availability of a large amount of source DNA sequences in which to find *Alu* elements and the advent of multiple, publicly available primate genome sequences against which to screen them.

AluHunter uses this screening process to bioinformatically isolate polymorphic insertions on a large scale. The project is an attempt to broadly characterize all *Alu* elements in GenBank sequences and isolate those that may be informative at the generic or subgeneric level.

2 METHODS

AluHunter consists of a suite of scripts written in Perl and Python that automate the process of isolating polymorphic *Alu* elements. *AluHunter* automatically retrieves novel non-human primate sequences from GenBank. It then identifies *Alu* elements in the sequences using RepeatMasker (Smit, *et al*. (1996–2010), http://repeatmasker.org), and stores their source information, sequences and 200 bp long flanking sequences in a database. Each *Alu* is then screened against one or more closely related genomes, via a BLASTN search of the genome for the *Alu* element’s flanking sequences. If the flanks are found in the genome, with a gap in between that is within 10% of the original *Alu* element’s size, the *Alu* is considered to be present or fixed in both the taxon of interest and the genome. Alternatively, if the flanks are found in the BLAST search with no gap or a small gap of <10% of the original *Alu* element’s size between them, the *Alu* is considered to be absent in the genome and therefore potentially polymorphic in the taxon of interest. The use of fuzzy size matching is to allow for sequence variation between taxa. Flanks that have multiple close matches in the genome—possibly indicating the presence of repetitive elements within the flanks—are removed from further analysis. Finally, information on polymorphic *Alu* elements is uploaded to an online MySQL database, which is accessible from http://www.aluhunter.com via a front end written in PHP.

On the website, a user can select his or her genus or genera of interest, and the site will display available information on *Alu* elements present in the

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

taxon that have been found to be absent in a closely related genome. Each *Alu* returned has a link to the source sequence in GenBank. Either singularly or in batch, users can download the *Alu* element sequences and flanking regions in FASTA format as well as primer sequences designed with Primer3 for typing the *Alu* elements (Rozen & Skaletsky, http://frodo.wi.mit.edu/primer3/). The primer sequences have been screened against a mispriming library to avoid designing primers that would match to repetitive DNA, and they can be tested against several primate genomes via linkage to UCSC's *in silico* PCR website (http://genome.ucsc.edu/). Users can also easily BLAST the *Alu* with or without flanks against genomes or GenBank via interfaces with NCBI's website.

As an example, a user interested in determining the evolutionary relationships of baboons (genus *Papio*) could select the genus and download primers to test for the presence of >3000 *Alu* elements present in baboons and absent in the rhesus macaque genome. PCR on samples from various baboon species using these primers can be visualized with gel electrophoresis. Sequencing the amplicons will confirm if the *Alu* element is present or absent. Though some will be fixed in baboons, other *Alu* elements will group some species or populations to the exclusion of others. A Dollo parsimony analysis can then be used on these binary characters to reconstruct the evolutionary history of the taxa. When tested in multiple populations, the *Alu* elements that are variable within a species can be used as markers in population genetic studies.

The algorithm is similar to that of Farwick *et al*. (2006), but differs in that its goal is identifying recently inserted Alu elements that are informative at the subgenus level. Since conservation of flanking regions for PCR primer design is more likely for closely related taxa, *AluHunter* does not have stringent requirements for conservation of flanks above the necessity of successful BLAST searches in the outgroup genome. This allows for the inclusion of unannotated source sequences and greatly increases the number of *Alu* elements found.

3 RESULTS

As of August 2011, there are 1 963 554 *Alu* elements present in 66 primate genera in the *AluHunter* database. These have been selectively screened against nine primate genomes, resulting in the isolation of 42 055 *Alu* elements (2.1% of the total) that are polymorphic at the generic or subgeneric level. These include:

- 31 785 *Alu* elements in *Nomascus* (gibbons) that are absent in the great ape genomes (*Homo sapiens, Pan troglodytes, Gorilla gorilla or Pongo pygmaeus*),
- 2417 *Alu* elements in *Papio* (baboons) that are absent in the rhesus macaque genome (*Macaca mulatta*),
- 2645 *Alu* elements in *Saimiri* (squirrel monkeys) that are absent in the marmoset genome (*Callithrix jacchus*),
- 2687 *Alu* elements in *Colobus* (colobus monkeys) that are absent in the rhesus macaque genome (*Macaca mulatta*),
- 1106 *Alu* elements in *Chlorocebus* (vervet monkeys) that are absent in the rhesus macaque genome (*Macaca mulatta*) and
- 526 *Alu* elements in *Callicebus* (titi monkeys) that are absent in the marmoset genome (*Callithrix jacchus*).

Ultimately, the *AluHunter* database aspires to be a classification of all *Alu* elements ever sequenced and deposited in GenBank. Despite its simple algorithm, its utility derives from the scale that its degree of automation allows. The database grows with GenBank, and will have increasing resolution as more genomes are added. The *Alu* elements isolated by the *AluHunter* program and available in its database are a useful resource for researchers studying the phylogenetic relationships or population genetics of primates.

ACKNOWLEDGEMENTS

The author would like to thank the NYU Molecular Anthropology Laboratory, especially Dr Andrew Burrell, Dr Todd Disotell, Jason Hodgson and Luca Pozzi for advice.

Funding: National Science Foundation Graduate Research Fellowship.

Conflict of Interest: none declared.

REFERENCES

Farwick,A. *et al*. (2006) Automated scanning for phylogenetically informative transposed elements in rodents. *Syst. Biol.*, **55**, 936–948.

Hillis,D.M. (1999) SINEs of the perfect character. *Proc. Natl Acad. Sci. USA*, **96**, 9979–9981.

Ray,D.A. *et al*. (2006) SINEs of a nearly perfect character. *Syst. Biol.*, **55**, 928–935.

Roy-Engel,A.M. *et al*. (2001) *Alu* insertion polymorphisms for the study of human genomic diversity. *Genetics*, **159**, 279–290.

Schmitz,J. *et al*. (2001) SINE insertions in cladistic analyses and the phylogenetic affiliations of Tarsius bancanus to other primates. *Genetics*, **157**, 777–784.

Smit,A.F.A. *et al*. (1996–2010) RepeatMasker Open 3.0. Available at http://www.repeatmasker.org.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016