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

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.


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 homoplasies, 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 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
As of August 2011, there are 1,963,554 Alu elements in the isolation of 42,055 selectively screened against nine primate genomes, resulting in 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),
- 2465 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


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