Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Gene 365 (2006) 11 – 20

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Whole genome computational comparative genomics: A fruitful approach for ascertaining *Alu* insertion polymorphisms

Jianxin Wang [a,1], Lei Song [a,1], M. Katherine Gonder [b], Sami Azrak [a], David A. Ray [c], Mark A. Batzer [c], Sarah A. Tishkoff [b], Ping Liang [a,*]

[a] *Department of Cancer Genetics, Roswell Park Cancer Institute, Elm and Carlton Streets, Buffalo, NY 14263, USA*
[b] *Department of Biology, University of Maryland, College Park, MD 20742, USA*
[c] *Department of Biological Sciences, Biological Computational and Visualization Center, Center for BioModular Multi-scale Systems, Louisiana State University, Baton Rouge, LA 70803, USA*

Received 28 April 2005; received in revised form 20 June 2005; accepted 7 September 2005
Available online 10 January 2006

Abstract

Alu elements are the most active and predominant type of short interspersed elements (SINEs) in the human genome. Recently inserted polymorphic (for presence/absence) *Alu* elements contribute to genome diversity among different human populations, and they are useful genetic markers for population genetic studies. The objective of this study is to identify polymorphic *Alu* insertions through an in silico comparative genomics approach and to analyze their distribution pattern throughout the human genome. By computationally comparing the public and Celera sequence assemblies of the human genome, we identified a total of 800 polymorphic *Alu* elements. We used polymerase chain reaction-based assays to screen a randomly selected set of 16 of these 800 *Alu* insertion polymorphisms using a human diversity panel to demonstrate the efficiency of our approach. Based on sequence analysis of the 800 *Alu* polymorphisms, we report three new *Alu* subfamilies, Ya3, Ya4b, and Yb11, with Yb11 being the smallest known *Alu* subfamily. Analysis of retrotransposition activity revealed Yb11, Ya8, Ya5, Yb9, and Yb8 as the most active *Alu* subfamilies and the maintenance of a very low level of retrotransposition activity or recent gene conversion events involving S subfamilies. The 800 polymorphic *Alu* insertions are characterized by the presence of target site duplications (TSDs) and longer than average polyA-tail length. Their pre-integration sites largely follow an extended "NT-AARA" motif. Among chromosomes, the density of *Alu* insertion polymorphisms is positively correlated with the *Alu*-site availability and is inversely correlated with the densities of older *Alu* elements and genes.

Keywords: Retrotransposition; Polymorphism; Bioinformatics; Comparative genomics

1. Introduction

Alu elements are the predominant type of short interspersed elements (SINEs) in the human genome, with over 1 million copies comprising ~10% of the total genome (Houck et al., 1979; Lander et al., 2001; Venter et al., 2001). The origin and amplification of *Alu* elements are evolutionarily recent events that coincided with the radiation of primates (Batzer and Deininger, 2002; Kapitonov and Jurka, 1996; Quentin, 1988; Shaikh and Deininger, 1996). *Alu* elements increase in number by retrotransposition, a process involving the insertion of reverse transcribed DNAs of *Alu*-derived transcripts back into the genome, apparently by hijacking the L1 retrotranspotion machinery (Boeke, 1997; Cost and Boeke, 1998; Dewannieux et al., 2003). Based on a hierarchical series of sequence mutations, *Alu* elements are classified into three major families designated as J, S, and Y, representing the oldest, intermediate, and youngest *Alu* sequences, respectively, and each of these families is further divided into one or more levels of subfamilies based on subfamily-specific diagnostic mutations (Batzer et al., 1990, 1996b; Jurka and Smith, 1988). It is estimated that approximately 5000 young *Alu* elements are specific to humans (Batzer and Deininger, 1991). Among these young *Alu*

Abbreviations: PCR, polymerase chain reaction; SINE, short interspersed element; WGSA, whole genome shotgun assembly; TSD, target site duplication; PHGS, public version of human genome sequences; CHGS, Celera human genome sequence; BLAST, basic local alignment search tool.

[*] Corresponding author. Tel.: +1 716 845 1556; fax: +1 716 845 1692.
E-mail address: Ping.Liang@Roswellpark.org (P. Liang).

[1] These authors contributed equally to this research.

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.09.031

elements, ~25% have inserted so recently that they are polymorphic among different human population groups, families, or even individuals with respect to their presence or absence in the genome (Batzer and Deininger, 2002).

Because *Alu* insertions are unique events that are identical-by-descent or free of homoplasy, they have been useful in genetic mapping and population genetics studies (Batzer et al., 1994; Batzer and Deininger, 1991; Perna et al., 1992; Roy-Engel et al., 2001; Salem et al., 2003, 2005a; Stoneking et al., 1997; Tishkoff et al., 2000). In addition, *Alu* elements are known to impact several aspects of the genome. For example, *Alu* insertions provide the evolutionary potential to enhance the coding capacity and versatility of the genome by creating novel proteins via insertion into coding regions or by creating alternatively spliced exons (Lev-Maor et al., 2003; Makalowski et al., 1994; Sorek et al., 2002). De novo *Alu* insertions can cause genetic diseases by insertion-mediated interruption of gene structures (Deininger and Batzer, 1999; Ganguly et al., 2003; Wallace et al., 1991).

Using various methodologies, over 1000 *Alu* insertions have been identified as polymorphic among diverse human populations. Earlier studies using genomic library screening with probes/primers specific for young *Alu* elements contributed to the discovery of a small number of polymorphisms (Arcot et al., 1995; Batzer et al., 1995; Batzer and Deininger, 1991; Roy et al., 1999). With the availability of the human genome sequence, a new and more fruitful approach was developed. Using this strategy, *Alu* elements belonging to young subfamilies were identified by computational sequence analysis, and oligonucleotide primers were designed based on the flanking regions for polymerase PCR-based assays to ascertain the polymorphism status of these candidates by screening DNA samples from diverse human populations. The first study using such a strategy identified 106 polymorphic *Alu* insertions out of 475 Ya5 and Yb8 insertions (Carroll et al., 2001). Subsequently, this method has been extensively used to analyze almost all Y subfamilies including Ya (Otieno et al., 2004), Yb (Carter et al., 2004; Roy-Engel et al., 2001), Yc (Roy-Engel et al., 2001; Garber et al., 2005), Yd (Xing et al., 2003), Yg and Yi (Salem et al., 2003), Ye (Salem et al., 2005b) and multiple Y subfamily members on the X chromosome (Callinan et al., 2003). These studies are responsible for the identification of the majority of the known polymorphic *Alu* insertions.

However, the search for polymorphisms using this strategy has so far been limited to the public version of the human genome sequence. In addition, the selection of candidate polymorphisms is biased towards certain relatively small and young subfamilies for which the numbers of candidates are manageable for PCR assays. Therefore, the currently identified polymorphic elements likely represent a partial list of all potential polymorphic *Alu* insertions that exist in current human populations. In fact, a very recently study that utilized the human trace genomic sequences representing different human individuals revealed a high proportion of new polymorphic *Alu* loci (Bennett et al., 2004). In this study, we developed an in silico comparative genomics approach for comparing the public and Celera versions of human genome sequences and identified several hundred new *Alu* insertion polymorphisms. Our data represents the largest set of polymorphic *Alu* insertions identified by a single study to date.

2. Materials and methods

2.1. Sources for genomic sequences

The human genomic sequence data used in this study are the public version (Lander et al., 2001) obtained from the UCSC site (April 12, 2003 freeze or hg15) at http://genome.ucsc.edu and the Celera version from the Celera Discovery System (August 2003 version) through private database subscription (http://cds.celera.com, Venter et al., 2001). The Celera sequences represent unconnected scaffolds grouped by chromosome. We also retrieved the Celera whole genome shotgun assembly (WGSA) sequences from GenBank (accessions AADD01000001-AADD01211493, Istrail et al., 2004). All sequences in fasta format were downloaded onto our local bioinformatics server for analyses.

2.2. In silico identification of Alu insertion polymorphisms

To identify polymorphic *Alu* insertions between the two human genome sequences (public human genome sequence, PHGS and the Celera human genome sequence, CHGS), we developed a strategy as illustrated in Fig. 1. Briefly, all *Alu* elements in both genome sequences plus 100 bp flanking sequences on both sides were identified by querying the genomic sequences with the *Alu* consensus sequences using a locally installed basic local alignment search tool (BLAST) program (Altschul et al., 1997). Each of these sequences from PHGS was used to query the corresponding chromosome in CHGS. If a perfect or close to perfect match at full length (*Alu* element plus flanking sequences with length ≥98% and identity ≥98%) is found, the *Alu* insertion is considered to be shared between the two genomes and is excluded from further analysis. Otherwise, if the best match is limited to only the *Alu* or flanking regions, indicating that there is no full-length match, the *Alu* insertion is considered to be a candidate for being polymorphic and its sequence is subject to another search.

In the second search, the two flanking sequences of the *Alu* element are joined and used to query CHGS. If we find only one perfect or close-to-perfect match, then this *Alu* is considered to be absent in CHGS, i.e. it is polymorphic between the two genomes. Thus, we were able to identify *Alu* loci that are present in the PHGS, but absent from the CHGS. The procedure is then repeated by exchanging the positions of the two genomes to identify *Alu* elements that are present in CHGS but absent from PHGS. All polymorphic loci identified through this automatic computer procedure were subjected to manual verification. For an *Alu* insertion to be considered polymorphic, we required both the existence of a unique perfect match to the joined flanking sequence (with the removal of one copy of the target site duplication) and the absence of the *Alu* element from the other genome at that specific locus.