Available online at www.sciencedirect.com
SCIENCE DIRECT®
Forensic Science International 153 (2005) 117–124
Forensic Science International
www.elsevier.com/locate/forsciint

Inference of human geographic origins using *Alu* insertion polymorphisms

David A. Ray[a,1], Jerilyn A. Walker[a,1], Ashley Hall[b], Barbara Llewellyn[c], Jack Ballantyne[b], Allen T. Christian[d], Kenneth Turteltaub[d], Mark A. Batzer[a,*]

[a]*Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA*
[b]*National Center for Forensic Science, University of Central Florida Post Office Box, 162367, Orlando, FL 32816-2367, USA*
[c]*Illinois State Police, Forensic Science Center at Chicago, 1941 W. Roosevelt Road, Chicago, IL 60608, USA*
[d]*Biology and Biotechnology Research Program, Lawrence Livermore National, Laboratory, PO Box 808, L-452 Livermore, CA 94551, USA*

Received 18 May 2004; received in revised form 26 October 2004; accepted 28 October 2004
Available online 7 April 2005

Abstract

The inference of an individual's geographic ancestry or origin can be critical in narrowing the field of potential suspects in a criminal investigation. Most current technologies rely on single nucleotide polymorphism (SNP) genotypes to accomplish this task. However, SNPs can introduce homoplasy into an analysis since they can be identical-by-state. We introduce the use of insertion polymorphisms based on short interspersed elements (SINEs) as a potential alternative to SNPs. SINE polymorphisms are identical-by-descent, essentially homoplasy-free, and inexpensive to genotype using a variety of approaches. Herein, we present results of a blind study using 100 *Alu* insertion polymorphisms to infer the geographic ancestry of 18 unknown individuals from a variety of geographic locations. Using a Structure analysis of the *Alu* insertion polymorphism-based genotypes, we were able to correctly infer the geographic affiliation of all 18 unknown human individuals with high levels of confidence. This technique to infer the geographic affiliation of unknown human DNA samples will be a useful tool in forensic genomics.

Keywords: Forensic genomics; *Alu*; Geographic affiliation; PCR

1. Introduction

Forensic DNA specimens are routinely matched to alleged criminal suspects in modern law enforcement. Frequently however, tools that narrow the potential pool of suspects are essential precursors to a positive identification in investigative forensics. The inferred ancestral origin of a DNA specimen is one type of evidence that can aid a criminal investigation. Human genetic variation and geographic population affiliation have been studied using many genetic systems, including mitochondrial [1–3], Y-chromosome [1,2], microsatellite [4,5], short tandem repeats (STR) [2,6–8], mobile elements [4,9–14], and single nucleotide polymorphisms (SNPs) [15–18].

Recently, Frudakis et al. [19] developed a SNP-based system for inference of ancestry for application to forensic casework. The initial system consisted of 56 SNP loci targeted from pigmentation and xenobiotic metabolism genes with ancestral diversity designed to identify individuals of European, African, and Asian descent [19]. Subsequently, Frudakis and DNAPrint™ Genomics, Inc. (Sarasota, FL)

* Corresponding author. Tel.: +1 225 578 7102; fax: +1 225 578 7113.
E-mail address: mbatzer@lsu.edu (M.A. Batzer).
[1] These authors contributed equally to this work.

0379-0738/$ – see front matter
doi:10.1016/j.forsciint.2004.10.017

have introduced commercial applications of various SNP-based systems as a forensic service to law enforcement agencies. Notably, DNAWITNESS^{TM} 2.0 was instrumental for inferring the geographic origin of the Louisiana serial killer in 2003 (www.dnaprint.com).

Although emerging SNP-based technologies have recently proven quite useful in law enforcement and will undoubtedly remain so in the future, SNPs have some limitations due the fact that they represent single base pair differences. Like most other genetic polymorphisms, SNPs can be merely identical-by-state; that is, they may have arisen as a result of an independent parallel forward or backward mutation resulting in genotype misclassification (homoplasy). Recent improvements in SNP-based approaches suggest that most of these problems can be overcome by carefully selecting the correct loci and the correct number of SNPs to use.

In this study, we introduce the use of insertion polymorphisms based on short interspersed elements (SINEs) as an alternative or a complement to existing systems. Mobile element insertion polymorphisms are essentially homoplasy-free characters, identical by descent [20–23], and easy to genotype in a variety of formats [4,24–27]. The ancestral state of a human mobile element insertion polymorphism is known to be the absence of the element at a particular genomic location [10]. *Alu* elements are approximately 300 nucleotides in length and represent the most abundant class of short interspersed mobile elements in the human genome with more than one million copies [21]. Most of these elements are ''fixed'', meaning that all individuals are homozygous for the insertion at a particular locus. However, members of several young *Alu* subfamilies such as Ya5, Ya8, Yb8, Yb9, Yc1, Yc2 and others, are polymorphic for insertion presence/absence [22,28–30] and different numbers of such markers have been shown to provide robust measurements of the relationships among various human populations [10,12,13,31]. These features make mobile element insertion polymorphisms virtual genomic fossils of ancestral lineage and, thus, a valuable tool for determining human geographic origins.

Here, we report the application of 100 *Alu* insertion polymorphisms as a forensic tool to ascertain the inferred geographic ancestry of unknown human DNA samples. In this blind study, we examined DNA specimens from 18 geographically diverse humans. For each sample, we used multi-locus genotypes from *Alu* insertion polymorphisms to infer geographic affiliation from among four major world populations.

2. Materials and methods

2.1. DNA samples

Eighteen anonymous human DNA samples were obtained under informed consent for this experiment by the Illinois State Police Forensic Science Center at Chicago and the National Center for Forensic Science, University of Central Florida in Orlando. The DNA from each sample was extracted from bloodstain cards or buccal swabs by the source laboratories (Illinois State Police and National Center for Forensic Science) and shipped to Louisiana State University (LSU) for genetic analysis using 100 *Alu* insertion polymorphisms [4,12] and a mobile element-based sex typing assay [26]. Investigators from each source laboratory had access to the physical description and geographic ancestry of the anonymous subjects while the analysis team at LSU remained blind to this data until the conclusion of the study.

2.2. Alu elements and PCR amplification

One hundred *Alu* insertion polymorphisms were used in this study. A complete list of the *Alu* elements oligonucleotide primers and amplification conditions is available as an electronic appendix (Supplementary Table 1) to this manuscript. It is also available at our website (http://www.batzerlab.lsu.edu) and at http://www.genome.org as supplemental material for Watkins et al. [4,12]. PCR reactions for agarose gel-based detection were carried out in 25 μl using 10 ng of DNA template, 1X PCR buffer II (Applied Biosystems, Inc.), 0.2 mM dNTPs, 200 nM each oligonucleotide primer, optimized $MgCl_2$, and one unit *Taq* DNA polymerase. Each sample was subjected to an initial denaturation of 1 min at 95 °C followed by 32 amplification cycles of denaturation at 95 °C for 30 s, optimized annealing for 30 s, followed by extension at 72 °C for 30 s. Amplicons were size-separated on a 2% agarose gel containing 0.2 μg/ml ethidium bromide and visualized by UV illumination (Fig. 1). Human gender identification was performed using sex chromosome specific mobile elements as previously reported by Hedges et al. [26].

2.3. Data analysis and structure 2.0 inference

Genotypic data were recorded for each allele as follows: an individual who was homozygous present for a given *Alu* locus was assigned the code 1, 1; homozygous absent, 0, 0; and heterozygous, 1, 0. A sample of the data is shown in Table 1 (the complete data set is available in Supplementary Table 2 and at our website, http://batzerlab.lsu.edu under publication #152). The geographic affiliation of the samples was inferred using Structure 2.0 [32–34]. This software package performs model-based clustering using genotypic data from unlinked markers to infer population structure. For each individual, Structure 2.0 estimates the proportion of ancestry from each of K clusters. We used a burn-in of 15,000 iterations and a run of 20,000 replications. The sample size was 715 individuals of known geographic ancestry [12], plus eighteen individuals of unknown ancestry for a total of 733. Because previous analyses of the same known data indicated the presence of four distinct populations [4], the expected number of populations (K) was set at four (European, African, Asian, or Indian). Three replicate