Mobile element-based forensic genomics

David A. Ray  
*Louisiana State University*

Jerilyn A. Walker  
*Louisiana State University*

Mark A. Batzer  
*Louisiana State University*

Follow this and additional works at: [https://digitalcommons.lsu.edu/biosci_pubs](https://digitalcommons.lsu.edu/biosci_pubs)

**Recommended Citation**  
Mobile element-based forensic genomics

David A. Ray¹,², Jerilyn A. Walker¹, Mark A. Batzer∗

Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803, USA

Available online 8 December 2006

Abstract

Mobile elements are commonly referred to as selfish repetitive DNA sequences. However, mobile elements represent a unique and underutilized group of molecular markers. Several of their characteristics make them ideally suited for use as tools in forensic genomic applications. These include their nature as essentially homoplasy-free characters, they are identical by descent, the ancestral state of any insertion is known to be the absence of the element, and many mobile element insertions are lineage specific. In this review, we provide an overview of mobile element biology and describe the application of certain mobile elements, especially the SINEs and other retrotransposons, to forensic genomics. These tools include quantitative species-specific DNA detection, analysis of complex biomaterials, and the inference of geographic origin of human DNA samples.

Keywords: Mobile element; Forensic genomics; SINEs

1. Introduction to mobile elements

Mobile elements are repetitive DNA sequences with the unique ability to move and/or make copies of themselves. They can comprise between 40 and 90% of genomes [1–3]. While individual insertions tend to be either neutral or deleterious, their prevalence in eukaryotic genomes indicates that their presence is, on the whole, selectively advantageous. Several studies suggest that they play important roles in genome structure, genome instability and gene expression [1,4–11]. For instance, the movement of these elements can change genomes by insertion mutagenesis [9,12–16], insertion-mediated genomic deletions [15,17–20], and transposition-mediated transduction [15,21–24]. Furthermore, the widespread presence of very similar DNA sequences throughout a genome promotes recombination, including both equal and unequal crossover events [9,15,25,26], segmental duplication [27], gene conversion [28–30], exon shuffling [23,31] and chromosomal rearrangements [32].

Mobile elements are divided into two classes based on their method of mobilization (Fig. 1) [33]. Class I mobile elements utilize a “copy and paste” method referred to as retrotransposition. In this process, the original DNA copy in the genome is first transcribed to RNA. The transcript is used as a template to form a DNA molecule that is then inserted into a new location in the genome via a process known as target primed reverse transcription [34]. Autonomous class I elements encode some or all of the enzymes needed to accomplish their mobilization. These retrotransposons are subdivided into several subgroups based on autonomy and sequence characteristics. The autonomous class I elements include LTR (long terminal repeat) retrotransposons that share many characteristics with retroviruses [35]. Also autonomous are the non-LTR retrotransposons, which are usually 4–6 kb
Fig. 1. General structure for the two classes of mobile elements. Class I elements of the LTR variety are flanked by both long terminal repeats (LTRs) and target site duplications (TSDs), which are produced during integration into the genome. These autonomous retrotransposons encode several polypeptides related to viruses. LINEs typically contain two open reading frames. Finally, SINEs do not contain any open reading frames. Instead, they use the enzymatic machinery of LINEs to mobilize in a genome. They are transcribed by RNA polymerase III which recognizes the A and B promoter sequences that make up part of the SINE sequence itself. SINEs also often have a repetitive motif at the 3′ end ((XX)n). Autonomous class II elements, the DNA transposons contain an open reading frame for transposase and are typically flanked by inverted terminal repeats (ITRs) and direct repeats (DRs).

<table>
<thead>
<tr>
<th>Mobile Element Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Class I</strong></td>
</tr>
<tr>
<td>LTR Retrotransposon</td>
</tr>
<tr>
<td>Virus-like ORFs</td>
</tr>
<tr>
<td>1 - 11 kb</td>
</tr>
<tr>
<td>TSD</td>
</tr>
<tr>
<td>LTR</td>
</tr>
<tr>
<td>TSD</td>
</tr>
<tr>
<td>LINE</td>
</tr>
<tr>
<td>ORF1</td>
</tr>
<tr>
<td>ORF2</td>
</tr>
<tr>
<td>(A)n</td>
</tr>
<tr>
<td>6 - 8 kb</td>
</tr>
<tr>
<td>5′UTR</td>
</tr>
<tr>
<td>3′UTR</td>
</tr>
<tr>
<td>SINE</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>(XX)n</td>
</tr>
<tr>
<td>0.1 - 0.5 kb</td>
</tr>
<tr>
<td>TSD</td>
</tr>
<tr>
<td><strong>Class II</strong></td>
</tr>
<tr>
<td>DNA Transposon</td>
</tr>
<tr>
<td>Transposase</td>
</tr>
<tr>
<td>1 - 3 kb</td>
</tr>
<tr>
<td>DR</td>
</tr>
<tr>
<td>ITR</td>
</tr>
<tr>
<td>ITR</td>
</tr>
</tbody>
</table>

A typical example is the LINE-1 (long interspersed element-1) repeat of mammals. LINE-1 elements contain two open reading frames (ORFs). The first, short ORF encodes a protein thought to have a nucleic acid binding function [36–39]. The second ORF encodes a protein with both endonuclease and reverse-transcriptase properties [40,41]. It is via this enzymatic machinery that a LINE is able to mobilize after transcription by RNA polymerase II [42]. The endonuclease of LINE-1 and other non-LTR retrotransposons recognize specific target sequences in order to initiate the insertion process [43,44].

Class II elements, the DNA transposons and their derivatives, are common in many organisms from bacteria to humans. First discovered by Barbara McClintock in maize [45], DNA transposons differ from class I elements in that they utilize a DNA intermediate when mobilizing. They may utilize a “cut and paste” (conservative) or “copy and paste” method (replicative). In the former case, an entire DNA segment that includes the coding sequences and associated non-coding regions is excised from where it resides and reinserted into the genome at a different location. This process is mediated by the action of transposase, which is encoded by the autonomous element. Usually, transposase binds to both the inverted repeats of the transposon and a target site in the genome where the transposon will reinsert. The target site is cut and the transposon is then inserted.

Non-autonomous versions exist for both classes I and II elements. These elements lack the coding sequences typical of autonomous elements. Instead, they are believed to “hijack” the machinery of autonomous elements to accomplish their mobilization [46,47]. Several examples include TRIMs (terminal-repeat retrotransposons in miniature) and LARDs (large retrotransposon derivatives) that are both recently described non-autonomous elements derived from LTR retrotransposons [48,49]. SINEs (short interspersed elements) are non-autonomous, non-LTR retrotransposons that occur in large numbers in a wide variety of vertebrate genomes [14,50–53]. Class II non-autonomous mobile elements include the MITEs (miniature inverted terminal repeat elements), Mers, and Acrobat elements among others [54–56].

This review will focus on SINEs. The widespread presence of SINEs in eukaryotes [14,50] and their relatively small size (usually between 100 and 500 bp) make them useful for a variety of different forensic approaches. Below, we will discuss some of them. The approaches that have been attempted include the simple identification of the taxon present in an unknown DNA sample to the identification of ancestry within a specific taxon.

2. Identification and quantitation of species-specific DNA

As various SINEs and LINEs expanded throughout the genomes of different organisms, generating over 100,000 mobile elements in many, genome-specific families with characteristic diagnostic mutations have arisen [10]. These large dispersed gene families serve as novel markers that identify the DNA from each species, thus...
Fig. 2. Species DNA detection using PCR and agarose gel electrophoresis. Lanes: (1) 100 bp ladder, (2) cow, (3) pig, (4) chicken, (5) horse, (6) sheep, (7) deer, (8) antelope, (9) rabbit, (10) duck, (11) dog, (12) cat, (13) rat, (14) mouse, (15) human and (16) NTC (no template control). (A) Cow-specific; (B) pig-specific; (C) chicken-specific; (D) amplifies in all ruminants.

providing specific genomic tags that can be used in conjunction with PCR to amplify specific subsets of genomic sequences unique to the genome or species of interest (for recent reviews see [57,58]). Mobile element-based PCR assays have been developed for the sensitive identification and quantitation of bovine, porcine, chicken and ruminant species DNA from mixed sources and food products [59,60]. To date, class-specific intra-element PCR assays are available for birds, order-specific assays are available for primates, ruminants and rodents, and species-specific assays are available for humans, cows, horses, pigs, dogs, cats, hamsters, guinea pigs, rats, rabbits, and chickens (Fig. 2).

3. Quantitative detection of species DNA in meats

Documented cases of “mad cow disease” (bovine spongiform encephalopathy) have received high publicity in recent years. The risk associated with infectious transmissible spongiform encephalopathy in humans has discouraged many individuals around the globe from consuming beef. Hindu populations also choose not to eat beef, while Jewish and Muslim populations choose to avoid consumption of pork, even in minute quantities, due to their religious beliefs. Many consumers prefer to include more chicken in their diet instead of beef or pork. In addition to infectious disease and religious concerns, many individuals are altering their eating behavior to include more chicken simply to reduce dietary fat intake in accordance with health trends. Physicians routinely recommend such dietary changes to patients with heart disease, diabetes or obesity, for example.

Previously, species-specific SINE-based and LINE-based PCR assays have been reported for the detection of ruminant-, pig-, and chicken-derived materials [59,60]. Quantitative SINE-based PCR assays have been developed for the sensitive detection and quantitation of bovine, porcine, chicken and ruminant species DNA from mixed sources and food products [60]. For example, the bovine SINE satellite 1.711B is thought to occupy 7.1% of the bovine genome [61] and the intra-SINE-based PCR assay using this repeat element can detect 0.005% bovine DNA from a mixed DNA sample [60]. The porcine SINE PRE-1 has an estimated 100,000 copies in the pig genome [62] and the intra-PRE-1-based PCR assay is able to detect even trace quantities (0.0005%) of porcine DNA from within a mixed-DNA sample [60]. A randomly selected food product labeled “chicken sausage” was determined to contained about 8% pork [60]. The chicken intra-CR1 SINE-based PCR assay can detect down to 0.05% chicken DNA and the intra-Bov-tA2 SINE-based PCR assay can detect as little as 0.1% ruminant species DNA (cow, sheep, deer, antelope, etc.) from mixed sources [60]. This highly sensitive species specificity makes these assays ideal for identification of beef, pork, chicken, and ruminant materials from human foods and animal feed mixtures.

4. Identification and quantitation of animal DNA

In addition to agricultural meat species intended for human consumption, certain situations make determining the source of a tissue, blood or DNA sample a necessity. For example, some violent crime scenes may include blood stains from not only human victims but from our animal companions as well. Other events such as earthquakes, tornadoes, and terrorist attacks create such volatile conditions that the biological evidence available for post-event genomic analysis may include a human and a nonhuman component. Thus, there is a need for sensitive detection and quantitation of other common domestic species from multi-species or unknown DNA samples.
To address this need, a series of class-specific (Aves), order-specific (Rodentia), and species-specific (equine, canine, feline, rat, hamster, guinea pig, and rabbit) PCR-based assays were developed using intra-element amplification of genome-specific SINEs and LINEs [63]. The minimum effective quantitation levels of these assays ranged from 1 to 0.001% in 10 ng of starting DNA template. Thus, even when small amounts of DNA are available, it is possible for forensic investigators to determine “who” may have been present at a scene whether that individual is human or non-human.

5. The Alu element

The most abundant class of SINE in the human genome is the Alu element (Fig. 3). This SINE family is derived from the 7SL gene and is specific to primate genomes. Alu elements are only ~300 bp long but, with more than one million copies, they account for more than 10% of the mass of the human genome [1,14]. Most Alu elements are “fixed” in the human genome, meaning that all individuals are homozygous for the insertion at a particular locus in the genome in which it resides. The expansion of Alu elements throughout primate evolution has created several recently integrated “young” subfamilies that are present in the human genome but are largely absent from nonhuman primates [14]. Some members of these young Alu subfamilies have inserted in the human genome recently enough that individuals remain polymorphic for insertion presence/absence [14,64–68]. Both fixed and polymorphic Alu elements have been utilized quite successfully as forensic tools for a variety of applications [69,70].

6. Mobile element-based human gender identification

One important forensic use for Alu elements is human gender identification [70]. Determination of gender from human DNA samples is a common problem in forensic laboratories. The most widely used approach is based on the Amelogenin locus, which yields different-sized PCR (polymerase chain reaction) amplicons for the X and Y chromosome versions of the Amelogenin gene [71]. However, this method can misidentify males as females in some cases due to a deletion in the AMEL Y region [72–74]. While the frequency of the deletion is relatively low, the crucial nature of forensic test results in circumstances such as rape and prenatal gender determination where there is risk for male-specific inherited disorders makes any source of error a legitimate cause for concern. This has led several researchers to recommend that Amelogenin should not be relied upon as the sole determinant of gender [72–75].

Fixed Alu insertions on either the X or the Y chromosome provide a way of identifying the respective chromosome. Loci AluSTXa and AluSTYa demonstrate 100% accuracy in gender identification (X and Y chromosome identification) in 778 human DNA samples from diverse populations [70]. The combination of these two markers provides added assurance that gender test results are accurate since independent mutations would have to occur in two separate genomic locations to affect the outcome of the test. This Alu-based method may also be useful as a precursor to more extensive Y chromosome marker analysis, such as Y-Plex™ 12 (ReliaGene Technologies, Inc., New Orleans, LA). In addition, since AluSTXa is fixed on the X chromosome and AluSTYa is fixed on the Y chromosome, fluorescent-based detection of these markers allows for quantitative resolution of male and female contributions to mixed DNA samples, such as might be found in sexual assault cases.

7. Human DNA identification and quantitation

Just as in the examples for other species described above, the identification of human DNA in a complex mix may also be required. The vast majority of human forensic casework involves the analysis of nuclear loci, making the quantitation of human nuclear DNA a paramount issue [76]. Alu elements have amplified to a copy number of over 1 million elements throughout primate evolution producing a series of subfamilies of Alu elements that appear to be of different ages [14]. Because of their high copy number, Alu elements are a rich source of human genetic markers. Several systems have been developed to quantify human DNA using PCR primers designed to amplify the core Alu sequence [77–81]. Dur-
ing intra-Alu PCR, primers are designed within the core body of the element to amplify multiple target copies of the Alu family. Most of these systems are specific to the primate lineage and are not necessarily human specific.

Intra-Alu-based PCR assays have been developed for human-specific DNA identification and quantitation based on Alu subfamilies Yb8 and Yd6 [69], taking advantage of the large diagnostic insertion or deletion characteristic of these respective subfamilies. The intra-Yb8-based PCR assay has a linear quantitation range of 10–0.001 ng of DNA and is capable of detecting 10 pg of human DNA from a mixed source sample [69]. The intra-Yd6-based PCR assay is less sensitive quantitatively with a detection limit of 100 pg of human DNA. But the intra-Yd6-based PCR assay is specific to humans, no background amplification occurs from non-human primate templates, and has proven useful as an initial screening tool [69]. The collective use of primate specific and human specific Alu-based PCR assays in forensic genomics provides a powerful tool for sensitive human DNA identification and quantitation from mixed sources.

8. Alu insertion polymorphisms and inference of ancestry

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 predictor evidence which can aid a criminal investigation. Polymorphic Alu insertions have been widely used to study human genetic variation in world populations [82–103]. These elements are ideal for studying human geographic origins because they are essentially homoplasy free characters [57,104], they are identical by descent [14,57,105], and the ancestral state of a mobile element insertion polymorphism is known to be the absence of the element at a particular genomic location [83]. These features make the analysis of Alu elements a robust tool for tracking human geographic ancestry.

The distribution and structure of human genetic diversity has been systematically evaluated in a dataset of 100 Alu insertion polymorphisms, mined from existing human genome databases and selected for the greatest allele frequency variation between populations [106]. In this analysis, 710 individuals representing 31 different populations across 4 continents were evaluated for each of the 100 Alu insertion polymorphisms [85]. The inferred geographic affiliation of the samples was calculated using the Structure 2.0 software package [107,108]. The Markov Chain Monte Carlo methodology used by Structure 2.0 incorporates a powerful analysis to group all individuals into the selected number of populations and then determine the probability that each individual belongs to any given group. In addition, the software has the ability to detect admixture between populations in individual genotypes going back three generations.

This existing dataset is a powerful forensic tool to ascertain the inferred geographic ancestry of unknown forensic DNA samples. The methodology is simple, requiring only standard thermal cyclers for PCR and the ability to run agarose gel electrophoresis. In a study conducted in collaboration with law enforcement officials, forensic DNA specimens were evaluated using the dataset of 100 Alu insertion polymorphisms and Structure 2.0. The system demonstrated accurate assignment of each specimen to the correct geographic affiliation [109].

In addition to existing datasets, Alu elements have continued to accumulate in various populations since the human dispersal from Africa estimated to have occurred 156,000 years ago [85]. Many of these “new” Alu insertions are likely to be polymorphic and occur in various world populations to different degrees. The presence of these population-specific insertions will allow for subsequent layers of subgroup affiliation tests to be conducted using a cascade-like approach in the analyses. Thus, once the initial Structure 2.0 analysis assigns the DNA specimen to a continental affiliation such as Europe, Africa, Asia or India, subsequent Structure 2.0 analyses, using only insertion loci that are useful within the sub-continental population groups, can be used to further isolate the sub-continental region. The inferred ancestral origin of a DNA specimen narrows the pool of potential suspects and can successfully impact a criminal investigation.

Forensic DNA samples are often only available in trace quantities. This has previously been a limitation to this type of multiple locus approach. Recent advancements in whole genome amplification (WGA) technologies which utilize Phi29 polymerase have virtually eliminated this obstacle. In fact, successful genotypes have been obtained using post-WGA DNA, obtained from residual skin cells salvaged from a fingerprint [110]. Therefore, the inferred ancestral origin of a human DNA sample can even be determined from trace evidence.

9. Conclusion

Mobile elements represent novel tools within the increasingly varied armamentarium available to
forensic scientists. Their utility ranges from simple detection schemes, such as gender and species identification, to complex multi-locus systems for inference of human ancestry. Continued development of mobile element-based genetic systems in combination with high-throughput DNA analysis technologies will undoubtedly further the application of these approaches in the future.

Acknowledgements

This research was supported by the Louisiana Board of Regents Governor’s Biotechnology Initiative GBI (2002-005) (MAB), National Science Foundation EPS-0346411 (M.A.B.) and the State of Louisiana Board of Regents Support Fund (MAB).

This paper is dedicated to the memory of Tony Carrano. During my (MAB) years at Lawrence Livermore National Laboratory I had the privilege of working in the Human Genome Center under the guidance of Tony. At that time Tony had the foresight to initiate a number of the forensic genomics efforts at Lawrence Livermore National Laboratory in an effort to create future areas of investigation for the Biology and Biotechnology Research Program. These early initiatives proved to be the underpinnings for many subsequent years of fruitful scientific investigations by individuals at Lawrence Livermore National Laboratory and elsewhere including my own laboratory. Tony was an exceptional scientist and was one of the best scientific administrators that I have ever known. His mentoring, dedication, leadership, and enthusiasm will be remembered and missed.

This paper is dedicated to the memory of Tony Carrano. During my (MAB) years at Lawrence Livermore National Laboratory and elsewhere including my own laboratory. Tony was an exceptional scientist and was one of the best scientific administrators that I have ever known. His mentoring, dedication, leadership, and enthusiasm will be remembered and missed.

References


[25] V.O. Kolosha, S.L. Martin, High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from...


[54] V.V. Kapitonov, J. Jurka, ACROBAT1, a nonautonomous DNA transposon from zebrafish, Repbase Rep. 2 (2002) 1.


D.A. Ray et al. / Mutation Research 616 (2007) 24–33


