Mobile DNA elements in primate and human evolution

Jinchuan Xing  
*University of Utah Health Sciences*

David J. Witherspoon  
*University of Utah Health Sciences*

David A. Ray  
*West Virginia University*

Mark A. Batzer  
*Louisiana State University*

Lynn B. Jorde  
*University of Utah Health Sciences*

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

**Recommended Citation**


This Article is brought to you for free and open access by the Department of Biological Sciences at LSU Digital Commons. It has been accepted for inclusion in Faculty Publications by an authorized administrator of LSU Digital Commons. For more information, please contact [ir@lsu.edu](mailto:ir@lsu.edu).
Mobile DNA Elements in Primate and Human Evolution

Jinchuan Xing,1 David J. Witherspoon,1 David A. Ray,2 Mark A. Batzer,3 and Lynn B. Jorde1*

1Department of Human Genetics, University of Utah Health Sciences Center, Salt Lake City, UT 84112
2Department of Biology, West Virginia University, PO Box 6057, Morgantown, WV 26506
3Department of Biological Sciences, Biological Computation and Visualization Center, Center for Bio-Modular Multi-scale Systems, Louisiana State University, Baton Rouge, LA 70803

ABSTRACT Roughly 50% of the primate genome consists of mobile, repetitive DNA sequences such as Alu and LINE1 elements. The causes and evolutionary consequences of mobile element insertion, which have received considerable attention during the past decade, are reviewed in this article. Because of their unique mutational mechanisms, these elements are highly useful for answering phylogenetic questions. We demonstrate how they have been used to help resolve a number of questions in primate phylogeny, including the human–chimpanzee–gorilla trichotomy and New World primate phylogeny. Alu and LINE1 element insertion polymorphisms have also been analyzed in human populations to test hypotheses about human evolution and population affinities and to address forensic issues. Finally, these elements have had impacts on the genome itself. We review how they have influenced fundamental ongoing processes like nonhomologous recombination, genomic deletion, and X chromosome inactivation. Yrbk Phys Anthropol 50:2–19, 2007. ©2007 Wiley-Liss, Inc.

KEY WORDS Alu; retrotransposon; primate phylogeny; human population genetics; forensics

Mobile DNA elements are discrete DNA sequences that have the remarkable ability to transport or duplicate themselves to other regions of the genome. This process has been occurring in virtually all organisms for many millions of years. As a result, mobile elements are a major component of human and nonhuman primate genomes, accounting for 40–50% of their contents (Lander et al., 2001; CSAC, 2005; Gibbs et al., 2007). The properties and applications of these mobile elements have been the subject of increasing interest during the past two decades. In this review, we will focus on several areas in which the study of mobile elements has made significant contributions to physical anthropology.

Because of their unique manner of propagating themselves in the genome, mobile elements are highly useful in tracing relationships of individuals, populations, and species. We will demonstrate how these elements have been used to resolve questions about primate systematics. We will then show how they have been useful in addressing questions about human origins, population affinities, and population history. The potential to use mobile elements as forensic tools, both in human and in nonhuman primates, will be discussed next. Finally, the presence of more than one million mobile elements in the primate genome has had a significant impact on processes such as nonhomologous recombination, genomic deletion and duplication, and gene conversion. The nature and consequences of this impact will be the final topic of the review. Before reviewing these applications, we begin with a brief discussion of the basic biology that underlies mobile elements and their behavior in the genome.

BASIC MOBILE ELEMENT BIOLOGY

Mobile elements can be divided into two different classes based on how they duplicate themselves within the genome. DNA transposons (Fig. 1A) mobilize through DNA intermediates, typically using a so called “cut and paste” mechanism (Smit and Riggs, 1996; Pace and Feschotte, 2007). During this process, transposase, an enzyme encoded within some transposons, first excises a DNA transposon from its original genomic location. It then generates a break in another genomic location, where it reinserts the transposon. A short stretch of identical DNA sequence (“target site duplications” or TSDs) is created on either end of the newly integrated transposon during this enzymatic process. In the primate lineage, DNA transposons have been shown to be inactive for the past 40 million years or so (Pace and Feschotte, 2007). Therefore, this review will focus on the other major class of mobile elements: retrotransposons.

Retrotransposons mobilize through RNA intermediates using a “copy and paste” mechanism (Ostertag and Kazazian, 2001; Deininger and Batzer, 2002). In this process, an RNA copy is first generated from the original retrotransposon and is subsequently reverse-transcribed back into DNA using an enzyme called reverse transcriptase. It is then inserted into a new location in the genome. As in the transposition of DNA transposons, target site duplications are usually generated during this process. Retrotransposons can be further subdivided into those elements that are autonomous, meaning that they encode their own replication machinery (e.g. Long Inter-
A (CCCTCT) element. SVA elements can be divided into five components: (1) proteins provide enzymatic machineries that are necessary for their propagation from L1 elements (Kajiwara and Okada, 2002; Dewannieux et al., 2003). A second family of nonautonomous retrotransposons was recently discovered in hominids and termed “SVA” for each of its three components [SINE-R, VNTR (Variable Number of Tandem Repeat), and Alu] (Shen et al., 1994; Wang et al., 2005) (Fig. 1C).

L1 elements (Fig. 1D), the most successful class of autonomous retrotransposons in mammalian genomes (Smit et al., 1995; Ostertag and Kazazian, 2001), arose as a family of repeated DNA sequences about 150 million years ago. In primate genomes, these elements have reached a copy number in excess of 500,000 elements (Lander et al., 2001; CSAC, 2005; Gibbs et al., 2007). The majority of these elements are truncated or defective copies that were originally produced by a relatively small number of full-length, retrotransposition-competent copies (Brouha et al., 2003; Han et al., 2007). These full-length L1 elements are about 6 kb in length and encode an RNA-binding protein as well as a second protein with endonuclease and reverse transcriptase activity. Because they can make copies of themselves and are likely the source of all other L1 elements in the genome, these L1 elements are termed “master” mobile elements. The human genome contains 80–100 of these retrotransposition-competent elements (Brouha et al., 2003). Curiously, only a few retrotransposition-competent L1 elements were recovered from the draft sequence of the rhesus macaque genome (Han et al., 2007).

Retrotransposon-based genetic systems were first applied as a research tool in the early 1990s (see Fig. 2) (Batzer and Deininger, 1991; Batzer et al., 1991; Perna et al., 1992; Minghetti and Dugaiczky, 1993; Murata et al., 1993; Batzer et al., 1994). Their use was highlighted by phylogenetic studies of salmon (Murata et al., 1993; Murata et al., 1996), whale (Shimamura et al., 1993; Batzer et al., 1994). Early applications of these systems also addressed questions about human origins and population affinities (Batzer et al., 1994, 1996; Stoneking et al., 1997). Since then, retrotransposons have become widely recognized as powerful tools for phylogenetic and population genetic studies (Shedlock and Okada, 2000; Shedlock et al., 2004; Schmitz et al., 2005; Ray et al., 2005).
Retrotransposon-based systems have three important advantages compared to other types of genetic loci (e.g., single nucleotide polymorphisms (SNPs), microsatellites, and restriction fragment length polymorphisms (RFLPs)). First, the probability that two retrotransposons will insert independently in the same position is essentially zero. Even if two retrotransposons did insert at the same location, they can usually be distinguished by the type of insertions and length of the flanking TSDs generated during their integration (Conley et al., 2005).

In addition, the precise removal of an entire retrotransposon after its fixation in the genome is extremely rare (van de Lagemaat et al., 2005) and is very unlikely to happen in multiple genomes. These two characteristics suggest that retrotransposons are essentially free of homoplasy, and that two individuals sharing the same retrotransposon insertion most likely acquired it from a common ancestor (identity by descent) (Saleh et al., 2005; Ray et al., 2006). This allows retrotransposon-based studies to infer relationships more accurately than other systems, in which different individuals may acquire the same genotype independently (i.e., identity by state/homoplasy). Second, the ancestral state of any given locus can be determined unambiguously as the absence of the retrotransposon insertion. This simplifies phylogenetic inference and presents another major advantage over other marker systems. Third, unlike most other molecular systems, the use of retrotransposons in phylogenetic and population genetic inference does not directly rely on comparisons of DNA sequence variation. This feature allows retrotransposons to complement traditional DNA sequence-based molecular studies.

Together, these features make retrotransposon-based systems especially useful for groups that have high levels of sequence homoplasy and for closely-related species for which there is little phylogenetic information in DNA sequences.

Nevertheless, like any other genetic system, retrotransposon-based systems are not perfect. Several factors, including lineage sorting, parallel or near-parallel insertion, and precise deletion of the element, may potentially create problems and confound the results (Hillis, 1999; Ray et al., 2006). Several studies have shown that most confounding events happen at very low rates, and basic cautions, including careful analysis and interpretation of the data, can limit their potential influence (Shedlock and Okada, 2000; Shedlock et al., 2004; Salem et al., 2005; Ray et al., 2006). Basic knowledge of the taxa under investigation and collection of a sufficient number of informative insertions can also increase confidence in results (Waddell et al., 2001).

Although in principle all retrotransposons can be used as phylogenetic or population genetic markers, SINEs have been the choice of markers in most studies for two reasons. First, the amplification of SINEs has been well studied in many taxa, and specific methods have been developed to retrieve SINE insertions from species for which little DNA sequence information is available (Okada et al., 2004). Second, the short length of SINEs makes them easier to manipulate and genotype than longer mobile elements such as LINEs, which typically require multiple PCR amplifications to genotype a single locus (Sheen et al., 2000; Myers et al., 2002; Vincent et al., 2003).

Mobile Elements in Primate Phylogenetic Studies

Primate phylogeny and mobile element-based primate phylogenetic studies

The order Primates is one of the most diverse mammalian orders, exceeded in number of species only by Rodentia and Chiroptera. Phylogenetic studies of primates can be traced back to the beginning of the modern systematics, when Linnaeus recognized the order in 1766 and divided it into four divisions with 35 species. Over the last two and a half centuries, primate systematics has been studied intensively by a variety of means, including the fossil record, morphology, behavioral and ecological studies, and molecular and genetic analysis (Rasmussen and Nekaris, 1998; Ross et al., 1998; Goodman, 1999; Schneider, 2000; Yoder, 2003; Kay et al., 2004; Schmitz et al., 2005). Today, more than 300 primate species are recognized (Groves, 2001), with new species and even genera still being identified (Jones et al., 2005; Sinha et al., 2005; Davenport et al., 2006).

The current consensus view of primate phylogeny (Disotell, 2003) divides order Primates into two suborders, Strepsirrhini and Haplorhini. Strepsirrhini include the Lorisiformes (loris) and the Lemuriformes (leums). Haplorhini include the Tarsiiformes (tarsiers) and Anthropoidea, which is further subdivided into the Platyrhini (New World monkeys, NWMs) and the Catarrhini, composed of Cercopithecidae (Old World monkeys, OWMs) and Hominoidea (apes and humans). Despite great progress in primate systematics, a number of questions remain unanswered or debated. Examples include the monophyly of the order Primates, the aye-aye affiliation, the tarsier affiliation, NWM family relationships, and the human–chimpanzee–gorilla trichotomy, just to name a few. The development of mobile element-based genetic systems has provided opportunities to resolve many of these questions.

Alu elements, which are primate-specific SINEs, have been used extensively in primate phylogenetic studies and have been investigated in nearly the entire order (see Fig. 3). The human genome draft sequence has been invaluable in determining locations and flanking sequences of Alu insertions (Schmitz et al., 2001; Salem et al., 2003c; Ray et al., 2005a). In addition, the wealth of knowledge of Alu biology (e.g., subfamily phylogenetic distributions) (Batzner and Deininger, 2002) has greatly facilitated marker-mining in different primate genomes. In the following text, we briefly review recent progress in retrotransposon-based primate phylogenetic studies.

Primate monophyly

It is generally accepted that order Primates forms a monophyletic group, which belongs to cohort Archonta along with orders Scandentia (tree shrews), Dermoptera (flying lemurs or colugos), and Chiroptera (bats). However, several recent molecular studies based on mitochondrial sequence suggested the clustering of Dermoptera with anthropoid primates (Murphy et al., 2001; Arnason et al., 2002; Schmitz et al., 2002a). A clade comprising flying lemurs and anthropoids has therefore been
proposed with the name Dermosimii (Arnason et al., 2002).

Mobile element insertions have been used to resolve the relationship between flying lemurs and primates (Schmitz et al., 2002a; Schmitz and Zischler, 2003). In these two studies, Schmitz et al. recovered three insertions present in the genomes of all tested primates but absent in the flying lemur genome. In addition, in a genome-wide analysis using Alu elements (present in all primates) as probes, no signal could be detected from the flying lemur genome, arguing against the inclusion of flying lemurs in order Primates. Therefore, it appears that a mitochondrial-specific nucleotide and amino acid composition bias may have caused confounding results in the earlier analysis (Schmitz et al., 2002a). These results demonstrate some of the advantages of retrotransposon-based systems in phylogenetic analysis.

**Tarsier affiliation**

Separated from other primates some 45 million years ago, tarsiers (genus *Tarsius*) have evolved a series of unique characters (Yoder, 2003). The affiliation of tarsiers is one of the most controversial issues in primate phylogeny and is fundamental for constructing primate infraorder relationships. Two different systems have been proposed depending on the tarsier affiliation: one divides order Primates into Prosimii (tarsiers, lemurs, and lorises) and Anthropoidea (monkeys and apes). Alternatively, if tarsiers are considered a sister clade to NWMs, OWMs, and Hominoids, then the Primates order can be divided into Strepsirrhini (lemurs and lorises) and Haplorhini (tarsiers, monkeys, and apes). Each of these two classifications is supported by a number of morphological studies. Analyses based on mitochondrial (Andrews et al., 1998; Schmitz et al., 2002b) and nuclear DNA (Goodman et al., 1998) were unable to resolve this question unambiguously [see (Yoder, 2003) for a review].

Four different SINE-based studies have yielded evidence supporting the Strepsirrhini and Haplorhini division. Zietkiewicz et al. (1999) determined the sequences of 37 Alu elements in the tarsier genome and compared them with the human and strepsirrhine Alu consensus sequence. This comparison indicated that tarsier Alu elements cluster with Alu subfamilies from the human lineage and support the sister relationship of tarsiers and anthropoids (monkeys and apes). In a second study, Schmitz et al. (2001) examined 118 human intronic Alu insertion loci. Three insertions were present in tarsier and anthropoids but absent in the strepsirhines. In a recent review, Schmitz et al. (2005) provide evidence for another Alu insertion in the human serine palmitoyl transferase gene, which is shared between tarsier and anthropoids, further supporting the Strepsirrhini and Haplorhini division. In a fourth study, Kuryshhev et al. (2001) discovered at least one unambiguous insertion...
Yearbook of Physical Anthropology—DOI 10.1002/ajpa

(FLAM-A) shared between tarsier and human but not strepsirrhines. Combined together, SINE insertion data provide some of the strongest evidence for the Strepsirrhini versus Haplorhini division.

Strepsirrhine phylogeny

Strepsirrhine primates are divided into two infraorders: Lemuriformes (lemurs) and Lorisoformes (lorises) (Disotell, 2003), although some authors consider aye-aye (Daubentonia madagascariensis) as a separate infraorder, Chiromyiformes (Groves, 2001). The aye-aye is one of the most morphologically unique primates, and its relationship with other strepsirrhines is rather mysterious (Yoder, 1997). Besides the aye-aye’s affiliation, a number of lower-level relationships among strepsirrhines are still under dispute.

In an attempt to resolve strepsirrhine phylogeny, Roos et al. (2004) examined SINE insertions in 20 strepsirrhine species. In a multiloci approach, two types of SINE insertions with known sequence were analyzed with Southern blots. Another 42 individual loci obtained from genomic sequence were genotyped with PCR, and a total of 61 SINE insertions were detected from these loci. With these SINE insertions, Roos et al. were able to infer a single phylogenetic tree with strong support. In addition, this tree was further supported by another 18 informative SINE insertion loci identified by Herke et al. (2007). Combined with the mitochondrial genome sequence, these results suggest a sister relationship of aye-aye and Lemuriformes and the monophyly of family Lorisidae. The inferred phylogeny supports the hypothesis that strepsirrhines originated in Africa and that Madagascar and Asia were colonized by single immigration events from Africa.

New World monkey phylogeny

NWMs (infraorder Platyrhini) are generally considered a monophyletic group consisting of 15 to 16 genera. These genera have been divided into two (Cebidae and Atelidae) (Disotell, 2003), three (Pitheciidae, Cebidae, and Atelidae) (Goodman et al., 1998), or four families (Pitheciidae, Cebidae, Aotidae, and Atelidae) (Groves, 2001) by different authors, and the branching order of these families has not been determined with confidence (Schneider, 2000). Different topologies can be produced from the same dataset depending on the analytic method (Steiper and Ruvolo, 2003).

Two mobile element-based studies were performed to resolve family relationships. In an earlier study (Singer et al., 2003), six informative elements were identified among 74 intronic Alu insertions. Besides three Alu insertions supporting the monophyly of Platyrhini, one insertion grouped genus Callithrix and Cebuella, one insertion supported the monophyly of the subfamily callitrichinae, and one insertion suggested a close affiliation for Aotus, Saimiri, and Cebus to the callitrichine monkeys. Because of the relatively small number of SINE insertions examined in this study, the inferred relationships are suggestive and have limited statistical support.

A second study by Ray et al. (2005b) is more comprehensive: a total of 183 Alu elements that integrated into various NWM genomes were used in the analysis. Among them, 124 Alu elements are present in at least two NWM species but not in all species, making them informative in parsimony analysis (parsimony-informative markers). This analysis produced a single most parsimonious tree with a consistency index of 1.0. (The consistency index, which varies from 0 to 1, is a measure of how well the observed data fit the phylogenetic tree.)

This tree supports a sister relationship between family Atelidae (spider, woolly, and howler monkeys) and Cebi- dae (marmosets, tamarins, squirrel, capuchin, and owl monkeys). Family Pitheciidae (titi and saki monkeys) then joined this two-family clade to form the infraorder. This study represents the strongest evidence to date regarding NWM family relationships.

Old World monkey phylogeny

OWMs (family Cercopithecidae) can be divided into two distinct subfamilies: Cercopithecinae (cheek-pouched monkeys) and Colobinae (leaf-eating monkeys) (Delson, 1992; Groves, 2001; Disotell, 2003). The higher-level relationships (e.g., subfamily, tribe) inferred from molecular studies (Page et al., 1999; Page and Goodman, 2001) are generally congruent with classifications based on morphological data (Delson, 1992; Goodman et al., 1998). However, considerable disagreement still exists for some relationships at the genus level, especially for subfamily Colobinae (Disotell, 2000).

The OWM phylogeny has been the subject of two Alu-based analyses (Xing et al., 2005, 2007). In a family-wide survey, 285 phylogenetically informative Alu insertions were collected from sixteen OWM taxa. A phylogenetic hypothesis with strong statistical support was inferred from these insertion events. The resulting relationships among all major clades are in general agreement with other molecular and morphological data sets. In the Asian colobine clade, a sister relationship between Nasalis and Pygathrix is strongly supported. This clade is subsequently joined by Trachypithecus, providing strong molecular evidence for the Asian colobine relationships.

The second study of OWMS focused on tribe Cercopithecini (guenons) (Xing et al., 2007). Two main phylogenetic questions concerning the Cercopithecini are: (1) whether terrestrial guenons (African green monkey, patas monkey, and L’hoesti’s monkey) form a monophyletic group and therefore the current genus Cercopithecus is paraphyletic; (2) what relationships exist among the major groups within Cercopithecini. A total of 179 Alu insertions were identified and genotyped in eleven cercopithecine genomes. These insertions were used to construct a robust phylogeny, which supports the following relationships: (1) all terrestrial guenons cluster in a single clade, supporting a single transition to a terrestrial lifestyle and the paraphyly of current genus Cercopithecus; (2) Allenopithecus is the basal lineage of the tribe, and the split between the arboreal Cercopithecus group and M. talapoin happened after the divergence of the terrestrial group. Another interesting observation is that a number of loci supported different clustering patterns within the arboreal guenons (genus Cercopithecus), suggesting a rapid basal radiation of these species and possible interspecific hybridizations after their speciation.

Human, chimpanzee, and gorilla trichotomy

The relationship among human, chimpanzee, and gorilla (i.e., the trichotomy problem) has been a particularly difficult, long-standing problem (Patterson et al., 2006), and a number of studies have attempted to
address this question. Although the consensus approach (Satta et al., 2000; O’Huigin et al., 2002) and mtDNA sequences (Horai et al., 1995) support the chimpanzee as the nearest living relative of humans, the evidence is not overwhelming and some results supported alternative scenarios (Hasagawa and Kishino, 1991; Ruvolo et al., 1991; Chen and Li, 2001). For example, Satta et al. (2000) analyzed sequences from 45 nuclear loci and found that 60% of the loci support the human–chimpanzee clade, whereas the remaining 40% support the two alternatives equally.

Salem et al. (2003c) characterized 117 AluY subfamily members along with 16 loci from other AluY subfamilies based on the human genome sequence. Using a total of 101 informative insertion loci, Salem et al. were able to infer a single most parsimonious tree with high levels of support (consistency index = 0.99; bootstrap support = 100% except one node). The resulting tree clearly clusters human and chimpanzee as a sister clade with gorilla as an out-group.

**POPULATION GENETIC STUDIES**

Just as retrotransposition of Alu and LINE-1 (L1) elements created an illuminating set of genetic markers for studies of primate evolution, their continued retrotransposition has generated a large set of more recent insertions useful in phylogenetic analyses: they are essentially free of homoplasy caused by convergence or reversion, the ancestral state is known, and genotypes can be determined using simple procedures and widely available equipment.

As markers for studying human population history, polymorphic Alu insertion loci have additional desirable properties. Recent Alu insertions are distributed widely and almost randomly throughout the genome, the majority appear to be selectively neutral or nearly so (Cordaux et al., 2006a), and they do not appear to strongly influence local recombination rates (IHMC, 2005; Myers et al., 2006). Thus Alu polymorphisms should accurately reflect demographic forces that have affected the entire genome. L1 insertions have not been studied in as much detail as Alu insertions, but they also appear to exhibit these properties (Myers et al., 2002; Boissinot et al., 2004; Song and Boissinot, 2007). Alu and L1 retrotransposition has been frequent enough to generate a steady stream of insertions (Deininger and Batzer, 2002), so human population history has been marked by new insertions. Taken together, these accumulated insertion polymorphisms constitute a large data set of independent genetic markers widely dispersed throughout the genome and evolving under a common and simple evolutionary model. This is exactly the kind of data needed to make detailed and reliable statistical inferences about human demographic history.

**Worldwide population genetic diversity and structure**

The frequency distributions of Alu and L1 insertion alleles at polymorphic insertion loci are typical of neutral genetic polymorphisms in humans (IHMC, 2005; Witherspoon et al., 2006). However, it must be noted that the allele frequency distribution for any type of genetic marker is sensitive to the methods used to identify loci that are polymorphic.

The earliest Alu and L1 insertion polymorphisms were identified incidental to surveys of loci of medical interest (Economou-Pachnis and Tsichlis, 1985; Dombroski et al., 1991; Wallace et al., 1991; Holmes et al., 1994). With the advent of the Human Genome Project, it became possible to quickly and efficiently find Alu and L1 insertions and identify the most recent ones by virtue of their high insertion frequency. Similarity to the reconstructed consensus sequences of the presumed “master” progenitor elements from which all these elements are generated. This is the first step in ascertaining these elements. Insertions whose sequences identify them as members of inactive Alu or L1 families are almost certainly so old that they have become fixed in the human population, so these—the vast majority of insertions—can be ignored. The much more manageable subset of loci that are likely to be recent insertions (and thus polymorphic for presence or absence) can then, in the second ascertainment step, be screened for polymorphism in panels of varying size and population diversity (Carroll et al., 2001; Myers et al., 2002; Salem et al., 2003b; Xing et al., 2003; Carter et al., 2004).

The initial identification step biases the data set toward insertions that have reached a high frequency in the population sample from which the public human genome sequence was obtained, and the second step introduces a modest bias toward insertions of intermediate frequency in the populations from which the screening panels were sampled. To eliminate these biases, molecular techniques have been devised specifically to find previously unknown low-frequency Alu and L1 polymorphisms (Boissinot et al., 2004). These techniques rely on the known sequences of Alu and L1 elements. The frequency distribution of the combined set of insertion polymorphisms ascertained by these different methods should approximate the distribution that would be observed with complete ascertainment of polymorphic loci in the largest ascertainment panels used.

Ascertainment biases can limit the power of genetic data sets to make inferences regarding demographic history. However, polymorphic loci with common minor alleles (>5%; “common polymorphisms” hereafter) are ideally suited for analyses of the structure of population structure. For such questions, common polymorphisms are more informative than rare polymorphisms (Bamshad et al., 2003; Witherspoon et al., 2006). Furthermore, under the assumption that most Alu and L1 polymorphisms are selectively neutral, intermediate-frequency insertion-present and insertion-absent (empty) alleles are treated identically by genetic drift and demographic forces. The alleles are therefore interchangeable for purposes of population structure analyses that rely on the signal of genetic drift, so a slight bias toward loci with higher frequencies of the insertion allele does not affect the result.

Alu and L1 insertion polymorphisms have been used to illuminate the patterns and distribution of human genetic diversity in many studies. In such studies, a set of common insertion polymorphisms is chosen, and genotypes at those loci are obtained for individuals sampled from diverse populations around the world. Within a population, these loci seldom deviate from Hardy–Weinberg equilibrium, underscoring their utility for population genetic analysis (Watkins et al., 2001, 2003).
The proportion of variance in allele frequencies that is due to differences between populations, \( F_{ST} \), depends on the choice of populations sampled and how they are grouped. \( F_{ST} \) estimates obtained from a set of \( Alu \) insertion genotypes range from 8 to 14%, for ~30 regional populations and three major continental groups (African, East Asian, and European), respectively (Watkins et al., 2001, 2003). \( F_{ST} \) estimates from a set of \( L1 \) insertion genotypes are similar (Sheen et al., 2000; Witherspoon et al., 2006). While relatively small, these \( F_{ST} \) values differ significantly from zero. Thus, although most genetic variation is due to differences between individuals within populations, the global human population has been structured rather than panmictic. Both \( Alu \) and \( L1 \) polymorphisms show higher average gene diversity (the proportion of individuals that are heterozygous at a locus, averaged across loci) in African populations (Watkins et al., 2001). This difference implies that African populations have maintained larger effective population sizes than non-African populations during recent evolutionary history (Stoneking et al., 1997).

These observations of nonzero \( F_{ST} \) and differences in genetic diversity between African and non-African populations imply that our genomes bear traces of past population migrations, geographic isolation, population fusion and fission, and changing population sizes. As early as 1994, Batzer et al. (1994) genotyped a set of four polymorphic \( Alu \) insertion loci in 664 individuals from 16 diverse populations. The four insertion alleles are found at different frequencies in each population, and this information can be used to compute the genetic distance (a measure of genetic dissimilarity) between the populations. Despite the small number of loci, a clear pattern emerged: populations that were geographically nearer to each other were genetically more similar. This positive correlation between genetic and geographic distance was confirmed and explored in greater detail with larger sets of \( Alu \) polymorphisms (Watkins et al., 2001, 2003) and corroborated with \( L1 \) polymorphisms (Witherspoon et al., 2006). These analyses showed that most of the correlation between genetic and geographic distance is due to the large genetic and geographic distances between African and non-African populations. While geographic distances between European, East Asian, and Indian populations vary over a wide range, their genetic distances are relatively more similar.

The pattern of genetic relationships between populations can be seen in a tree of populations constructed from the allele frequencies (see Fig. 4). The non-African populations cluster together and are separated from the African populations by a long branch. While a correlation between geographic and genetic location is apparent among non-African populations (European, Indian, and East Asian populations do form clusters), the genetic distances among all of these populations are relatively small. For example, the Indian populations are roughly the same geographic distance from the European and African populations, but the Indian populations are genetically much closer to the European and East Asian populations than to the African populations (see also Watkins et al., 2003). Principal components analyses of \( Alu \) and \( L1 \) data in these populations yield qualitatively similar results. In particular, the first principal component, which typically accounts for most of the variance in the data, separates sub-Saharan African populations from the others (Watkins et al., 2003; DJW, unpublished results). The second principal component typically distinguishes East Asian from European populations.

**Recent African origins hypothesis**

These findings support the recent African origins (RAO) hypothesis. This hypothesis states that the most recent population that was ancestral to all modern humans arose in Africa during the late Paleolithic. Migrants from that population then colonized the rest of the world less than 100,000 years ago, and, in the strict version of the RAO, replaced premodern humans without significant genetic mixing (Stringer and Andrews, 1988; Tishkoff and Verrelli, 2003).

Under the RAO hypothesis, anatomically modern human populations in Africa evolved in place, accumulating a substantial level of genetic diversity. At some later time, the rest of the world was colonized by a small
number of migrants drawn from the northeastern region of the African population. These migrants represented only a subset of the genetic diversity of the parent population, and still more of that diversity was lost due to genetic drift in the small populations that emigrated and spread rapidly around the world. The RAO hypothesis explains the higher genetic diversity in African versus non-African populations as the consequence of this “bottleneck” in population size, in contrast with the larger population size maintained by the African population. The long branch separating African from non-African populations likewise reflects the genetic drift that would have occurred in the small emigrating population. The relatively high genetic similarity among non-African populations (despite their wide geographic distribution) is attributed to their relatively recent founding. The older African subpopulations show greater genetic differentiation (see Fig. 4, longer branches in African cluster) because they have been drifting apart for much longer.

The frequency of Alu and L1 insertions is lower, on average, in African populations than in non-African ones (Watkins et al., 2003; Witherspoon et al., 2006). Since the ancestral state at these loci is known to be the absence of the insertion, an ancestral population must have existed that lacked these insertions entirely. That population would be more similar to African populations than to non-African groups, which is consistent with the RAO scenario above. However, at least part of this frequency difference may be due to the ascertainment bias for Alu and L1 insertions that have reached moderate to high frequency in non-African populations (Watkins et al., 2001; Witherspoon et al., 2006). Moreover, the relationship between the frequencies of ancestral alleles and ancestral populations is not simple and depends in part on mutation rates (Rogers et al., 2007). Therefore, this observation must be considered with caution.

Classification

Human population history can be inferred from the patterns of human genetic diversity only because the human gene pool is not thoroughly mixed. It is subtly structured, and that structure is a result of our complex demographic history, from our ancestral African roots through the most recent population fissions, fusions, and migrations. Since cultural, linguistic, and ethnic traits are also transmitted across generations, they are often correlated with the subtle genetic differences between groups. In particular, various categorizations of humans into “races” have been based on cultural and phenotypic traits; so they are correlated to some extent with detectable population genetic structure. This has raised the question whether genetic data can be used to reliably classify individuals into groups that correspond with self-identified ancestry or ethnicity. This question is of particular importance in biomedical research, since undetected population structure within a research population can lead to spurious results (Marchini et al., 2004).

Genetic structure within a population will manifest itself as departures from the patterns expected in a homogeneous population—e.g., violations of Hardy–Weinberg equilibrium and unexpected associations between alleles at independent or loosely-linked loci. The method implemented in the program Structure (Pritchard et al., 2000) assigns individuals to subpopulations in a way that minimizes those violations within subpopulations. The method thus identifies and classifies genetically related groups of individuals within a sample population. Romualdi et al. (2002) applied this method to polymorphic Alu insertion loci in humans. They concluded that the classification of individuals into populations was error-prone and that the continental population groups defined a priori (Africa, Asia, Europe, Australia and the Americas) were genetically indistinct. However, only 21 polymorphic Alu loci were analyzed. Subsequent analyses using the same methods and 100 polymorphic Alu loci (Bamshad et al., 2005) or 75 polymorphic L1 loci (Witherspoon et al., 2006) found clear evidence of continental population structure, but only when at least 50 loci were used. Data sets consisting of fewer loci (unless selected initially for large inter-population differences) typically lack the statistical power to detect the level of population structure that is present in continental human populations (Rosenberg et al., 2005).

The reliability with which individuals can be assigned into groups that correspond to their populations of origin depends not only on the number of loci, but also on the populations sampled. If relatively distinct populations are sampled (e.g., sub-Saharan Africans, East Asians, and Northern Europeans), 100 polymorphic Alu loci provide enough information to classify individuals into those three categories with nearly 100% accuracy (Bamshad et al., 2003). However, if more geographically intermediate populations are added (e.g., from the Indian subcontinent), the accuracy of classification into these four a priori groups suffers, even with 175 Alu and L1 loci combined (Witherspoon et al., 2006). This is due in part to a failure of the groupings to reflect the actual genetic structure in the data, and partly due to the challenge inherent in making finer distinctions between more similar individuals.

Figure 5 shows the results of a Structure analysis based on 175 Alu and L1 loci reported earlier (Watkins et al., 2003; Witherspoon et al., 2006). The optimal number of populations required to explain the population structure in the data is five, instead of the four that might have been presumed a priori (i.e., European, East Asian, sub-Saharan African, and Indian subcontinental). To adequately model the unusual differentiation of the Irula noncaste population of India requires a fifth population. However, other subpopulations that might seem equally unique (e.g., the Khonda Dora, another non-caste Indian population; or the group of Pygmy populations) are not modeled as separate populations. This result is similar to that of Rosenberg et al. (2002), in which a small isolated Pakistani population, the Kalash, formed a distinct group in a Structure analysis.

The block-like structure pattern in Figure 5 might suggest that human populations are discrete units, like islands: homogeneous within their boundaries and clearly distinguishable from their neighbors. To some extent, this appearance is due to the lack of geographically intermediate populations in this data set. An exhaustive sampling of world populations would almost certainly show a smoother gradation between neighboring populations (Witherspoon et al., 2006). However, even the apparently distinct genetic differences between the groups in Figure 5 do not imply that there are reliable categorical differences between individuals from different groups. Using the same set of 175 Alu and L1 insertions genotyped in the same individuals, Witherspoon et al. (2007) showed that two individuals sampled from different continental regions (European, sub-Saharan
African, East Asian, and Indian subcontinental) were often (25% frequency) genetically more similar than a pair drawn from within the same region. The number of loci used to compute genetic similarity (equivalently, relatedness) is critical: as more polymorphic loci are used, it becomes less likely that an individual will resemble a member of another population more than a member of his or her own population (see Fig. 6). Even with thousands of loci, however, that probability does not decrease to zero. When dealing with phenotypes whose genetic variations are determined by fewer than 20 loci, group labels are likely to be poor predictors of individual phenotypes (Witherspoon et al., 2007).

**Analyzing individual human populations and their histories**

As more genetic data are collected, ever-finer details of human population history can be resolved. The genetic effects of great distances and geographic barriers were the first to be detected, but more subtle barriers to gene flow exist, such as linguistic and social boundaries. In theory, a strong enough social barrier between populations, maintained long enough, could leave a genetic trace. This appears to be the case in some populations of India, where the caste system restricted intermarriage between groups for thousands of years. Although alleles and haplotypes are widely shared among castes, small genetic differences do exist, and these can be detected with sufficiently large data sets (Majumder et al., 1999; Bamshad et al., 2001). These differences probably reflect a combination of social strictures and migration of West Eurasian peoples into India around the time that the caste system arose.

**Alu** insertion polymorphisms have been used in many other samples and populations to test hypotheses concerning migrations and the sources of modern populations (Batzer et al., 1996; Sherry et al., 1997; Stoneking et al., 1997; Novick et al., 1998; Majumder et al., 1999; Comas et al., 2000; Bamshad et al., 2001; de Pancorbo et al., 2001; Nasidze et al., 2001; Roy-Engel et al., 2001; Antunez-de-Mayolo et al., 2002; Romualdi et al., 2002; Xiao et al., 2002; Bamshad et al., 2003; Bulayeva et al., 2003; Gonzalez-Perez et al., 2003; Mastana et al., 2003; Vishwanathan et al., 2003; Watkins et al., 2003; Barbalic et al., 2004; Comas et al., 2004; Cotrim et al., 2004; Dornelles et al., 2004; Maca-Meyer et al., 2004; Mansoor et al., 2004; Dunn et al., 2005; Mateus Pereira et al., 2005; Terreros et al., 2005; Herrera et al., 2007). In general, the results of studies of mobile element variation are in agreement with those based on other genetic systems, including mitochondrial (Budowle et al., 1999; Jorde et al., 2000; Bamshad et al., 2001), Y-chromosome (Jorde et al., 2000; Bamshad et al., 2001), microsatellite (Jorde et al., 1997; Bamshad et al., 2003), short tandem repeats (STR) (Shriver et al., 1997; Jorde et al., 2000; Budowle et al., 2001; Butler et al., 2003), and single nucleotide polymorphisms (SNPs) (Sachidanandam et al., 2001; Matise et al., 2003; Reich et al., 2003; Salisburry et al., 2003).

**FORENSIC APPLICATIONS**

The application of DNA to problems in forensics is an active field of research. For example, the Federal Bureau of Investigation’s Combined DNA Index System (CODIS) is a standard set of 13 STR loci that is routinely employed to match a suspect’s genetic material to the DNA found at a crime scene. Early in an investigation, before any definite suspects have been confirmed, however, tools that narrow the potential pool would be desirable. One method to reduce the list of possible suspects is to infer the ancestral origin of a DNA specimen found at a crime scene.

**Inferring human geographic origins**

A specific application to the forensic problem of determining ancestral origin of an unknown DNA sample was recently the subject of a study by Ray et al. (2005a). As background for this study, 710 individuals representing 31 different populations across four continents were evaluated for each of the 100 **Alu** insertion polymorphisms (Watkins et al., 2003). This database of genetic variation was then applied to determine the ancestry of 18 anonymous DNA samples culled from two forensic laboratories. The 18 individuals were genotyped at the same 100 **Alu** insertion loci, and the inferred geographic affiliation of the samples was calculated using the
samples in a forensic analysis. The methodology is also technically simple, requiring only standard thermal cyclers for PCR and the ability to run and visualize agarose gel electrophoresis.

Because many recent Alu insertions will be polymorphic for their presence or absence and have variable insertion frequencies among world populations (Watkins et al., 2003), the presence of these population-indicative insertions allows for hierarchical testing of subgroup affiliations. For example, once the initial Structure analysis assigns the DNA specimen to a continental affiliation such as Europe, Africa, Asia, or India, subsequent analyses, using only insertion loci that are useful within the subcontinental population groups, can be used to further isolate the subcontinental region. Identification of these subpopulation-specific loci is a difficult undertaking. Fortunately, a recent study (Cordaux et al., 2007) demonstrated that population-indicative Alu insertions can be identified preferentially. We believe that the application of this targeted identification will result in the rapid identification of additional markers that are likely to be more informative for specific population groups. Incorporation of these loci will reduce the number of loci needed for accurate identification of unknown samples.

Mobile element-based human sex identification

Another potential forensic application of Alu elements is human sex identification (Hedges et al., 2003). Determination of sex from human DNA samples is a common procedure in forensic laboratories. The most widely used approach is based on the Amelogenin locus (one of the standard CODIS loci), which yields different-sized PCR amplicons for the X and Y chromosome versions of the Amelogenin gene (Sullivan et al., 1993). However, this method can misidentify males as females in some cases due to a deletion in the AMEL Y region that reaches an allele frequency as high as 8% in some populations (Santos et al., 1998; Steinlechner et al., 2002; Thangaraj et al., 2002). Any source of error is a legitimate cause for concern. In fact, several researchers recommend that Amelogenin should not be relied upon as the sole indicator of sex (Santos et al., 1998; Brinkmann, 2002; Steinlechner et al., 2002; Thangaraj et al., 2002).

Fixed Alu insertions on either the X or the Y chromosome provide an alternative method of identifying the sex chromosomes present in a sample. Hedges et al. (2003) discovered two loci, AluSTXa and AluSTYa, that consistently differ with regard to insertion presence/absence on the X and Y chromosomes. They examined 778 human DNA samples from diverse populations and achieved 100% accuracy. The power of this test over the standard Amelogenin test derives from the statistical improbability of two independent mutations occurring in two separate genomic locations. Thus, the combination of these two markers (with or without the standard Amelogenin test) provides added assurance that sex test results are accurate.

Human DNA identification and quantitation

Another forensic task that must be considered is the identification and quantitation of human DNA in a complex mix (Chakraborty et al., 1999). Because of their high copy number, Alu elements represent a naturally amplified marker for the identification of primate and human DNA samples in a forensic analysis. The methodology is also technically simple, requiring only standard thermal cyclers for PCR and the ability to run and visualize agarose gel electrophoresis.

Because many recent Alu insertions will be polymorphic for their presence or absence and have variable insertion frequencies among world populations (Watkins et al., 2003), the presence of these population-indicative insertions allows for hierarchical testing of subgroup affiliations. For example, once the initial Structure analysis assigns the DNA specimen to a continental affiliation such as Europe, Africa, Asia, or India, subsequent analyses, using only insertion loci that are useful within the subcontinental population groups, can be used to further isolate the subcontinental region. Identification of these subpopulation-specific loci is a difficult undertaking. Fortunately, a recent study (Cordaux et al., 2007) demonstrated that population-indicative Alu insertions can be identified preferentially. We believe that the application of this targeted identification will result in the rapid identification of additional markers that are likely to be more informative for specific population groups. Incorporation of these loci will reduce the number of loci needed for accurate identification of unknown samples.

Mobile element-based human sex identification

Another potential forensic application of Alu elements is human sex identification (Hedges et al., 2003). Determination of sex from human DNA samples is a common procedure in forensic laboratories. The most widely used approach is based on the Amelogenin locus (one of the standard CODIS loci), which yields different-sized PCR amplicons for the X and Y chromosome versions of the Amelogenin gene (Sullivan et al., 1993). However, this method can misidentify males as females in some cases due to a deletion in the AMEL Y region that reaches an allele frequency as high as 8% in some populations (Santos et al., 1998; Steinlechner et al., 2002; Thangaraj et al., 2002). Any source of error is a legitimate cause for concern. In fact, several researchers recommend that Amelogenin should not be relied upon as the sole indicator of sex (Santos et al., 1998; Brinkmann, 2002; Steinlechner et al., 2002; Thangaraj et al., 2002).

Fixed Alu insertions on either the X or the Y chromosome provide an alternative method of identifying the sex chromosomes present in a sample. Hedges et al. (2003) discovered two loci, AluSTXa and AluSTYa, that consistently differ with regard to insertion presence/absence on the X and Y chromosomes. They examined 778 human DNA samples from diverse populations and achieved 100% accuracy. The power of this test over the standard Amelogenin test derives from the statistical improbability of two independent mutations occurring in two separate genomic locations. Thus, the combination of these two markers (with or without the standard Amelogenin test) provides added assurance that sex test results are accurate.

Human DNA identification and quantitation

Another forensic task that must be considered is the identification and quantitation of human DNA in a complex mix (Chakraborty et al., 1999). Because of their high copy number, Alu elements represent a naturally amplified marker for the identification of primate and human DNA samples in a forensic analysis. The methodology is also technically simple, requiring only standard thermal cyclers for PCR and the ability to run and visualize agarose gel electrophoresis.
human-specific DNA. Several systems have been developed to identify and quantify human DNA using PCR primers designed to amplify the core Alu sequence (Sifis et al., 2002; Nicklas and Buel, 2003a,b,c, 2005). During 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 the Yb8 and Yd6 subfamilies of Alu (Walker et al., 2003b). These techniques can detect and quantify human DNA ranging from 10 to 0.001 ng, even in samples that involve mixtures with nonhuman contaminants (Walker et al., 2003a).

**Primate identification**

The application of forensic techniques is not limited to human-specific problems. The identification of illegally taken wildlife and determination of management units in conservation efforts are important tasks that rely on similar methodologies. Just as humans have their own specific subfamilies of Alu elements, other primate lineages have as well. New World primates, for example, harbor Alu subfamilies that are distinct from all other primates (Ray and Batzer, 2005). Thus, any primate DNA could potentially be identified and quantified using the same techniques as long as one knows the Alu subfamilies in question.

Techniques that lead in this direction have already been developed. For example, in laboratories dealing with primate genetics, it is critical that researchers ascertain that they are handling DNA from the appropriate taxa. Very often researchers collect or receive DNA that was collected in a noninvasive manner (i.e., “divorced” tissues such as hair or feces) (Kohn et al., 1995; Taberlet et al., 1997; Matsubara et al., 2005). This is especially common in investigations of the illegal wildlife trade and identification of seized products (Matsubara et al., 2005; Yan et al., 2005; Domingo-Roura et al., 2006). Even when laboratories deal in their own “homegrown” DNA via cell culture, cross-contamination can occur among cell cultures and within concurrent large-scale DNA extractions from multiple species. Furthermore, simple mishandling of well-documented samples may lead to mislabeling or cross-contamination.

Just as taxonomists and museum curators have long relied on detailed analysis of morphological characters in the form of dichotomous keys to identify unknown samples, molecular biologists can now develop their own molecular keys to do the same. Previously, mitochondrial haplotypes had been used as a method for species identification (Kohn et al., 1995; Palumbi and Cipriano, 1998; Hebert et al., 2004). Unfortunately, complications involving maternal inheritance, primer failure, and nuclear translocation of mitochondrial genes have limited the applicability of this approach (Moritz and Cicero, 2004; Lorenz et al., 2005).

Under extreme situations (e.g., highly degraded, extremely small quantities of DNA samples), high copy-number mitochondrial DNA might be the only option for forensic studies. However, several factors make a SINE-based dichotomous key for identifying samples attractive for most primate forensic studies. First, genotyping SINEs require only simple equipment to which even modest molecular biology laboratories have access. Second, the technique is reliable even in the hands of a novice because no special technical expertise is required. Third, unlike plastic morphological characters or potentially confusing sequence data, this type of SINE-based data consists solely of presence–absence states. Such data are easy to interpret in a hierarchical fashion to rapidly narrow the possibilities regarding the identity of anonymous DNA samples. Fourth, in contrast to mitochondrial DNA methods, “universal” primers are used in a simple PCR assay to amplify loci from most relevant taxa. Finally, unlike “single locus” (e.g., mainly mitochondrial DNA) systems, SINE-based dichotomous keys identify terminal branches based on multiple unlinked loci instead of the single linkage group exemplified by the maternally inherited mitochondrial genome. One such SINE-based dichotomous key was recently developed for a variety of primates (Herke et al., 2007).

**THE IMPACT OF MOBILE ELEMENTS ON PRIMATE GENOME EVOLUTION**

Although initially considered “junk DNA,” mobile elements have proven to have profoundly influenced primate genomic evolution. In this section, we will review several aspects of mobile elements’ impact.

**Mobile elements and human diseases**

Mobile elements can cause human diseases both during and after their integration (Deininger and Batzer, 1999; Batzer and Deininger, 2002; Callinan and Batzer, 2006; Chen et al., 2006). The insertion of mobile elements can directly disrupt genes or alter gene expression and thus lead to disease. After mobile-element integration, nonhomologous recombination between similar elements can generate disease-causing deletions. The amplification rate of mobile elements has varied dramatically during the primate evolution leading to the human lineage (Hedges et al., 2004, 2005). It is estimated that the current rate of Alu amplification in the human genome is of the order of one Alu insertion in every 200 births (Deininger and Batzer, 1999). To date, alterations caused by mobile elements have been identified in more than 100 human genetic disorders, including cases of hemophilia type A and B, familial breast cancer, and neurofibromatosis type 1 [see (Deininger and Batzer, 1999; Batzer and Deininger, 2002; Callinan and Batzer, 2006; Chen et al., 2006) for reviews].

**Mobile element exonization and mobile element-mediated exon shuffling**

One of the most direct impacts of mobile elements is their contribution to the formation of new genes and gene families. Some mobile elements contribute new genetic material that eventually becomes functional in itself or becomes part of a larger gene via a process termed “exonization” (Lev-Maor et al., 2003). Examples include: (1) the human SETMAR gene, formed from the fusion of a domain that exhibits histone methyltransferase activity (the SET domain) and the entire transposase-coding region of a Hsmar1 transposon (the MAR domain) (Cordaux et al., 2006b) and (2) the human BC200 gene, which is derived from an Alu element and encodes a brain-specific non messenger RNA (Kuryshchev et al., 2001). Many additional examples have also been reported (Sorek et al., 2002; Lorenc and Makalowski, 2003; Krull et al., 2005).
Another mechanism by which mobile elements contribute to gene creation is termed sequence transduction or 3′ transduction (Holmes et al., 1994). During this process, genomic materials downstream from a retrotransposon are transcribed and mobilized along with the retrotransposon. This may result in the duplication of exons to the new location or the alteration of gene expression at a new location through the movement of promoter and enhancer sequences. This process, which was initially reported in cell culture retrotransposition assays (Morish et al., 1999) and subsequently in the human genome, leads to duplication of exons and even whole genes (Rozmahel et al., 1997; Ejima and Yang, 2003; Xing et al., 2006).

**Mobile elements and primate genomic structural variation**

In addition to their direct impact on gene expression, mobile elements also represent one of the important sources of structural genomic variation. Retrotransposon insertions have contributed to a 15–20% expansion of the human genome compared to strepsirrhine genomes (Liu et al., 2003). Several other mechanisms by which mobile elements shape primate genomes have been reported recently, including retrotransposon insertion-mediated deletion in the genome (Callinan et al., 2005; Han et al., 2005), mobile element-mediated nonhomologous recombination (Sen et al., 2006), and mobile element-mediated gene conversion (Kass et al., 1995; Roy et al., 2000; Otem et al., 2003a; Vincent et al., 2003; Otieno et al., 2004; Zhi, 2007).

**Functional role of “junk DNA”**

Whether seemingly “selfish” mobile elements possess any function has been a fundamental question since mobile elements were first recognized as a major component of mammalian genomes. Their prevalence in eukaryotic genomes suggests that mobile elements may be on the whole, selectively advantageous, and several studies suggest that they play important roles in genome structure and gene expression (Brosius and Gould, 1992; Vidal et al., 1993; Schmid, 1998; Hamdi et al., 2000; Deininger and Roy-Engel, 2002; Labrador and Corces, 2002; Nouaud et al., 2003). One of the more compelling examples is the involvement of L1 elements in the human X chromosome inactivation process [Lyon’s “repeat hypothesis” (Lyon, 1998)]. L1 elements compose ~30% of the human X chromosome, and recent studies suggest that L1 elements may function as “way stations” for X inactivation signals (Bailey et al., 2000; Hansen, 2003; Ross et al., 2005). Other roles that have been proposed for mobile elements include double-strand DNA break repair (Morris et al., 2002), the creation of new CpG islands (short stretches of DNA, which have a high frequency of C-G dinucleotide sequences and are usually found near the promoter regions of genes) (Schmid, 1998), and the stimulation of translation during cell stress and viral infection through the regulation of double-stranded RNA-activated protein kinase (Chu et al., 1998). However, the putative functional role of mobile elements as a whole is still a topic of intense debate.

**CONCLUSIONS**

As this review has shown, mobile elements are proving highly useful in answering questions about primate phylogeny, human population affinities, and forensics. In addition, because of their unique mechanistic properties and ubiquity in the genome, mobile elements exert important effects on recombination, deletion and duplication, DNA sequence variation, and gene expression.

Many taxa, including a number of nonhuman primates, have not yet been characterized for mobile element variation. Such studies will improve the potential to resolve thorny phylogenetic questions, and they can be used to address issues about species conservation and identification. Another question that can be addressed by mobile elements is the relationships of primates with other mammals (i.e., the superordinal taxonomic position of primates).

In humans, a reference set of 100 Alu insertion polymorphisms has been established and typed in more than 700 individuals (Watkins et al., 2003). An additional reference set of 75 L1 insertion polymorphisms has also been compiled and analyzed (Witherspoon et al., 2006). Additional human populations can now be genotyped for these polymorphisms in comparative studies of human population origins and history. All of these efforts will be enhanced by new technical developments, such as microarray-based analysis of mobile elements. Clearly, much remains to be learned about the evolution and consequences of this fascinating component of our genome.

**ACKNOWLEDGMENTS**

The authors thank the editor and the three anonymous reviewers for their constructive and valuable comments. In addition, valuable comments and assistance were provided by Liz Marchali, Alan Rogers, Kazima Bulayeva, and W. Scott Watkins.

**LITERATURE CITED**


Murata S, Takasaki N, Saitoh M, Tachida H, Okada N. 1996. Details of retropositional genome dynamics that provide a rationale for a generic division: the distinct branching of all the pacific salmon and trout (Oncorhyncus) from the Atlantic salmon and trout (Salmo). Genetics 142:915–926.


MOBILE ELEMENTS IN PRIMATE AND HUMAN EVOLUTION


Yearbook of Physical Anthropology—DOI 10.1002/ajpa


MOBILE ELEMENTS IN PRIMATE AND HUMAN EVOLUTION


