Journal of Cellular Biochemistry 102:110–121 (2007)

Most Human Alu and Murine B1 Repeats Are Unique

Boris Umylny,[1] Gernot Presting,[2] Jimmy T. Efird,[3] Boris I. Klimovitsky,[1] and W. Steven Ward[1]*

[1]Institute of Biogenesis Research, John A. Burns School of Medicine, University of Hawaii, Honolulu, Hawaii 96822
[2]Department of Molecular Biosciences and Bioengineering, John A. Burns School of Medicine, University of Hawaii, Honolulu, Hawaii 96822
[3]Asia-Pacific Institute of Tropical Medicine and Infectious Diseases and Biostatistics and Data Management Facility, John A. Burns School of Medicine, University of Hawaii, Honolulu, Hawaii 96822

Abstract Alus and B1s are short interspersed repeat elements (SINEs) indirectly derived from the 7SL RNA gene. While most researchers recognize that there exists extensive variability between individual elements, the extent of this variability has never been systematically tested. We examined all Alu elements over 200 nucleotides and all B1 elements over 100 nucleotides in the human and mouse genomes, and analyzed the number of copies of each element at various stringencies from 22 nucleotides to full length. Over 98% of 923,277 Alus and 365,377 B1s examined were unique when queried at full length. When the criterion was reduced to half the length of the repeat, 97% of the Alus and 73% of the B1s were still found to be a single copy. All single and multi-copy sequences have been mapped and documented. Access to the data is possible using the AluPlus website http://www.ibr.hawaii.edu. J. Cell. Biochem. 102: 110–121, 2007.

Key words: Alu; B1; SINE; human repeated elements; mouse repeated elements

Human Alus [Deininger et al., 1981] and mouse B1s [Krayev et al., 1980] are short interspersed elements (SINEs) indirectly derived from 7SL RNA, a signal recognition particle involved in translation of eukaryotic secreted proteins [Ullu and Tschudi, 1984]. These repeats occupy a significant portion of the human (10.7%) and mouse (2.7%) genomes [Lander et al., 2001; Waterston et al., 2002]. The human Alu is a dimer consisting of two similar elements, linked by a short poly-A chain with a total length of approximately 300 nucleotides (nt) [Jurka and Milosavljevic, 1991; Quentin, 1992]. The mouse B1 is a monomer of approximately 140 nt in length, with an internal 29-nucleotide duplication [Labuda et al., 1991]. The Alu 5′ monomer (FLAN) shares significant homology with certain proto-B1 (pB1) sequences [Quentin, 1994]. It is believed that Alu and B1 sequences are propagated by a reverse transcriptase encoded by the L1 family of long interspersed repeat elements (LINEs) [Schmid, 1998]. While there is no confirmed function for Alu or B1 elements, various stress conditions increase both expression [Chu et al., 1998] and L1-mediated retroposition [Hagan et al., 2003], suggesting the possibility of function [Schmid, 1998].

Alu and B1 sequences are preferred methylation targets [Hellmann-Blumberg et al., 1993; Kochanek et al., 1993; Jeong and Lee, 2005], accounting for 33% of the total genomic methylation sites [Schmid, 1998]. In the genome, Alu and B1 sequences appear to be preferentially distributed in GC-rich regions [Lander et al., 2001; Waterston et al., 2002]. The B1 distribution of the mouse genome exhibits a greater correlation with the Alu content of the orthologous areas of the human genome than with the immediate GC-density [Waterston et al., 2002]. This suggests that genomic features, which are correlated with but distinct from GC-content, may determine Alu/B1 distribution [Waterston

Grant sponsor: NIH; Grant number: HD28501; Grant sponsor: NIH; Grant number: G12RR003061.

*Correspondence to: Dr. W. Steven Ward, PhD, University of Hawaii at Manoa, Institute for Biogenesis Research, 1960 East-West Road, Honolulu, HI 96822.
E-mail: wward@hawaii.edu

Received 29 November 2006; Accepted 3 January 2007
DOI 10.1002/jcb.21278

et al., 2002]. On chromosomes 21 and 22, a positive correlation exists between the distribution of Alus and exons [Blinov et al., 2001]. Notably, Alus are not uniformly distributed across chromosomes [Grover et al., 2003, 2004].

Alu subfamilies are classified by age. The oldest, Jo and Jb, were derived from a single ancestral gene 81 million years ago [Kapitonov and Jurka, 1996]. The intermediate S subfamilies (Sx, Sp, Sq, and Sc) have an estimated age of approximately 35–48 million years [Jurka and Milosavljevic, 1991]. The youngest group, previously known as Sb, includes the Y subfamilies [Rowold and Herrera, 2000]. It is estimated that some of the Y Alus may be as young as 3–4 million years [Kapitonov and Jurka, 1996], and some of these may still be undergoing active retroposition [Jurka et al., 2002].

Both the human and mouse sequencing projects reported a large variation between individual Alus and B1s, ranging from 1% to 40% from the consensus sequence [Lander et al., 2001; Waterston et al., 2002]. A recent classification of approximately half of all human Alus resulted in the identification of over 200 subfamilies [Price et al., 2004]. These results imply that significant variability exists in the sequences of 7SL SINEs. In this manuscript, we verify whether there exists sufficient variability in Alu and B1 sequences to enable generation of sequence specific primers and probes.

MATERIALS AND METHODS

Software

Bioperl was used with Ensembl's release 31 Perl API libraries [Hubbard et al., 2005], GNU bash version 3.00.00 and Perl version 5.8.5 for scripting, GNU gcc version 3.3.4 for handling procedures that required enhanced performance and MYSQL version 12.22 for storing Ensembl data as well as the primary data repository. Analyses were performed on an Intel platform running SUSE version of Linux (distributed by Novell).

Identification of Alu and B1 Repeats

RepeatMasker version 3.1.0 from http://www.repeatmasker.org [Smit et al., 1996–2004] was used to identify Alus and B1s in the FASTA files from human and mouse genomes release 31 from Ensembl's website (http://www.ensembl.org/index.html). RepeatMasker requires two external packages—WU-BLAST [Gish, 1996–2004] to compare sequences and Repbase libraries [Jurka et al., 2005] that contain SINE repeat consensus sequences. The repeat libraries release from January 12, 2005 and WU-BLAST version 1.05 were used in our analyses. The addresses and lengths of Alu and B1 sequences were taken directly from the RepeatMasker's "out" files. These addresses include the poly-A tail at the ends of the sequences. Addresses and lengths of genes were taken directly from Ensembl release 31. A total of 1,160,797 human Alus and 498,420 mouse B1s were identified. These values corresponded to 9.9% of the human and 2.2% of the mouse genome and compared favorably with recent estimates reported in the literature [Waterston et al., 2002].

The repeats were identified and loaded into a MYSQL database using custom Perl and bash scripts and 'C' coded executables. The reports were generated from the database using custom Perl scripts. All custom programs and scripts as well as database schema are available from http://www.IBR.hawaii.edu.

Our search for Alu sequences greater than 200 nt and B1 sequences greater than 100 nt identified 923,280 Alus and 365,382 B1s. A previous report identified 930,250 Alu sequences with lengths greater than 200 nt [Dagan et al., 2004], closely matching the above estimates for the number of Alu repeats. A total of 5 B1 repeats contained at least 1 'N' nucleotide—indicating incomplete sequencing. These five were excluded from the data set. Further, three Alus contained at least one 'N' nucleotide and were subsequently excluded from the data set.

Scanning the Human and Mouse Genomes for Copies of Alu and B1 Repeats

As the focus of this study was on near-full length sequences that might potentially have a function, we restricted our search to Alus and B1s greater than 200 and 100 nt, respectively. Alu sequences over 200 nt and B1 sequences over 100 nt were extracted from Ensembl's FASTA files using their address as determined by RepeatMasker. Each Alu and B1 element was then compared to its respective genome using BLAST [Altschul et al., 1997] from NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) to search for matches. The word size was set to

200 for Alus and 100 for B1s and other parameters were left at their default values.

BLAST default behavior is to filter low-complexity repeat sequences, including poly-A tails found on many Alus and B1s. To avoid these artifacts, we used BLAST with the filtering flag off. In all references to percent of sequence length the length of the Query, not the Target sequence will be considered.

Self-Validation of BLAST Results

RepeatMasker uses WU-BLAST to identify repeats. In a separate approach, we used NCBI-BLAST to align these repeats with the entire human and mouse genomes. Using two different implementations of BLAST provided an important control for the results produced by both programs. RepeatMasker reported a set of sequences it identified as Alus and B1s. When these sequences were aligned against the source genomes using NCBI-BLAST, every sequence was matched at least once against itself and all matches were limited to the set of sequences identified by WU-BLAST.

Validation

We randomly selected 3,500 B1s from the 358,574 B1 sequences identified as single copy at full length and 100% identity and matched them against the entire mouse genome using a custom pattern matching application. Each of the 3,500 B1s corresponded to a single locus in the mouse genome. We also randomly selected 10,000 single-copy Alus as well as 675 3-copy B1s and 686 4-copy Alus, all based on full length, 100% identity. All were checked using the same custom pattern matching application. In each case, the results were identical to the ones reported by BLAST.

Scanning the Alu and B1 Repeat Sequences With Respect to One Another

The relatively large 200/100 nt word size allowed the alignment of every Alu and B1 repeat with the entire source-genome, but also introduced some limitations. For example, when using BLAST to identify sequences identical to relatively short SINE queries, a word size of 100 can prevent detection of elements differing by a single nucleotide. To address this problem BLAST was run using a 22 nt word size on all selected Alu and B1 sequences (i.e., those over 200 and 100 nt) with respect to one another. As we were primarily interested in 100% identity matches, gapped alignment was suppressed. As a control for this analysis, we also ran an un-gapped self-BLAST with 22 nt word size on human and mouse RefSeq transcripts [Pruitt et al., 2005]. To prepare control data that would be comparable to Alu and B1 databases, the long sequences of human and mouse RefSeq FASTA files were fragmented into sequences of 295 nt for human and 131 nt for mouse. These numbers were selected to match the average lengths of Alus and B1s in the data sets. Next, NCBI-BLAST was used to perform an un-gapped alignment of these fractions against the unfractionated RefSeq databases with 22 nt word size. All parameters were identical to those used for Alu/B1 self-BLAST.

Determining the Positions Relative to Genes and the GC Content of the DNA

The population of SINEs over 200/100 nt was sorted into four libraries: SINEs located within (i) introns (intronic), (ii) exons (exonic), (iii) 500 nt of genes (close), or (iv) more than 500 nt from any gene (distant). For each library, the percentage of GC-nucleotides was computed based on the total number of GC and AT nucleotides. If other nucleotides were found within the sequence (e.g., 'N'), these nucleotides were not used in the calculation. The GC content of the chromosomal DNA, defined as the DNA within one repeat length 5′ and 3′ to the SINE, was also computed. These calculations were performed using custom software. As a control, the same program was used to compute the GC content of the entire human and mouse genomes. The results were 41% GC content for human and 42% for the mouse, which is identical to that reported [Lander et al., 2001; Waterston et al., 2002].

Determining Age-Based Distribution of Alu Repeats

Using results of the NCBI-BLAST at a 22 nt word size, we classified all repeats based on the number of times they were found in the data set at 100% identity over 200 nt. The 200 nt length, which represents approximately 68% of the average Alu length in our dataset, was chosen to allow for mutations. These repeats were classified as single copy, 2–10 copy, 11–100 copy, and greater than 100 copy. Within these categories we used RepeatMasker's Repbase classification to separate repeats into

broad age-based categories—oldest, intermediate and youngest—corresponding to the J, S, and Y subfamilies.

Detection of Full Length Exact Alu Copies

To detect full length exact copies of Alus we search the database of BLAST hits for all Alus (source) that matched at full length and 100% identity to different Alus (target) of the same length. The data were then validated by verifying that the target Alus also matched the source Alus at full length and 100% identity.

Segmental Duplication

To check if perfectly matched Alus resulted from segmental duplications, we adapted a previously published approached [Cheung et al., 2003]. We generated 5,000 nucleotide segments around the perfectly matched elements by extending the sequences 2,500 nucleotides in both directions. The sequences were then self-aligned and matches longer than 4,000 nucleotides at 90% or greater identity were identified as results of possible segmental duplications.

Statistical Analysis

Standard linear model and exact permutation statistical procedures were used to analyze the data in this study. Continuous variables were characterized by a point estimate of central tendency. Underlying normality was assessed using multivariate log-normal plots. When necessary, an appropriate normalizing and/or variance stabilizing transformation were applied to the data. P-values < 0.05 were considered to be statistically significant. When appropriate, statistical models were adjusted for multiplicity using a sequential Bonferroni correction.

RESULTS

Most Alu and B1 Nucleotides Are in Sequences of Greater Than 200 and 100

We selected the entire populations of human Alu sequences of length greater than 200 nt and mouse B1 sequences of length greater than 100 nt. These minimum length requirements were imposed because, if SINEs do have a normal physiological function as non-coding RNA [Liu et al., 1995; Chu et al., 1998; Hagan et al., 2003; Vila et al., 2003], the function would more likely be retained in repeats that are closer to their consensus length. The average length for the selected Alus was computed to be 295 nt and for the selected B1s 131 nt. We found that 90% of Alu nucleotides occur in sequences greater than 200 nt (Fig. 1A), and 83% of B1 nucleotides occur in sequences of greater than 100 nt (Fig. 1B).

Most Human Alus and Mouse B1s Are Single-Copy Sequences

We searched the entire human genome for Alus greater than 200 nt using BLAST [Altschul et al., 1990] and found that, when requiring 100% identity over full length, over 98% are single-copy sequences (Table I). Of the remainder, most tended to be 2–10 copy sequences and a small percentage of Alus had up to 450 copies. We refer to every repeat sequence that at a given set of criteria contains more than one exact copy in the genome as a multi-copy sequence. We also mapped the entire population of mouse B1

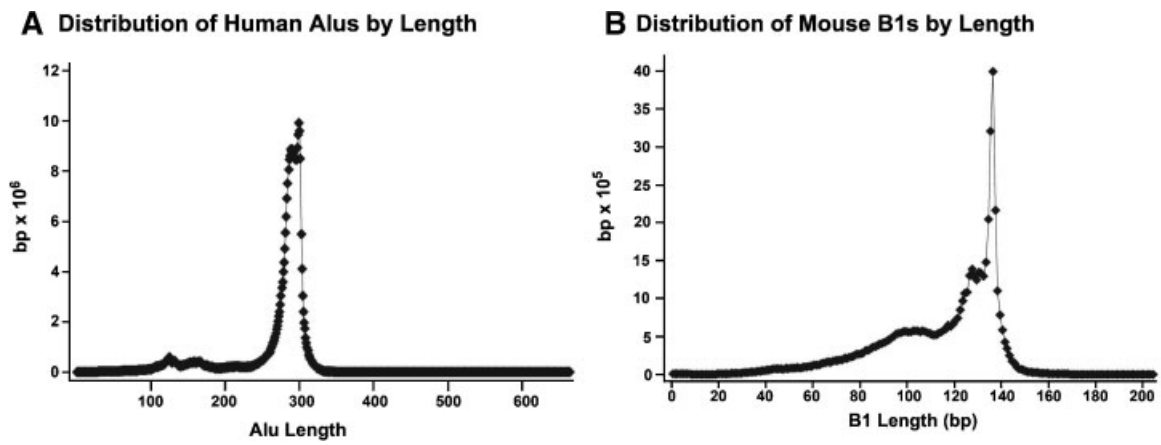

Fig. 1. Distribution of (**A**) Alu and (**B**) B1 sequences by length in nucleotides.

TABLE I. Copy Number of Human Alus and Mouse B1s

	Human		Mouse	
	No. seqs.	% Alu repeats	No. seqs.	% B1 repeats
Total	923,277		365,377	
Matched	923,277	100	365,377	100
Copy no.				
Single	911,156	98.7	358,574	98.14
2–10	11,487	1.2	5,649	1.55
11–100	489	0.05	1,149	0.031
101–200	103	0.01	5	0.001
201–1000	42	0.005	0	0
Highest	450		190	

Copy number of human Alus and mouse B1s from the respective genomes matched at 100% identity over full length.

sequences of length greater than 100 nt to their genomic locations. Our search of the complete mouse genome for B1 sequences using the same criteria, that is, 100% identity over full length, found that over 98% were single-copy sequences. As in the case of human Alus, most of the repeated B1s tended to occur 2–10 times and a small percentage of B1s had up to 190 copies (Table I).

Using a word size of 200 nt for the Alu BLAST and 100 nt for the B1 BLAST search was an efficient method for finding full length, 100% matches in the genome, but limited our ability to identify shorter matches as well as SINEs with a single mutation. We therefore performed a second set of experiments in which we used BLAST with a 22 nt word size to compare every Alu and B1 sequence against the entire population of the respective repeat elements. The percent of Alus and B1s that were single copy was plotted as the criterion was gradually relaxed from a 100% match at full length to a 100% match at 22 nt (Fig. 2). We found that when the match size approached half of the average size of the repeat, 97% of the Alus and 73% of the B1s were still single copy. No Alus and only 7.3% of B1 sequences were unique at a match length of 22 nt.

For comparison, we performed the same analysis on a data set of known mRNA sequences taken from the RefSeq database of known transcripts, as described in Materials and Methods (Fig. 2). At full length, Alus and B1s contain a larger proportion of single-copy sequences than RefSeq fractions of the same length. However, at shorter lengths the relationship between individual Alu and B1 elements becomes obvious. Below a certain length—approximately 24% (70 nt) for Alus and 30% (40 nt) for B1s—the percentage of

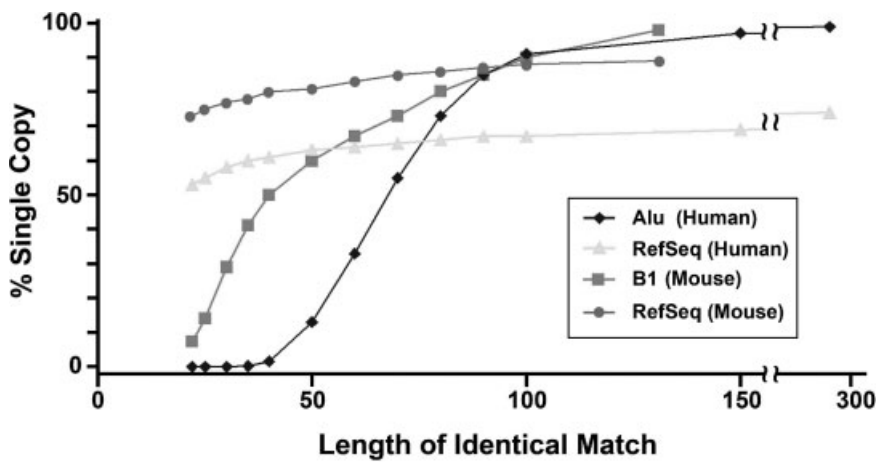

Fig. 2. Single-copy content based on the length of the match. The criterion for the length of identical match was gradually decreased from full length to 22 nt in human Alus (diamonds), and mouse B1s (squares). These were compared to human RefSeq transcripts (triangles) and mouse RefSeq transcripts (circles).

single-copy SINE sequences declines dramatically, while the percentage of single-copy RefSeq fractions declines at a much slower pace. For example, below 35 nt nearly all Alus are multi-copy sequences (Fig. 2). While the poly-A tails and Alu linker regions could be the causes of these very short matches, we believe that further analysis of identical portions of Alus and B1s might prove interesting.

Distribution of the Multi-Copy SINES

Our criteria for single-copy SINEs implied that the corresponding criteria for multi-copy SINEs were very strict; that is, two sequences had to contain 100% identity along the full length to be called copies. However, even at this high level of stringency, we found a significant number of these repeat. We examined these multi-copy repeats in several ways. First, we questioned whether multi-copy Alus were restricted to the youngest Alu subfamilies, since these were the most actively transposing elements in recent evolutionary history. We found that while most highly repeated Alus with more than 11 copies belong to the youngest subfamily, those with 2–10 copies, which represent the bulk of the multi-copy elements, have a distribution that is quite similar to the distribution of all Alu repeats (Tables II and III). The age-based distribution of Alus with 2–10 copies was not found to differ significantly from the age-based distribution of all Alus ($P = 0.23$). Over 62% of all Alus with multiple copies belong to the oldest and intermediate subfamilies (Table II). Furthermore, even at 200 nt, 94% of the youngest Alus were single-copy sequences (Table III).

Distribution of Multi-Copy Alus and B1s by Chromosome

We examined the distribution of multi-copy 7SL RNA SINEs by chromosome. In human and mouse, chromosome Y is enriched for multi-copy Alu (Fig. 3) and B1 (Fig. 4) elements. However, both single and multi-copy Alus and B1s are found on all chromosomes. For example, Alu 1:57,691, which is 208 nt long and has 450 matches (Fig. 5) is distributed on all chromosomes, with the number of copies being nearly proportional to the length of the chromosome. This particular sequence also intersects or lies in proximity of 140 different genes (Table IV).

Some Multi-Copy SINEs Are Truncated Forms of Larger SINE

We analyzed the structure of the multi-copy Alu with the highest copy number, the 209 nt Alu 1:57691, which is repeated exactly 450 times at full length and 100% identity. We found that all 449 copies were larger than Alu 1:57691 (Fig. 6), indicating that Alu 1:57691 is nested within 449 other, larger Alus. When we analyzed the number of times each of the 449 copies of this Alu was, itself, repeated exactly, we found that approximately one third (157) of them are single-copy elements (Fig. 6). The remaining 292 sequences could be divided into two multi-copy categories. Repeats in the first category are truncated versions of larger copies and vary in length (e.g., 4:45326). Elements in the second category, represented as the steps in Fig. 6, are exact, full-length copies of each other. One example is Alu 5:50201, which is repeated exactly 67 times. Interestingly, the 67 copies of this Alu are distributed over all chromosomes (Fig. 5).

Alu and B1 Sequences Appear to be GC-Rich and Their Surrounding Sequences Have GC-Content Comparable to the Genome

The sequences flanking human Alus (see Materials and Methods) have GC-content equivalent to that of the human genome (Table V, $P > 0.05$), while the Alu sequences appear to have significantly higher GC-content

TABLE II. Distribution of Human Multi-Copy Alus by Age and Copy Number

		% of total multi-copy Alu repeats		
Age	Subfamily	2–10 Copies	11–100 Copies	>100 Copies
Oldest	J	18	1	0
Intermediate	S	58	2	0
Youngest	Y	24	97	100
	Total %	83	7	10

A sequence is defined as multi copy if it had more than one 100% identity match over 200 nt.

TABLE III. Distribution of Human Alus by Age

	Subfamily		% of all Alu repeats	
Age	Name	% Single copy	Total	Single copy
Oldest	J	99	22	23
Intermediate	S	98	63	63
Youngest	Y	94	15	14

A sequence is defined as a single copy if it had only one 100% identity match over 200 nt.

($P < 0.0001$). Mouse B1s, likewise have GC-content significantly higher than the genome ($P < 0.0001$), while the GC-content of the surrounding sequences may actually be slightly lower than the average for the genome ($P = 0.04$).

Full Length, Exact Alu Copies do not Appear to be the Result of Segmental Duplications

We differentiated between what we term nested exact copies and full length exact copies of Alus. A nested exact copy is an exact copy of a full length Alu that is also contained within a different Alu at another chromosomal location (for e.g., Alu 1:57691 is nested within 449 other, longer Alus, as shown in Fig. 6). A full length exact copy of an Alu is an Alu located at another chromosomal location that is an exact copy of the first, without having any additional Alu sequences (for e.g., Alu 5:50201 is repeated exactly 67 times throughout the genome, 100%, full length, and none of these copies have additional recognizable Alu sequences, as shown in Fig. 6). We detected a total of 10,139 Alu sequences that are full length, exact copies of other Alus at full length 100% identity. These elements appear to be well distributed with respect to genes (Table VI) and subfamilies (Table VII). Our check for duplicated DNA segments indicated that propagation of only 112 of these Alus could be attributed to segmental duplications. The copy numbers of these full length, exact copies were lower than for the total number of repeated Alus (compare Tables I and VIII).

AluPlus Website

The database of SINE relationships used in this study is available online via the AluPlus system (http://www.ibr.hawaii.edu). This system provides access to the integrated database of nucleotide sequence-based relationships between 7SL RNA-derived sequences for humans (Alus) and mouse (B1s) combined with Ensembl-based genomic feature information. A list of Alus or B1s that match the search criteria with their genomic locations, complete nucleotide sequences and copy number is generated with each search. Additionally, a search for a particular Alu/B1 sequence will provide information regarding genomic features (genes, introns, and exons) that all copies of that sequence intersect. The user also may supply a number of nucleotides that define proximity around known genes. Alus/B1s that lie closer to

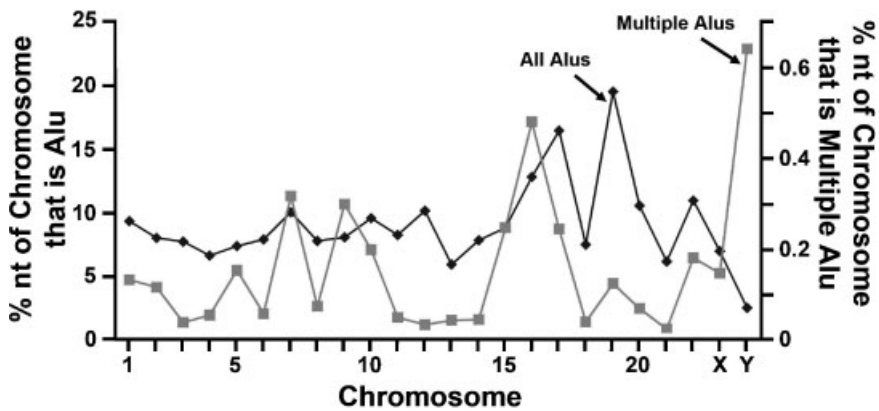

Fig. 3. Distribution of human Alus by chromosome. Distribution of all human Alus greater than 200 nt and all multi-copy human Alus by chromosome. The diamonds are the percentage of all Alus on a particular chromosome (left hand scale). The squares are the percentage of multi-copy Alus on a particular chromosome (right hand scale). While chromosome Y does not contain an exceptional number of Alus, nearly all are multi-copy.

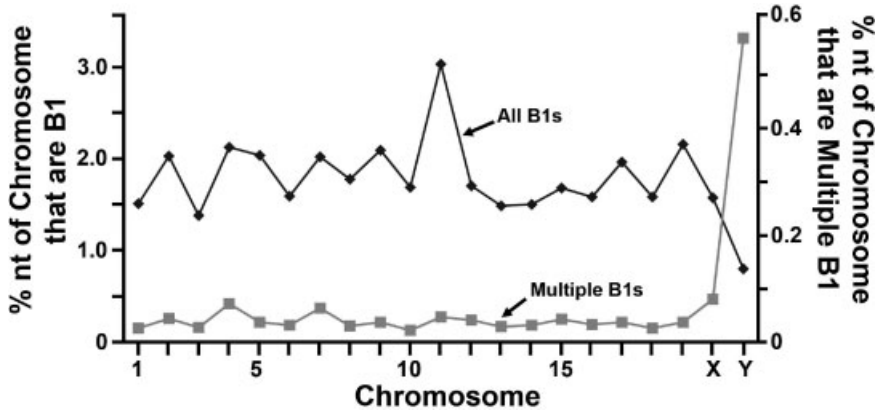

Fig. 4. Distribution of mouse B1s by chromosome. Distribution of all mouse B1s greater than 100 nt and all multi-copy mouse B1s by chromosome. The diamonds are the percentage of all B1s on a particular chromosome (left hand scale). The squares are the percentage of multi-copy B1s on a particular chromosome (right hand scale). While chromosome Y does not contain an exceptional number of B1s, nearly all are multi-copy.

the genes than the user-defined proximities will be flagged. Currently, four methods are provided to access the data: search by repeat sequence, search by gene sequence, search by gene id(s) and browse by copy number.

DISCUSSION

We have demonstrated that most Alu and B1 sequences are unique by using BLAST to look for exact matches of each element throughout the entire genome. The criteria for finding 100% matches at full length is admittedly strict, even at 50% of their lengths, greater than 97% of Alus and 73% of B1s were identified as single-copy sequences. While older Alu sequences were expected to be over-represented among the single copy repeats, we found that the youngest Alus are also predominantly single copy (94%, Table III). Alus with the highest copy number tended to be younger (Table II), however, more than half of all multi-copy Alus corresponded to the oldest and intermediate (J and S) subfamilies (Table II). In this analysis, we used Repbase to classify Alus and B1s [Jurka et al., 2005]. While a more recent approach for classifying Alus has been developed [Price et al., 2004], our use of RepeatMasker was limited to identifying repeat sequences and classifying them into broad age-based categories (young, intermediate, and old). For these limited purposes, we found Repbase to be

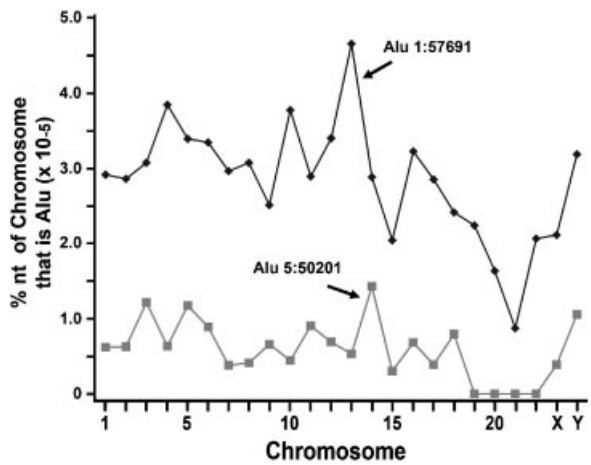

Fig. 5. Distribution of the Alu sequences 1:57691 and 5:50201 copies by chromosome. Distribution of the multi-copy Alu sequences 1:57691 and 5:50201 copies by chromosome. The Y-scale is the percentage of the particular chromosome that is identified by these sequences.

TABLE IV. Distribution of Alu Sequences 1:57,691 and 5:50,201 by Category of Position Relative to Genes

	No. of position type	
Category	Alu 1:57,691	Alu 5:50,201
Close	5 (4)	1
Intronic	126 (125)	23
Exonic	9 (2)	1
Far	193	42

'Close' indicates that the sequence was within 500 nt of at least one gene; Intronic' means that the sequence was within at least one intron; Exonic' means that the sequence was within at least one exon; Distant' means the sequence was more than 500 nt from the closest gene. Numbers in parenthesis are actual numbers of unique addresses. A single unique address occasionally corresponds to multiple genomic features (e.g., one address could be in proximity of more than one gene, or overlap more than one exon or intron).

an adequate classification scheme. In our analysis of Alu 1:57691, the Alu with the highest number of exact copies, all copies were found to be from the same younger family, AluYa5. This is expected based on the current models for Alu and B1 propagation.

These data have implications for our understanding of 7SL RNA-derived SINEs within both human and mouse genomes, and for the expression patterns of genes. Indeed, 7SL RNA SINEs conceivably could be classified as single copy and multi-copy sequences. While single-copy sequences outnumber the multi-copy sequences, both might provide useful tools for the understanding of genome structure and function. The distribution patterns of both human Alus and mouse B1s show an accumulation in the relative number of multi-copy sequences on chromosomes Y of both species (Figs. 3 and 4). More than 71% of all B1s on mouse chromosome Y are multi-copy. For humans, the number is slightly greater than 25%. This is distinctly different from 1% to 5% on other mouse and 0.3% to 4% on other human chromosomes. This enrichment of chromosome Y with multi-copy sequences may be related to previous observations that chromosome Y appears to be enriched for younger AluY sequences [Jurka et al., 2002].

As illustrated in Figure 6, the few Alu repeats that do fulfill our strict criteria for being multi-copy reveal an interesting pattern. Most of the multi-copy repeats were truncated versions of larger Alus, suggesting that some Alus were copied exactly and then portions were mutated at least once. However, some like Alu 5:50,201 were repeated exactly several times. These

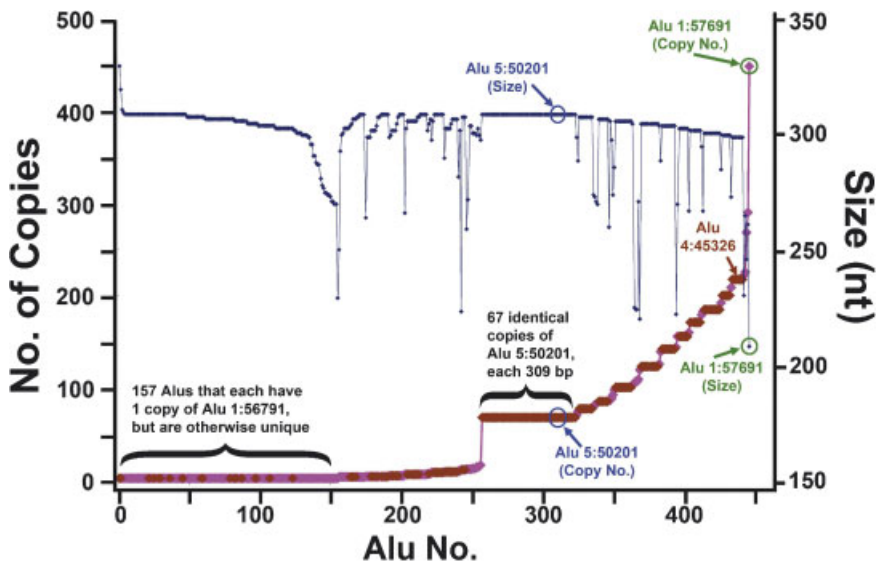

Fig. 6. Genomic Distribution of Alu 1:57691. Exact multi-copy Alus are sometimes truncated versions of slightly larger Alu sequences, but are also repeated as complete exact copies. To represent the complex nature of multi-copy Alus, we plotted each of the 450 exact copies of Alu 1:57691 with respect to size (blue line, right scale) and with respect to how many times each of these copies are, themselves, exactly copied (purple line, left scale). Of the 450 Alu sequences that contain exact replicas of Alu 1:57691, 157 w ere single copy by our strict criteria (far left). Of particular interest are steps, which represent Alus that are exact, complete copies of each other. The largest step contains 67 complete copies of Alu 5:50201 (an arbitrary representative of this smaller family), a 309 nt Alu, each of which also contains an exact replica of the smaller, 208 nt Alu 1:57691. Finally, to further illustrate the relationship between exact repeats, we plotted all the Alus that also contain an exact repeat of Alu 4:45326 as brown points. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

TABLE V. GC Content of Human Alus and Mouse B1s

	Intronic	Exonic	Close	Far
Alu sequence	52%	52%	52%	51%
Surrounding sequence	41%	43%	44%	40%
% of sequences	44%	0.5%	1.0%	54%
B1 sequence	50%	50%	51%	50%
Surrounding sequence	39%	41%	41%	39%
% of sequences	45%	0.6%	1.2%	53%

GC content of human Alus and mouse B1s and their surrounding sequences. The surrounding sequences were defined as one length of the repeat in both directions.

TABLE VII. Distribution of Exact, Full Length Alu Copies With Respect to Subfamilies

Subfamily	Distribution (%)	Distribution of all Alus (%)
J	22	22
S	57	63
Y	21	15
Total	100	100

exact repeats were distributed throughout the genome on 20 chromosomes (Fig. 5). In this case, it appears that the truncated version of the original Alu was transposed 67 times.

Comparisons between percentage single copy of Alus/B1s and corresponding RefSeq fractions indicate that at greater than 50% length, Alus and B1s have a higher percentage of single-copy elements than the RefSeq segments. The relative prevalence of multi-copy sequences among the RefSeq fractions might be explained by exon shuffling [Margulies and McCluskey, 1985] and/or gene duplication. A higher percentage of single-copy mouse RefSeq fractions compared to human RefSeq fractions also might indicate that exon shuffling and/or gene duplication is more prevalent in the human genome. Another possibility is the presence of non-Alu repeats within RefSeq sequences. However, analysis using NCBI-BLAST with filtering on produced results very similar to Figure 2 (unpublished data).

Alus are over expressed in cancerous tissues [Vila et al., 2003; Gibbons and Dugaiczyk, 2005] and in other stressful conditions [Chu et al., 1998; Hagan et al., 2003]. By determining that most Alus are single-copy sequences, it is now possible to identify individual Alus that are over- or under-expressed under certain conditions and trace them back to their genomic locations. An ability to differentiate between the Alus may lead to a better understanding of regulation of both expression and splicing.

Other studies have shown significant differences in methylation of Alus in haploid and developmental cells [Hellmann-Blumberg et al., 1993; Kochanek et al., 1993], as well as significant methylation of mouse B1 sequences [Jeong and Lee, 2005]. Preliminary data from the work at hand suggests significant changes in expression of B1-containing genes in developmental and spermatogenic cells (unpublished data). This change in expression, combined with an ability to explicitly determine changes in methylation of specific B1 sequences, may help answer questions regarding Alu/B1 function. For example, the human Alu sequence 1:57691 is represented on all chromosomes (Fig. 5) and copies of it lie within, or in close proximity to, 140 genes (Table IV). If Alu sequences are preferred methylation targets [Schmid, 1998], the distribution of multi-copy sequences throughout the genome could be providing a common regulatory mechanism for a large number of genes. The apparently random distribution of addresses corresponding to multi-copy sequence 1:57691 also may clarify the retroposition mechanism for these repeats. In this respect, the genome wide distribution of

TABLE VI. Distribution of Exact, Full Length Alu Copies With Respect to Genes, Introns, and Exons

Genomic location	Number of elements	Distribution (%)	Distribution of all Alus (%)
Close	151	1.5	1.0
Distant	6,587	65.0	54.0
Introns	3,343	33.0	44.5
Exons	58	0.5	0.5
Total	10,139	100	100

Close elements are within 500 nucleotide of a gene, distant elements are more than 500 nucleotides from a gene.

TABLE VIII. Copy Number of Exact, Full Length Alu Copies

Copy number	No. of Alus
2	7,222
3	1,590
4	544
5	215
6–10	358
11–50	143
67	67

1:57,691 indicate that the target address for retroposition is independent of the original source.

Another interesting question is raised by the presence of 10,139 Alu elements that have at least one exact copy. These elements are present in exons, introns, around genes, and in more remote areas of the genome (Table VI). They are also well distributed with respect to subfamilies (Table VII). In fact, it is interesting that both the old (J) and intermediate (S) are well represented among the perfectly matched multi-copy Alus. Based on our results, only 112 of these Alus can be accounted for by segmental duplications. It is therefore possible that approximately 2,200 J and 5,800 S Alus are either relatively recent or they have been protected from mutation for tens of millions of years.

These data present a novel, though not unexpected view of Alu and B1 repeats in the human and mouse genomes. While it was not unexpected that each Alu and B1 element would have a unique portion, by comparing each sequence to the entire genome using BLAST we quantified the extent of this identity. We have also created a website to allow access to this information. Some uses of this website are: (i) search by repeat sequence. If the researcher has part of a repeat sequence, he/she may search the database using this sequence and retrieve a list of repeats that match the sequence. An example would be [Hellmann-Blumberg et al., 1993] where a primer was used to identify a subset of Alus. Using AluPlus the researchers may obtain the list of Alus hybridized to the primer and information about these repeats can be accessed; (ii) search by gene id(s). Given a list of genes (e.g., from a microarray experiment), the user is able to obtain a list of repeats that lie within or just outside these genes. The distance that a sequence may lie outside the gene is controlled by proximity parameter. Several studies have shown that Alus and Alu-containing genes are over expressed in oncogenesis [Vila et al., 2003; Gibbons and Dugaiczyk, 2005]; our studies have indicated that coding B1s also are over expressed in embryonic and spermatogenic cells (unpublished data).

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the University of Hawaii Dell Cluster system under the management of the Department of Information and Computer Sciences with funding from NIH-NCCR P20RR016467 and NSF-EPS02-37065.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215: 403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Blinov VM, Denisov SI, Saraev DV, Shvetsov DV, Uvarov DL, Oparina N, Sandakhchiev LS, Kiselev LL. 2001. Structural organization of the human genome: Distribution of nucleotides, Alu-repeats and exons in chromosomes 21 and 22). Mol Biol Mosk 35:1032–1038.

Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, Tsui LC, Scherer SW. 2003. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol 4:R25.

Chu WM, Ballard R, Carpick BW, Williams BR, Schmid CW. 1998. Potential Alu function: Regulation of the activity of double-stranded RNA-activated kinase PKR. Mol Cell Biol 18:58–68.

Dagan T, Sorek R, Sharon E, Ast G, Graur D. 2004. AluGene: A database of Alu elements incorporated within protein-coding genes. Nucleic Acids Res 32: D489–D492.

Deininger PL, Jolly DJ, Rubin CM, Friedmann T, Schmid CW. 1981. Base sequence studies of 300 nucleotide renatured repeated human DNA clones. J Mol Biol 151: 17–33.

Gibbons R, Dugaiczyk A. 2005. Phylogenetic roots of Alu-mediated rearrangements leading to cancer. Genome 48:160–167.

Gish W. 1996–2004. WU-BLAST.

Grover D, Majumder PP, B-Rao C, Brahmachari SK, Mukerji M. 2003. Nonrandom distribution of alu elements in genes of various functional categories: Insight from analysis of human chromosomes 21 and 22. Mol Biol Evol 20:1420–1424.

Grover D, Mukerji M, Bhatnagar P, Kannan K, Brahmachari SK. 2004. Alu repeat analysis in the complete human genome: Trends and variations with respect to genomic composition. Bioinformatics 20:813–817.

Hagan CR, Sheffield RF, Rudin CM. 2003. Human Alu element retrotransposition induced by genotoxic stress. Nat Genet 35:219–220.

Hellmann-Blumberg U, Hintz MF, Gatewood JM, Schmid CW. 1993. Developmental differences in methylation of human Alu repeats. Mol Cell Biol 13:4523–4530.

Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J,

Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Birney E. 2005. Ensembl 2005. Nucleic Acids Res 33:D447–D453.

Jeong KS, Lee S. 2005. Estimating the total mouse DNA methylation according to the B1 repetitive elements. Biochem Biophys Res Commun 335:1211–1216.

Jurka J, Milosavljevic A. 1991. Reconstruction and analysis of human Alu genes. J Mol Evol 32:105–121.

Jurka J, Krnjajic M, Kapitonov VV, Stenger JE, Kokhanyy O. 2002. Active Alu elements are passed primarily through paternal germlines. Theor Popul Biol 61:519–530.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467.

Kapitonov V, Jurka J. 1996. The age of Alu subfamilies. J Mol Evol 42:59–65.

Kochanek S, Renz D, Doerfler W. 1993. DNA methylation in the Alu sequences of diploid and haploid primary human cells. EMBO J 12:1141–1151.

Krayev AS, Kramerov DA, Skryabin KG, Ryskov AP, Bayev AA, Georgiev GP. 1980. The nucleotide sequence of the ubiquitous repetitive DNA sequence B1 complementary to the most abundant class of mouse fold-back RNA. Nucleic Acids Res 8:1201–1215.

Labuda D, Sinnett D, Richer C, Deragon JM, Striker G. 1991. Evolution of mouse B1 repeats: 7SL RNA folding pattern conserved. J Mol Evol 32:405–414.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Liu WM, Chu WM, Choudary PV, Schmid CW. 1995. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. Nucleic Acids Res 23:1758–1765.

Margulies DH, McCluskey J. 1985. Exon shuffling: New genes from old. Surv Immunol Res 4:146–159.

Price AL, Eskin E, Pevzner PA. 2004. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Res 14:2245–2252.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence. RefSeq. a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:D501–D504.

Quentin Y. 1992. Origin of the Alu family: A family of Alu-like monomers gave birth to the left and the right arms of the Alu elements. Nucleic Acids Res 20:3397–3401.

Quentin Y. 1994. A master sequence related to a free left Alu monomer (FLAM) at the origin of the B1 family in rodent genomes. Nucleic Acids Res 22:2222–2227.

Rowold DJ, Herrera RJ. 2000. Alu elements and the human genome. Genetica 108:57–72.

Schmid CW. 1998. Does SINE evolution preclude Alu function? Nucleic Acids Res 26:4541–4550.

Smit A, Hubley R, Green P. 1996–2004. RepeatMasker Open-3.0 http://www.repeatmasker.org.

Ullu E, Tschudi C. 1984. Alu sequences are processed 7SL RNA genes. Nature 312:171–172.

Vila MR, Gelpi C, Nicolas A, Morote J, Schwartz S, Jr., Schwartz S, Meseguer A. 2003. Higher processing rates of Alu-containing sequences in kidney tumors and cell lines with overexpressed Alu-mRNAs. Oncol Rep 10: 1903–1909.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.