Widespread establishment and regulatory impact of Alu exons in human genes

Shihao Shenb,1, Lan Linb,1, James J. Caf, Peng Jiangb, Elizabeth J. Kenkelb, Mallory R. Stroikb, Seiko Satob, Beverly L. Davidsonb,d,e, and Yi Xingb,f,2

Departments of 1Biostatistics, 2Internal Medicine, 3Molecular Physiology and Biophysics, 4Neurology, and 5Biomedical Engineering, University of Iowa, Iowa City, IA 52242; and 6Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77845

Edited* by Wing Hung Wong, Stanford University, Stanford, CA, and approved January 12, 2011 (received for review August 28, 2010)

The Alu element has been a major source of new exons during primate evolution. Thousands of human genes contain spliced exons derived from Alu elements. However, identifying Alu exons that have acquired genuine biological functions remains a major challenge. We investigated the creation and establishment of Alu exons in human genes, using transcriptome profiles of human tissues generated by high-throughput RNA sequencing (RNA-Seq) combined with extensive RT-PCR analysis. More than 25% of Alu exons analyzed by RNA-Seq have estimated transcript inclusion levels of at least 50% in the human cerebellum, indicating widespread establishment of Alu exons in human genes. Genes encoding zinc finger transcription factors have significantly higher levels of Alu exonization. Importantly, Alu exons with high splicing activities are strongly enriched in the 5′-UTR, and two-thirds (10/15) of 5′-UTR Alu exons tested by luciferase reporter assays significantly alter mRNA translational efficiency. Mutations in the exon junction sequences corresponding to its upstream, downstream, and skipping junctions (Fig. 1f) modulate translational efficiency, such as the creation or elongation of upstream ORFs that repress the translation of the primary ORFs. This study presents genomic evidence that a major functional consequence of Alu exonization is the lineage-specific evolution of translational regulation. Moreover, the preferential creation and establishment of Alu exons in zinc finger genes suggest that Alu exonization may have globally affected the evolution of primate and human transcriptomes by regulating the protein production of master transcriptional regulators in specific lineages.

Results
RNA-Seq Analysis of Alu Exons in the Human Brain Transcriptome. To investigate the splicing activities of Alu exons, we analyzed a deep RNA-Seq data set of the human cerebellum totaling 123 million single-end reads (13). Using the University of California, Santa Cruz (UCSC) Known Genes annotation (14), we extracted 627 Alu-derived internal exons whose flanking exons were constitutively spliced. For each Alu exon, we collected a set of three exon–exon junction sequences corresponding to its upstream, downstream, and skipping junctions (Fig. 1f). We mapped the RNA-Seq reads to these exon–exon junctions (Materials and Methods). Because the flanking exons of most Alu exons were non-repeat-derived, we were able to map exon–exon junction reads unambiguously. In total, 287 Alu-derived exons had at least one read mapped to one of the three junctions, including 127 exons with at least five reads and 82 exons with at least 10 reads mapped to one of the three junctions. The majority of the 287 exons (197, 69%) had at least one read mapped to the upstream or downstream junction, indicating exon inclusion in the transcripts. For each of the 287 exons, we calculated its transcript inclusion level in the cerebellum (i.e., the percentage of transcripts including the exon among all transcripts including or excluding the exon) using its junction read counts (see the formula in Materials and Methods). For example, the Alu exon in zinc finger protein 445 (ZNF445) had an estimated inclusion level of 79% in the human cerebellum (Fig. 1f).

To confirm the RNA-Seq estimates of Alu exon inclusion levels, we randomly selected 46 exons with at least one junction read indicating exon inclusion for RT-PCR analysis. Of these 46 exons, 34 exons (74%) originated completely from Alu elements; the remaining 12 exons (26%) were derived from mergers of Alu and non-Alu sequences. We examined their splicing patterns in the human cerebellum using semiquantitative RT-PCR (see the list of Table S1 and their gel pictures in Fig. S1). For ex-
ample, the Alu exon in ZNF445 was confirmed to be included in the majority of its transcript products (Fig. 1B). Of all 46 exons tested, 45 (98%) showed exon inclusion in the cerebellum (Fig. S1A). From these 45 exons, we further selected 17 exons with weak exon inclusion PCR bands and confirmed their splicing into transcripts by additional RT-PCR experiments with one primer within the Alu exon and the other primer in a flanking constitutive exon (Fig. S1B). We classified the 46 exons into five categories ranging from no exon inclusion to constitutive splicing (i.e., 100% exon inclusion), based on the RT-PCR results. We observed a significant positive correlation between the exon inclusion levels estimated by RNA-Seq and by RT-PCR (Spearman’s $r = 0.60; P = 1.1e-5$; Fig. 1C). This correlation was stronger among Alu exons with at least five reads mapped to one of the three exon–exon junctions ($r = 0.64$) (underlined gene symbols in Fig. 1C). This result was expected, because the increased RNA-Seq read coverage allows more reliable estimates of exon inclusion levels (15). For the remainder of this paper, all analyses involving exon inclusion levels are restricted to Alu exons with at least five reads mapped to one of the three junctions. Of the 127 Alu exons meeting this criterion, 36 (28%) had an estimated exon inclusion level of at least 50% (referred to as “highly included” Alu exons). These results indicate that a significant portion of Alu exons have acquired strong splicing signals to be spliced into the majority of their genes’ transcripts in the cerebellum.

Ubiquitous or Tissue-Specific Splicing Patterns of Alu Exons. To assess whether the cerebellum-spliced Alu exons are ubiquitously spliced in a broad range of tissues or specifically spliced in the cerebellum, we analyzed a second RNA-Seq data set of the human liver with a total of 90 million reads (13, 16). We found 85 Alu exons with at least five reads mapped to one of the three exon–exon junctions in both data sets. For these 85 exons, the estimated transcript inclusion levels in the cerebellum and liver were strongly correlated (Pearson’s $r = 0.61; P = 4.8e-10$), indicating that most cerebellum-spliced Alu exons also were spliced in the liver. To examine directly the variations of Alu exon splicing among human tissues, for the 46 exons tested by RT-PCR in the cerebellum, we expanded our RT-PCR analysis to nine additional tissues (Materials and Methods). Of the 45 Alu exons spliced in the cerebellum, five displayed notable changes in splicing patterns in various human tissues; the rest had similar splicing patterns in all tissues analyzed (Fig. S1 and Table S1). For example, the Alu exon in ZNF445 had high inclusion levels in all 10 tissues (Fig. 1B).

Alu Exonization Events Are Strongly Enriched in Zinc Finger Transcription Factors. The expanded list of Alu exons and the ability to estimate transcript inclusion levels by RNA-Seq allowed us to identify gene families and functional categories enriched for the creation and establishment of Alu exons. We compared 35 genes with highly included Alu exons with a background list of 16,530 cerebellum-expressed genes (Materials and Methods). Using the functional annotation tool Database for Annotation, Visualization and Integrated Discovery (DAVID) (17), we identified two significantly enriched Gene Ontology (GO) terms with a Benjamin-corrected false discovery rate (FDR) of <0.05: “KRAB box transcription factor” ($P = 1.8e-4; \text{FDR} = 3.6e-3$) and “zinc finger transcription factor” ($P = 1.4e-3; \text{FDR} = 1.4e-2$). Both GO terms refer to zinc finger (ZNF) transcription factors, a large family of transcription factors in the human genome. These transcription factors typically are characterized by an N-terminal protein interaction domain, most commonly the Kruppel-associated box (KRAB) domain and the C2H2 ZNF DNA-binding domain in the C-terminal region (18). Interestingly, ZNF genes underwent rapid expansion and adaptive evolution during primate and human evolution (18, 19). Therefore, they have been considered key contributors to lineage-specific transcriptome regulation in primates and humans (18, 20).

To confirm the splicing of Alu exons in ZNF genes, we conducted RT-PCR analysis of 12 exons that had an estimated inclusion level of >50% or higher (Table S2). All 12 exons were validated as being spliced into the transcripts, including nine exons with at least medium inclusion level in the cerebellum, according to RT-PCR. To avoid misinterpretation of RT-PCR results because of nonspecific amplification of paralogous ZNF genes, the identities of all PCR products were confirmed by sequencing. Although the specific biological functions of most ZNF genes remain obscure (18), some of the Alu-exon-containing ZNF genes have been implicated in disease or gene regulation. For example, ZNF445 has been shown to activate the transcriptional activity of activator protein 1 (AP1) (21). Zinc finger protein 706 (ZNF706) has a sex-specific gene expression pattern (22) and is up-regulated in larynx tumors (23). Zinc finger protein 81 (ZNF81) is one of the three X-chromosome ZNF genes associated with nonsyndromic X-linked mental retardation (24). Histone H4 transcription factor (HINFP, also known as zinc finger protein 743, ZNF743) encodes a ZNF protein that interacts with methyl-CpG-binding protein 2 (MBD2) and is involved in transcriptional and epigenetic regulation (25).

We performed additional analyses to investigate the enrichment of Alu exons in ZNF genes. For independent confirmation of the result of the DAVID analysis based on GO annotations, we collected a list of 551 ZNF genes in the human genome from Huntley and colleagues (26) (Materials and Methods). We found that 25% (8/35) of genes with a highly included Alu exon were ZNF genes, compared with only 3% among other cerebellum-expressed genes, a more than sevenfold enrichment ($P = 7.1e-6$; Fig. S2).
one-sided Fisher exact test; Fig. 2A). Of all nine Alu exons in ZNF genes with at least five reads mapped to at least one exon–exon junction, the average estimated transcript inclusion level was 72%, compared with 28% for Alu-derived exons in non-ZNF genes ($P = 0.004$; one-sided Wilcoxon test; Fig. 2B). This observation was not an artifact caused by higher expression level or increased RNA-Seq coverage of ZNF genes in the cerebellum. In fact, according to RNA-Seq gene expression estimates based on the reads per kilobase of exon model per million mapped reads (RPKM) metric (27) (Materials and Methods), among Alu-exon–containing genes, ZNF genes had lower expression levels on average than non-ZNF genes in the cerebellum ($P = 0.007$; one-sided Wilcoxon test) (Fig. 2C).

**Frequent Alu Exonization Is Characteristic of Primate-Specific Genes.** Because many ZNF genes are primate-specific, we asked whether the preferential creation and establishment of Alu exons is a general characteristic of phylogenetically young genes. Based on the PhyloPat (28) phylogenetic classification of human genes (Materials and Methods), we grouped all human genes into four mutually exclusive groups with increasing phylogenetic ages: primate, mammalian, euteleostomi, and metazoan. Among genes expressed in the cerebellum, 9.3% (65) of primate genes contained internal Alu exons, compared with 4.4% (121) of mammalian genes, 4.4% (267) of euteleostomi genes, and 4.9% (346) of metazoan genes ($P < 1e-5$ for all one-sided Fisher exact tests between primate genes and any other age group) (Fig. S2A). We also observed a higher percentage of genes containing highly included Alu exons in the cerebellum among the primate genes (0.95%, 0.10%, 0.08%, and 0.17% in the four age groups, respectively; $P < 1e-5$ for all one-sided Fisher exact tests) (Fig. S2A). Moreover, among all Alu exons with at least five reads mapped to at least one of the three exon–exon junctions in the cerebellum, 75% (six of eight) in the primate gene group had at least a 50% inclusion level, a percentage higher than that of any other age group (Fig. S2B). Overall, we observed a significant anti-correlation between the phylogenetic ages of Alu-exon–containing genes and the inclusion levels of Alu exons in the cerebellum ($P = 0.003$, linear regression of gene age and Alu exon inclusion level). We observed the same trend in the liver RNA-Seq data set (Fig. S2 C and D).

Consistent with this trend, the creation and establishment of Alu exons were enriched in primate-specific ZNF genes as compared with ancient ZNF genes (Fig. S3). Given that a large fraction of ZNF genes are primate-specific, this result suggests that the enrichments of Alu exons in ZNF genes and in primate-specific genes are coupled to a certain extent. Nonetheless, the preference in ZNF genes over non-ZNF genes holds even after controlling for phylogenetic age (Fig. S3). Indeed, we found highly included Alu exons in both evolutionarily ancient and primate-specific ZNF genes. For example, ZNF445 is conserved among most mammalian species and has a 5′-UTR Alu exon that regulates its translational efficiency (Fig. 3D). Despite the ancient origin of this gene, the strong splicing activity of this Alu exon appears to have been acquired recently, because the comparison of the human and nonhuman primate orthologous genomic regions and the consensus sequence of the corresponding AluSc subfamily revealed a human/chimpanzee-specific C-to-T substitution that created the “GT” 5′ splice site. Interestingly, the genomic region of ZNF445 appeared under positive selection during recent human evolution according to SNP-based scans of positive selection signals (29). Collectively, our results indicate that Alu exonization has played a role in both the ongoing evolution of ancient ZNF genes and the recent expansion of the gene family in the primate and human lineages.

**Preferential Establishment and Regulatory Impact of Alu Exons in the 5′-UTR.** We found a strong enrichment of Alu exons in the 5′-UTR. We compared the occurrence and transcript inclusion levels of Alu exons in non-ZNF genes and any other age group) ($P = 0.003$, linear regression of gene age and Alu exon inclusion level). We observed the same trend in the liver RNA-Seq data set (Fig. S2 C and D).

**Consistent with this trend, the creation and establishment of Alu exons were enriched in primate-specific ZNF genes as compared with ancient ZNF genes (Fig. S3).** Given that a large fraction of ZNF genes are primate-specific, this result suggests that the enrichments of Alu exons in ZNF genes and in primate-specific genes are coupled to a certain extent. Nonetheless, the preference in ZNF genes over non-ZNF genes holds even after controlling for phylogenetic age (Fig. S3). Indeed, we found highly included Alu exons in both evolutionarily ancient and primate-specific ZNF genes. For example, ZNF445 is conserved among most mammalian species and has a 5′-UTR Alu exon that regulates its translational efficiency (Fig. 3D). Despite the ancient origin of this gene, the strong splicing activity of this Alu exon appears to have been acquired recently, because the comparison of the human and nonhuman primate orthologous genomic regions and the consensus sequence of the corresponding AluSc subfamily revealed a human/chimpanzee-specific C-to-T substitution that created the “GT” 5′ splice site. Interestingly, the genomic region of ZNF445 appeared under positive selection during recent human evolution according to SNP-based scans of positive selection signals (29). Collectively, our results indicate that Alu exonization has played a role in both the ongoing evolution of ancient ZNF genes and the recent expansion of the gene family in the primate and human lineages.

---

**Fig. 2.** Enrichment of Alu exons in ZNF genes. (A) The percentage of ZNF genes is significantly higher in genes containing highly included Alu exons than in other cerebellum-expressed genes without highly included Alu exons. (B) Alu exons in ZNF genes have higher overall transcript inclusion levels than Alu exons in non-ZNF genes. (C) ZNF genes with Alu exons do not have higher overall expression levels in the cerebellum compared with non-ZNF genes with Alu exons. All exons in panels A–C have at least five reads mapped to one of the three exon–exon junctions. Error bars indicate the 95% confidence interval.
exons within different regions of protein-coding genes [5′-UTR, coding sequence (CDS), 3′-UTR]. We restricted this analysis to exons for which transcript locations could be determined unambiguously using the UCSC Known Genes annotation (14). Among all internal Alu exons collected from the UCSC Known Genes database, 290 (54%) are in the 5′-UTR, 248 (46%) are in the CDS, and two are in the 3′-UTR (Fig. 3A). Of the 99 Alu exons with at least five reads mapped to one of the three exon–exon junctions (i.e., two inclusion junctions and one skipping junction, thus already slightly biased toward cerebellum-included Alu exons), 64 (65%) are in the 5′-UTR, 34 are in the CDS, and only one is in the 3′-UTR (Fig. 3A). Moreover, of the 17 highly included Alu exons according to RNA-Seq data, an even higher percentage of exons are located in the 5′-UTR (82%) (Fig. 3A). The percentages of 5′-UTR exons in these lists of Alu exons are substantially higher than in the general population of alternatively spliced exons in human genes. For example, among the high-confidence alternatively spliced cassette exons from the Alternative Splicing Annotation Project 2 (ASAP2) database (30), 673 (19.9%) are in the 5′-UTR, 2,700 (80.0%) are in the CDS, and two (0.1%) are in the 3′-UTR. Similarly, in the Human Transcriptome Database for Alternative Splicing (H-DBAS) compiled from full-length cDNAs (31), 17.5% of alternatively spliced exons are in the 5′-UTR. In the UCSC Known Genes database (14), 14.3% of alternatively spliced internal cassette exons are in the 5′-UTR (Fig. 3A). In summary, from alternatively spliced cassette exons to Alu exons to highly included Alu exons, we observed a consistent increase in the percentage of 5′-UTR exons and decrease in the percentage of CDS exons. Moreover, Alu exons in the 5′-UTR had higher average transcript inclusion levels than exons in the CDS (30% vs. 11%, P = 0.05, one-sided Wilcoxon test; Fig. 3B). Together, these results suggest that the 5′-UTR not only is a hotspot for the initial creation of new Alu exons but also favors the subsequent establishment of strong splicing activities. It should be noted that this analysis was restricted to internal spliced Alu exons in the UTRs, although Alu elements also could contribute to sequences within terminal exons, including polyadenylation sites (32).

The high level of Alu exon establishment in the 5′-UTR raised an interesting question about the regulatory impact of these exons. It is well known that the 5′-UTR contains regulatory signals of mRNA stability and protein translation (33). In several genes the alternative splicing of 5′-UTR exons regulates protein translation (34–36). To determine if Alu exonization in the 5′-UTR could affect the translation of host mRNAs broadly, we selected 15 Alu exons and compared the translational efficiency of the exon inclusion 5′-UTR versus the exon skipping 5′-UTR using a dual luciferase reporter construct pIRES-Luc2 (Materials and Methods). For each 5′-UTR isoform, the resulting reporter construct expressed both the firefly luciferase and the Renilla luciferase bicistronically. The Renilla luciferase was fused downstream of mRNA stability and protein translation (33). In several genes the alternative splicing of 5′-UTR exons. It is well known that the 5′-UTR has a dual regulatory role in transcription factors [34], which was calculated as the ratio between the Renilla luciferase and the luciferase bicistronically. The Renilla luciferase was fused downstream of the cloned 5′-UTR isoform, whereas the firefly luciferase translation was driven independently by the internal ribosomal entry site (IRES) and was not regulated by the cloned 5′-UTR (Fig. 3C). The translational efficiency of each 5′-UTR construct was calculated as the ratio between the Renilla luciferase and the firefly luciferase activities. In four genes encoding ZNF transcription factors [ZNF445, zinc finger protein 808 (ZNF808), ZNF81, HINFP/ZNF743], the Alu exon significantly altered the translational efficiency in HeLa cells when included in the 5′-UTR (Fig. 3D). We also observed a significant effect of the Alu exon on the translational efficiency of the non-ZNF genes aldo-keto reductase family C, member A1 (AKR1A1), PDZ domain containing 1 (PDZK1), leucine-rich repeat containing 56 (LRRRC56), nitric oxide synthase interacting protein (NOSIP), regulator of chromosome condensation and BTB domain containing protein 1 (RCBTB1), and ribonuclease P protein subunit p38 (RPP38) (Fig. 3D). In total, 10 of the 15 tested Alu exons altered translational efficiency. We tested nine exons in HEK293 cells and obtained similar results (Fig. S4).

Molecular Mechanisms of Translational Regulation by Alu Exons. Newly created 5′-UTR Alu exons potentially could regulate translational efficiency through a variety of molecular mechanisms, such as upstream ORF (uORF), IRES, secondary structure, and 5′-UTR length (37). We decided to test hypotheses reflecting two prevalent mechanisms of translational regulation by 5′-UTR elements: one would decrease and the other would increase translational efficiency. In the first hypothesis, Alu exons may introduce or elongate uORFs before the primary start codons, thus representing the translation of the primary ORFs (38, 39). In the second hypothesis, Alu exons could introduce cellular IRES elements to the 5′-UTRs and thus increase translational efficiency (37).

To test the first hypothesis involving uORFs, all 15 Alu exons tested by luciferase assays (Fig. 3D) were scanned for differences in uORFs between the Alu exon inclusion and exon skipping 5′-UTRs. We identified three genes in which the Alu exons either created a single new uORF (NOSIP and RPP38) or elongated an existing uORF (ZNFS1). In NOSIP, the Alu exon introduced a new uORF to the 5′-UTR (Fig. 4A). We used site-directed mutagenesis to disrupt the uORF by mutating the upstream AUG (uAUG) start codon within the Alu exon to UUG (a651), AAG (166a), or UAC (g67c). All three mutant constructs completely reversed the translational repression by the Alu exon (Fig. 4A). Similar results were obtained for the RPP38 Alu exon (Fig. S5). In ZNFS1, the Alu exon caused a frame shift of an existing uORF (51 nt), resulting in a fourfold longer uORF (267 nt) (Fig. 4B). The Alu exon inclusion 5′-UTR caused a >60% decrease in translational efficiency compared with the exon skipping 5′-UTR (Fig. 4B). We introduced three different stop codons, UAA, UAG, or UGA, within the Alu exon to shorten the uORF back to its original length (51 nt). All three mutations resulted in a full recovery of the translational efficiency (Fig. 4B). Together, these mutational studies reveal a general mechanism of translational repression by Alu exons through the creation or elongation of uORFs in the 5′-UTR.

To test the second hypothesis involving potential creation or disruption of IRES by Alu exons, a bicistronic reporter system was adapted using pRF-Luc2 as the empty vector backbone to generate a series of IRES-activity reporter constructs (Materials and Methods and SI Materials and Methods). Test sequences were

![Fig. 4. Alu exons repress translation by creating or elongating uORFs.](Image)
cloned and inserted between the upstream Renilla luciferase stop codon and the downstream firefly luciferase start codon. IRES activity of the test sequence was indicated by the ratio between firefly luciferase and Renilla luciferase activities. We selected three genes [ZNF808, HINFP, and protein arginine methyltransferase 7 (PRMT7)] to test in this system. These genes represented increased (ZNF808), decreased (HINFP), or constant (PRMT7) translatable efficiency after Alu exon inclusion (Fig. 3D). Among all tested 5′-UTR isoforms, most constructs showed a very low firefly/Renilla luciferase ratio that was not above the empty vector (pRF-Luc2) control (Fig. S6). Only the Alu exon inclusion 5′-UTR construct of ZNF808 showed a significant firefly/Renilla ratio comparable to the Encephalomyocarditis virus (EMCV)-IRES-1 positive control and much higher (approximately fourfold) than its Alu exon skipping counterpart (Fig. 5 and Fig. S6). This result is consistent with the observation that the Alu exon inclusion 5′-UTR of ZNF808 showed a significant increase in translatable efficiency (Fig. 3D), suggesting a potential IRES created by Alu exon inclusion. However, to confirm the identity of the IRES element definitively and to define its exact location in the Alu exon inclusion 5′-UTR requires substantial additional experimental efforts (39, 40).

Together, these experiments reveal potential molecular mechanisms (uORF, IRES) by which Alu exons modulate translational location in the Alu exon inclusion 5′-UTR (Fig. 3D). Among all tested 5′-UTR isoforms, most constructs showed a very low firefly/Renilla luciferase ratio that was not above the empty vector (pRF-Luc2) control (Fig. S6). Only the Alu exon inclusion 5′-UTR construct of ZNF808 showed a significant increase in translatable efficiency (Fig. 3D), suggesting a potential IRES created by Alu exon inclusion. However, to confirm the identity of the IRES element definitively and to define its exact location in the Alu exon inclusion 5′-UTR requires substantial additional experimental efforts (39, 40).

Discussion

The origin and evolution of new exons have attracted considerable interest in recent years (6). Although lineage-specific exons are common in mammalian genomes, identifying those exons that have genuine biological functions is a major challenge. Although a growing list of studies report new exons with strong splicing activities and functional roles, a global search for such new exons remains difficult. Most genomic technologies, such as EST sequencing, which is highly sensitive to mRNAs that are short (7) or low-abundance mRNAs, may not be able to detect these new exons. In this study, we used RNA-Seq to systematically identify Alu exons with high splicing activities, in the 5′-UTR exons as the most conserved among primates and other mammals, we demonstrate experimentally that the primate-specific Alu exons in their 5′-UTRs significantly alter the translatable efficiency of the host mRNAs. These results suggest an interesting evolutionary scenario: that the creation and establishment of Alu exons could impact the primate and human transcriptomes globally by fine-tuning the protein production of master transcriptional regulators in a lineage-specific manner.

Our study has broad implications for understanding gene regulation and transcriptome evolution in primates and humans. Importantly, we observed a significant enrichment of Alu exons, especially those with high splicing activities, in the 5′-UTR. Moreover, of the 15 Alu exons tested by luciferase reporter assays, 10 exons altered mRNA translational efficiency when included in the 5′-UTR. These results provide genomic evidence that a major functional consequence of Alu exonization is the lineage-specific evolution of the 5′-UTR and translational regulation. Consistent with our data pointing to the importance of the 5′-UTR in primate evolution, a recent study revealed 5′-UTR exons as the hotspot of human-specific acceleration of nucleotide substitutions (42). Together, these findings suggest that the evolution of the 5′-UTR is an essential aspect of human genome evolution and may contribute to the acquisition of species-specific traits. Our findings on the widespread translational impact of Alu exons also provides insight into the function of 5′-UTR alternative splicing events in general. Our data imply that alternative splicing of 5′-UTR exons could be a prevalent mechanism of gene regulation in humans that affects phenotypes or modulates disease pathogenesis.

Materials and Methods

RNA-Seq and RT-PCR Analysis of Alu Exons. The locations of Alu elements in the human genome were downloaded from the UCSC Genome Browser database (43). The locations of internal spliced exons in human genes were taken from the UCSC Known Genes database (14). To eliminate long exonic regions possibly resulting from intron-retention events, we removed exons longer than 250 bp as in ref. 9. We defined an exon as Alu-derived if the Alu element covered at least 25 bp of the exon and more than 50% of the total exon length. To avoid complications in RNA-Seq analysis arising from complex alternative splicing patterns of flanking exons, we focused on “simple” Alu exons with constitutive flanking exons.

We downloaded Illumina RNA-Seq data of human cerebellum and liver from published datasets. The cerebellum dataset consists of 123 million reads for six human cerebellum samples (13). The liver dataset consists of 90 million reads derived from two studies (13, 16). We mapped RNA-Seq reads to the human genome (hg18) and all exon–exon junctions supported by the UCSC Known Genes annotations (14), using the software ELAND allowing up to 2 bp mismatches. Each mapped exon–exon junction sequence required at
least 5 bp from any side of the exon junction. We removed reads that mapped to either the human genome (hg18) or multiple junctions. For each Alu exon, its transcript inclusion level in a given sample was calculated using the number of reads that uniquely mapped to its upstream junction (UJ), downstream junction (DJ), and skipping junction (SJ) as [(UJ+DJ)(UJ+DJ+2*SJ)] as in ref. 13. The overall expression levels of human genes in the cerebellum were calculated using the RPKM metric (27) within the constitutive exons. Cerebellum-expressed genes were defined as genes with at least one unique cerebellum RNA-Seq read mapped to their constitutive exons.

Total RNA samples of 10 human tissues were purchased from Clontech. For each tested Alu exon, we designed a pair of forward and reverse PCR primers at flanking conserved Alu sequences. The RT-PCR PAGE gel images were analyzed by densitometry using the ImageQuant TL software (GE). Final Alu exon inclusion levels were grouped into five categories: no exon inclusion (0%), minor (1–30%), medium (30–70%), major (70–99%), and constitutive (100%). To confirm further the exon inclusion events of Alu exons with weak exon inclusion PCR bands, we also designed a pair of PCR primers with one primer located within the Alu exon and the other primer located in a flanking constitutive exon. All RT-PCR primer sequences are described in Tables S3 and S4.

**Collection of ZNF Genes in the Human Genome.** We collected a list of 551 UCSC Known Genes loci encoding ZNF transcription factors, using a catalog of ZNF genes compiled by Huntley et al. (26). Specifically, we intersected the ZNF genes in Huntley et al. with the UCSC Known Genes annotations, with the requirement that at least 70% of the genomic region of a ZNF gene defined by Huntley et al. be covered by a UCSC Known Genes transcript.

**Phylogenetic Age Analysis.** We used the PhyloPat database (28) to determine the phylogenetic ages of human genes. PhyloPat classifies the phylogenetic lineages of human genes using Ensembl orthology annotations (28). We grouped all human genes into four mutually exclusive groups based on the PhyloPat classification: primate, mammalian, euteleostomi, and metazoan. For example, primate genes refer to human genes present in primate species but absent from other nonprimate mammalian species.

**Dual Luciferase Reporter Vector Construction and Dual Luciferase Reporter Assay.** The psiCHECK-2 vector (Promega) was modified to construct the dual luciferase reporter pIREs-Luc2 and pRF-Luc2 vector backbones. The pIREs-Luc2 reporter was used to assess the translational efficiency of Alu exon inclusion or skipping 5′-UTRs. The pRF-Luc2 reporter was used to test the potential IRES activity of the 5′-UTR isoforms. Details of the reporter vector construction and dual-luciferase reporter assay are supplied in SI Materials and Methods.

**Acknowledgments.** We thank Dr. Russ Carstens for the pIREs-Blast-delta-40 vector, and Jennifer Dozier for technical assistance. This work was supported by National Institutes of Health Grants R01HG004634 and R01GM088342 (to Y.K.) and P01NS050210 (to B.L.D.), and by the Roy J. Carver Trust (to B.L.D.).

References

5. Makalowski W, Mitchell GA, Labuda D (1994) Alu sequences in the coding regions of human genes compiled by Huntley et al. (26). Specifically, we intersected the ZNF genes in Huntley et al. with the UCSC Known Genes annotations, with the requirement that at least 70% of the genomic region of a ZNF gene defined by Huntley et al. be covered by a UCSC Known Genes transcript.
6. Gerber A, Oshlack A, Buckingham J, Hubbard T, Birney E (2007) RNA-Seq read mapped to their constitutive exons. Cerebellum-expressed genes were defined as genes with at least one unique cerebellum RNA-Seq read mapped to their constitutive exons.