Species-Specific Exon Loss in Human Transcriptomes

Jinkai Wang,†,1 Zhi-xiang Lu,†,1 Collin J. Tokheim,1 Sara E. Miller,2 and Yi Xing*1
1Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles
2Department of Internal Medicine, University of Iowa
†These authors contributed equally to this work.
*Corresponding author: E-mail: yxing@ucla.edu.
Associate editor: Joshua Akey

Abstract
Changes in exon–intron structures and splicing patterns represent an important mechanism for the evolution of gene functions and species-specific regulatory networks. Although exon creation is widespread during primate and human evolution and has been studied extensively, much less is known about the scope and potential impact of human-specific exon loss events. Historically, transcriptome data and exon annotations are significantly biased toward humans over nonhuman primates. This ascertainment bias makes it challenging to discover human-specific exon loss events. We carried out a transcriptome-wide search of human-specific exon loss events, by taking advantage of RNA sequencing (RNA-seq) as a powerful and unbiased tool for exon discovery and annotation. Using RNA-seq data of humans, chimpanzees, and other primates, we reconstructed and compared transcript structures across the primate phylogeny. We discovered 33 candidate human-specific exon loss events, among which six exons passed stringent experimental filters for the complete loss of splicing activities in diverse human tissues. These events may result from human-specific deletion of genomic DNA, or small-scale sequence changes that inactivated splicing signals. The impact of human-specific exon loss events is predominantly regulatory. Three of the six events occurred in the 5′ untranslated region (5′-UTR) and affected cis-regulatory elements of mRNA translation. In SLC7A6, a gene encoding an amino acid transporter, luciferase reporter assays suggested that both a human-specific exon loss event and an independent human-specific single nucleotide substitution in the 5′-UTR increased mRNA translational efficiency. Our study provides novel insights into the molecular mechanisms and evolutionary consequences of exon loss during human evolution.

Key words: exon loss, evolution, RNA-seq, splicing, primate.

Introduction
A major interest in evolutionary biology is to elucidate the origin of evolutionary novelties. Evolution can create novel functions and regulatory programs through a variety of mechanisms, such as creation of new genes (Knowles and McLysaght 2009; Wu et al. 2011), exaptation of nonfunctional sequences including transposable elements (Muotri et al. 2007; Feschotte 2008), and small-scale sequence changes that alter gene regulation or protein function (Enard et al. 2002; Haygood et al. 2007; Wang et al. 2007; Maricic et al. 2013). Additionally, lineage-specific loss of functional genetic materials can play a role in adaptive evolution (Wang et al. 2006; McLean et al. 2011). For example, a number of genes were inactivated in humans after the divergence from chimpanzees, and these gene loss events were thought to play an active role in the evolution of human-specific traits (Wang et al. 2006). Moreover, human-specific loss of regulatory DNA occurred in hundreds of genomic loci and may contribute to unique human traits such as the expansion of brain size and the loss of penile spines (McLean et al. 2011).

Recently, the evolution of exon–intron structures and gene splicing patterns has emerged as an important mechanism for the evolution of gene functions and species-specific regulatory networks (Xing and Lee 2006; Barbosa-Morais et al. 2012; Merkin et al. 2012). The vast majority of genes in multicellular eukaryotes have multiple exons, and alternative splicing of multiexon genes can produce distinct mRNA and protein isoforms from a single gene locus (Nilsen and Graveley 2010). These alternative isoforms can be viewed as “internal paralogs” of a gene to allow for an accelerated rate of gene evolution (Boue et al. 2003; Modrek and Lee 2003). The most well-studied type of splicing evolution is the creation of new exons (Sorek 2007). During evolution, new exons can be added to existing functional genes through exonization of nonexonic sequences or duplication of existing exons (Kondrashov and Koonin 2001; Letunic et al. 2002; Wang et al. 2005; Zhang and Chasin 2006). A series of studies have systematically analyzed exon creation in higher eukaryotes using transcriptome data generated by cDNA/expressed sequence tag (EST) sequencing (Modrek and Lee 2003; Wang et al. 2005; Zhang and Chasin 2006; Alekseyenko et al. 2007; Corvelo and Eyras 2008), exon microarray (Lin et al. 2008, 2009), and RNA sequencing (RNA-seq; Shen et al. 2011). Collectively, these studies demonstrate that exon creation is widespread during recent primate and human evolution, and provides an important means for recruiting nonfunctional sequences such as transposable elements into functional genes. For example, Zhang and Chasin (2006) used a multispecies comparison to identify over 2,000 exon creation events during primate and human evolution. Sela et al. (2007)
identified over 1,500 human exons derived from primate-specific Alu retrotransposons, and these exons are by definition primate-specific. Through subsequent evolution, some of these new exons can acquire functional roles. For example, a primate-specific new exon of the gene encoding the RNA editing enzyme ADARB1 (also known as ADAR2), created through the exonization of an intronic Alu element, inserts a peptide segment into the catalytic domain of ADARB1 and alters the catalytic activity of the protein product (Gerber et al. 1997). Another study shows that primate-specific exons in the 5′ untranslated region (5′-UTR) of human genes often contain cis elements of translational regulation and can alter the translational efficiency of the mRNA (Shen et al. 2011).

The opposite scenario to exon creation is species-specific exon loss in functional genes. A classic example is the human-specific exon loss in the CMAH gene, which encodes a hydroxyxylase responsible for the biosynthesis of N-glycolyl-neuraminic acid (Neu5Gc) from N-acetyllneuraminic acid (Neu5Ac) (Chou et al. 1998, 2002; Irie et al. 1998; Hayakawa et al. 2001). Neu5Ac and Neu5Gc are sialic acids that play important roles in cell–cell and cell–pathogen interactions (Schauer 2009). A 92-bp exon of CMAH was deleted from the human genome via Alu-mediated replacement of the genomic DNA (Hayakawa et al. 2001). This causes the frameshift of the downstream coding sequence, turning a functional CMAH gene encoding an active enzyme into a pseudogene. Consequently, human cells lack the ability to synthesize Neu5Gc, and this results in profound biochemical and physiological differences between humans and nonhuman primates (Varki 2001, 2010). For example, the loss of Neu5Gc and the resulting increase in Neu5Ac levels in human cells have affected humans’ interactions with Neu5Gc- or Neu5Ac-binding pathogens. Consequently, humans are more resistant to Neu5Gc-binding pathogens but more susceptible to Neu5Ac-binding pathogens, and this has important implications for human evolution (Varki 2001, 2010).

Despite this fascinating example, our knowledge of exon loss during recent human evolution is extremely limited. Unlike exon creation, which has been studied extensively in the past decade (Modrek and Lee 2003; Wang et al. 2005; Zhang and Chasin 2006; Alekseyenko et al. 2007; Corvelo and Eyras 2008; Lin et al. 2008, 2009; Shen et al. 2011), there are few genome-wide studies of exon loss (Alekseyenko et al. 2007) especially for recent exon loss events in the human lineage. This is understandable, considering the difficulty of identifying human-specific exon loss events using genome and transcriptome data. It is well known that the presence of splice site sequences in the genome is not a reliable indicator of exon splicing (Alekseyenko et al. 2007). To confidently discover exon loss events in the human lineage, we need to identify exons not seen in human transcripts but present in the transcriptomes of chimpanzees and other nonhuman primates. Historically, transcriptome data and exon annotations are significantly biased toward humans over nonhuman primates. In fact, comparative annotation using human mRNA sequences has long been an important strategy for annotating nonhuman primate genomes (Chimpanzee Sequencing and Analysis Consortium 2005; Florea et al. 2005). This ascertainment bias makes it challenging to discover human-specific exon loss events. Two early studies used the human-chimpanzee genome alignment to identify chimpanzee genomic regions that match exons in chimpanzee gene models or nonprimate vertebrate mRNAs but are deleted in the human genome (Sen et al. 2006; Hahn et al. 2007). This strategy identified several putative human-specific exon loss events. However, there was no direct evidence to support the splicing of these exons in nonhuman primates, and almost all reported events were in hypothetical genes later removed from gene annotation databases. It must be noted that any detection strategy based on human-specific deletion of genomic DNA is inherently limited, because it will miss exon loss events due to small-scale sequence changes that inactivated splicing signals in humans.

In this study, we carried out a genome-wide discovery of human-specific exon loss events, by taking advantage of RNA-seq as a powerful technology for exon discovery and annotation. RNA-seq enables accurate characterization and comparison of transcript structures across closely related species (Blekman et al. 2010; Perry et al. 2012). Using RNA-seq data, one can identify exons and predict transcript structures de novo even in the absence of prior transcriptome annotations (Guttman et al. 2010; Trapnell et al. 2010). Here, we used deep RNA-seq data of human and nonhuman primate tissues to reconstruct and compare transcript structures across the primate phylogeny. By identifying exons absent in human RNAs but present in the RNAs of chimpanzees and outgroup species, we discovered 33 candidate human-specific exon loss events, among which six exons passed stringent experimental filters for the complete loss of splicing activities in diverse human tissues. These data allowed us to investigate the mechanisms and potential consequences of exon loss during human evolution.

Results

Computational Discovery of Candidate Human-Specific Exon Loss Events from RNA-seq Data

We developed a computational pipeline to discover candidate human-specific exon loss events (fig. 1A). Briefly, we used a published deep RNA-seq data set of six tissues including brain (cerebral cortex or whole brain without cerebellum), cerebellum, heart, kidney, liver, and testis of human, chimpanzee, orangutan, and rhesus macaque (Brawand et al. 2011), as well as known transcript annotations of human and mouse genes (see details in Materials and Methods). For each tissue and species, we consolidated RNA-seq reads from all biological replicates and mapped reads to the respective genomes using TopHat (Trapnell et al. 2009). We then used Cufflinks (Trapnell et al. 2010) to perform de novo reconstruction of transcript structures. To compare transcript structures across species, we matched genomic regions between chimpanzee and other species using the UCSC pairwise genome alignments (Materials and Methods). From all internal (spliced) exons in chimpanzee transcripts, we identified
those that were missing from human transcripts (known transcript annotations + Cufflinks transcript reconstructions) but present in the transcripts of at least one outgroup species (orangutan, rhesus macaque, or mouse). To distinguish exon loss events from gene loss events, we required that the flanking exons in the chimpanzee transcripts must have matching exons in the orthologous human genes. To further ensure that we comprehensively captured human exons and the chimpanzee exons were indeed absent from human transcriptomes, we analyzed independent human RNA-seq data covering 19 tissues, including 16 tissues in the Human Body Map 2.0 data set and 3 tissues from different anatomical compartments of the human placenta (Kim et al. 2012; see Materials and Methods). We used Cufflinks to reconstruct transcript structures for this expanded set of human tissues and searched chimpanzee exons against the Cufflinks-reconstructed human exons. In total, 264 potential human-specific exon loss events were identified using this computational procedure (fig. 1A).

We performed additional computational analyses to remove false positive findings of exon loss. Our computational pipeline could produce false positives due to a variety of reasons, such as missing segments in the current human genome assembly, errors in the UCSC genome alignments, or the failure of Cufflinks to detect spliced exons from human RNA-seq data. To address these three potential sources of...
false positives, we designed three corresponding computational filters (see details in Materials and Methods). We applied these three filters in a sequential manner and they removed 137, 77, and 17 exons respectively. After these filters, we obtained a final list of 33 candidate human-specific exon loss events for further experimental analyses.

For proof of concept, we checked if our computational pipeline could identify the well-known human-specific exon loss event in the CMAH gene (Chou et al. 1998, 2002; Hayakawa et al. 2001). Indeed, our analysis of human and nonhuman primate RNA-seq data successfully captured this exon loss event. As seen from the RNA-seq read density plot and the Cufflinks transcript reconstructions (fig. 1B), this CMAH exon and its flanking exons had robust RNA-seq signals in chimpanzee, orangutan, and rhesus tissues. In contrast, in human tissues there was no signal for the exon of interest, whereas the flanking exons still had robust RNA-seq signals, indicating human-specific exon loss.

**Experimental Validation of Candidate Human-Specific Exon Loss Events in Diverse Human and Nonhuman Primate Tissues**

Next, we performed experimental validation of the 33 candidate human-specific exon loss events. Of these 33 events, four events appeared to be due to human-specific deletion of the genomic DNA that contained the exon. For these four events, we performed polymerase chain reaction (PCR) analyses of genomic DNAs from human and nonhuman primates to confirm the human-specific genomic deletion. In the remaining events, the chimpanzee exon had an aligned orthologous region in the human genome, but there was no evidence for exon splicing based on human RNA-seq data and known transcript annotations. For these events, we performed reverse transcription polymerase chain reaction (RT-PCR) analyses of a large panel of human, chimpanzee, and rhesus tissues (supplementary table S1, Supplementary Material online) to verify the presence or absence of the exons in mRNA transcripts. Specifically, for each exon in each species, we designed a maximum of two primer pairs, with one pair amplifying the mRNA sequence between the exon and its upstream exon and the other pair amplifying the mRNA sequence between the exon and its downstream exon. We declared an exon to be absent from a given RNA sample if both primer pairs failed to amplify the expected RT-PCR products. We should point out that most of the human tissues represented in the original RNA-seq data set were from a limited number of individuals. For example, the 16 tissues in the Human Body Map 2.0 data set were all from single donors. Thus, to ensure that the identified exon loss events reflected true species differences in exon usage, we carried out our validation experiments on human RNA samples pooled from a large number of individuals (supplementary table S1, Supplementary Material online).

Our experimental analysis confirmed the complete loss of exon splicing of six exons in all tested human tissues, whereas their orthologous exons were detected in RNAs from chimpanzees and other nonhuman primates (table 1 and supplementary table S2, Supplementary Material online). Of these six exon loss events, three (in RIIAD1, SLC7A6, and CMAH) were due to human-specific genomic deletion. For example, in SLC7A6 the alignment of the human and nonhuman primate genomes indicated human-specific deletion of an approximately 2-kb genomic region, which removed the exon from the human gene (fig. 2A). Our PCR analysis of genomic DNA indicated that this genomic deletion was a fixed change during human evolution, because we observed homozygous deletion of this region in all eight tested human individuals from major human populations (fig. 2B and supplementary table S1, Supplementary Material online). In contrast, RT-PCR analyses showed that the exon was present in all tested chimpanzee and rhesus tissues (fig. 2C and D). In fact, RNA-seq data suggested that this exon was always 100% included in the transcripts in chimpanzee and rhesus tissues (supplementary table S3, Supplementary Material online). Together, these data suggest that a constitutive SLC7A6 exon in the most recent common ancestor of humans and chimpanzees was lost during human evolution. We confirmed the same pattern of human-specific genomic deletion for the exons in RIIAD1 and CMAH (supplementary figs. S1 and S2, Supplementary Material online). Additionally, we checked the Database of Genomic Variants (DGV, http://dgv.tcag.ca/, last accessed November 28, 2014) (MacDonald et al. 2014), which contains structural variants of the human genome reported by the 1000 Genomes Project. None of the three events is polymorphic in DGV, further confirming the notion that these three genomic deletion events are fixed in human populations. The other three events (in KANK1, ZMYND11, and LOC150568) were in conserved regions between the human and chimpanzee genomes. Of these, two exon loss events (in KANK1 and ZMYND11) were apparently due to single nucleotide changes that disrupted the splice sites in humans. In KANK1, we confirmed the absence of exon splicing in all tested human tissues (fig. 3A), whereas the exon was present in almost all chimpanzee and rhesus tissues (fig. 3B and C). The alignment of the human and nonhuman primate genomes revealed a G to A change in the first intronic nucleotide downstream of the exon, which disrupted the highly conserved GT dinucleotide of the 5′-splice site (fig. 3D). Likewise, we observed a human-specific G to C nucleotide change that disrupted the 3′-splice site of the ZMYND11 exon (supplementary fig. S3, Supplementary Material online). We also validated the exon loss event in the noncoding RNA gene LOC150568 (supplementary fig. S4, Supplementary Material online). The 5′- and 3′-splice sites of this exon remained intact in the human genome. Therefore, this exon loss event may be due to other human-specific sequence changes that abolished exon splicing in the human lineage. However, we should caution that because its splice sites are intact in the human genome, we cannot completely rule out the possibility that this exon may be spliced in other cell types or developmental stages not tested in this study.

Twenty-seven of the 33 candidate events had RT-PCR signals in human tissues. We should note that our RT-PCR detection of exon splicing by placing one primer on the candidate exon is highly sensitive, because we will detect...
<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Chimpanzee Exon Regiona</th>
<th>Exon Length</th>
<th>Location</th>
<th>Mechanism</th>
<th>Consequence</th>
<th>Exon Splicing Levels in Nonhuman Primatesb</th>
<th>Supported Tissuesc</th>
<th>Gene Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>KANK1</td>
<td>chr9:581223–581348</td>
<td>126</td>
<td>5'-UTR</td>
<td>Splice site disruption</td>
<td>Exon loss removes 36, 33, and 12 bp uORFs</td>
<td>Constitutive</td>
<td>Widely includedd</td>
<td>Involved in the control of cytoskeleton formation by regulating actin polymerization</td>
</tr>
<tr>
<td>ZMYND11</td>
<td>chr10:244360–244475</td>
<td>116</td>
<td>CDS</td>
<td>Splice site disruption</td>
<td>Exon loss removes an alternative isoform subject to NMD</td>
<td>Alternative (0–16%)</td>
<td>Widely includedd</td>
<td>Corepressor of transcription (Probable)</td>
</tr>
<tr>
<td>RIIAD1</td>
<td>chr1:130786076–130786187</td>
<td>112</td>
<td>5'-UTR</td>
<td>Genomic deletion</td>
<td>Exon loss removes 48 and 117 bp uORFs</td>
<td>Constitutive</td>
<td>Expressed and included in brain, testis, kidney, and heart</td>
<td>Protein phosphorylation (Predicted)</td>
</tr>
<tr>
<td>SLC7A6</td>
<td>chr16:8044023–8044073</td>
<td>51</td>
<td>5'-UTR</td>
<td>Genomic deletion (Alu mediated)</td>
<td>Exon loss shortens a 264 bp uORF to 213 bp</td>
<td>Constitutive</td>
<td>Widely includedd</td>
<td>Uptake of amino acids</td>
</tr>
<tr>
<td>CMAH</td>
<td>chr6:25603207–25603298</td>
<td>92</td>
<td>CDS</td>
<td>Genomic deletion (Alu mediated)</td>
<td>Exon loss inactivates a coding gene to a pseudogene</td>
<td>Constitutive</td>
<td>Widely includedd</td>
<td>Cytidine monophospho-N-acetylneuraminic acid hydroxylase</td>
</tr>
<tr>
<td>LOC150568</td>
<td>chr2a:105599278–105599442</td>
<td>165</td>
<td>In noncoding RNA</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Alternative (25–100%)</td>
<td>Expressed and included in brain, heart, and testis</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

aCoordinates are based on chimpanzee panTro2 assembly.
bThe exon splicing levels are based on the RNA-seq data of chimpanzee, orangutan and rhesus macaque (supplementary table S3, Supplementary Material online).
cBased on RNA-seq data and RT-PCR results from this study.
dExpressed and spliced in all the chimpanzee and rhesus tissues used for RT-PCR analyses.
the exon signal as long as the exon is spliced (even at low levels) in some of the individuals that comprised the pooled tissue sample. Thus, we performed additional analyses to investigate the splicing levels of these exons in human and nonhuman primate tissues. Specifically, we randomly selected 16 of these 27 exons and measured their exon inclusion levels in human, chimpanzee, and rhesus tissues by semiquantitative RT-PCR using primers designed on flanking exons. We estimated exon inclusion levels based on the band intensities of RT-PCR products corresponding to exon inclusion and skipping isoforms. We found that most of these exons had zero or very low exon inclusion levels in all or the vast majority of human tissues, whereas their orthologous exons had much higher inclusion levels in chimpanzee and rhesus tissues (supplementary table S4, Supplementary Material online; also see supplementary fig. S5A–C, Supplementary Material online, for a few representative examples). These lowly included exons were not found in existing human transcript annotations or in the RNA-seq data of diverse human tissues. Together, our results suggest that most of these 27 exons had a significant reduction in splicing levels during human evolution or loss of splicing in most human tissues, and may have interesting evolutionary implications. Nonetheless, for our further studies we focused on the six exons with complete loss of exon splicing in all human tissues.

**Characteristics of Human-Specific Exon Loss Events**

Of the six events with complete loss of exon splicing in all human tissues, five were from protein-coding genes with known functions whereas one was from a noncoding RNA gene (LOC150568). The five protein-coding genes are involved in a variety of cellular functions such as the control of cytoskeleton formation (KANK1), transcriptional regulation (ZMYND11, also known as BS69), protein phosphorylation (RIIAD1), amino acid transport (SLC7A6), and cell metabolism (CMAH). We examined the splicing levels of these exons in nonhuman primate tissues using RNA-seq data (supplementary table S3, Supplementary Material online). In four events (KANK1, RIIAD1, SLC7A6, and CMAH), the exon was constitutively spliced (i.e., 100% exon inclusion) in nonhuman primate tissues. The exon in ZMYND11 was spliced at low levels in nonhuman primates. The exon in LOC150568 was also an alternative exon with low exon inclusion levels in nonhuman primates whereas the expression of this noncoding RNA appeared to be tissue-restricted. The observation that four of the six exons were constitutive exons in nonhuman primates...
indicates that human-specific exon loss could remove exons with high splicing activities, resulting in the use of distinct major mRNA isoforms in human and chimpanzee tissues.

We investigated the potential impacts of these exon loss events on gene expression and function. Of the five events in protein-coding genes, two were in coding regions (CMAH and ZMYND11) whereas the other three were in the 5'-UTR. The exon loss event in CMAH inactivated the gene and played an important role in the adaptive evolution of humans (Varki 2001, 2010). The exon in ZMYND11 was located in the coding region and the inclusion of this exon produced a minor mRNA isoform with a premature termination codon subject to the nonsense-mediated decay (NMD) pathway. Therefore, this exon likely represented splicing noise in nonhuman primates, which was removed from the human gene. The other three events (KANK1, RIIAD1, and SLC7A6) were constitutive internal exons in the 5'-UTR. All three events were predicted to remove or shorten upstream open reading frames (uORFs) within the 5'-UTR of human genes (table 1). uORFs have a widespread function in repressing mRNA translation and the strength of such translational repression is positively correlated with the uORF length (Calvo et al. 2009; Chatterjee and Pal 2009; Barbosa et al. 2013). Therefore, the three human-specific exon loss events in the 5'-UTR may increase the translational efficiency of these genes in the human lineage. Of note, our prior study of exon creation events shows that newly created primate-specific exons are preferentially located in the 5'-UTR and regulate translational efficiency through uORFs or other types of cis-regulatory elements of mRNA translation (Shen et al. 2011). Here, our new data indicate that both exon creation and loss events during human evolution have a strong preference toward the 5'-UTR and may impact gene regulation by modulating the rate of mRNA translation.
To test whether human-specific exon loss in the 5′-UTR could affect mRNA translational efficiency, we selected the exon loss event in SLC7A6 for detailed analyses. This event is interesting for several reasons. SLC7A6 is a broadly expressed gene in human and nonhuman primate tissues and encodes an amino acid transporter involved in the cellular uptake of dibasic amino acids and certain neutral amino acids (Broer et al. 2000). The exon is constitutively spliced in all nonhuman primate tissues but completely lost in humans due to genomic deletion. In the SLC7A6 mRNA from nonhuman primates, there was a uORF of 88 codons in length with the stop codon of the uORF located 50 bp downstream of the start codon of the canonical ORF (fig. 4A). The human-specific exon loss event removed a 51 bp exon, causing an in-frame deletion within the uORF and shortened the uORF from 88 codons to 71 codons.

We hypothesized that this human-specific exon loss event increased the translational efficiency of the SLC7A6 mRNA by shortening the uORF. To test this hypothesis, we made different versions of the SLC7A6 5′-UTR and analyzed their translational efficiency using a dual luciferase reporter construct pIRES-Luc2 as described previously (Shen et al. 2011). For each 5′-UTR sequence, the resulting reporter construct expressed both the Firefly luciferase and the Renilla luciferase bicistronically. The Renilla luciferase was fused downstream of the cloned 5′-UTR sequence, whereas the Firefly luciferase translation was driven independently by an internal ribosome entry site (IRES) and was not regulated by the cloned 5′-UTR. Because the stop codon of the uORF in the endogenous SLC7A6 gene is located downstream of the canonical translation start site, to test the effect of the uORF on 5′-UTR translational efficiency using luciferase assays, we artificially introduced a stop codon in the 5′-UTR upstream of the Renilla luciferase translation start site, as indicated by the triangle in the schematic diagram. *P ≤ 0.05; **P ≤ 0.01; ***P ≤ 0.001; NS, no significance; paired t-test. Error bar: standard error of the mean (SEM). (C) Results of the dual-luciferase reporter assays when there is no stop codon for the uORF in the 5′-UTR.

**Fig. 4.** Human-specific exon loss in SLC7A6 increases mRNA translational efficiency by shortening the uORF. (A) Schematic diagram of the gene structures of human, chimpanzee, and rhesus macaque SLC7A6. The coding regions are bolded. The uORFs are shown above the gene structure diagrams. The genomic region deleted in humans is illustrated by dashed box. (B) Dual-luciferase reporter assays to measure the translational efficiency of different SLC7A6 5′-UTR constructs in the HeLa cell line. Because the stop codon of the uORF in the endogenous SLC7A6 gene is located downstream of the canonical translation start site, to test the effect of the uORF on 5′-UTR translational efficiency using luciferase assays, we artificially introduced a stop codon in the 5′-UTR upstream of the Renilla luciferase translation start site, as indicated by the triangle in the schematic diagram. *P ≤ 0.05; **P ≤ 0.01; ***P ≤ 0.001; NS, no significance; paired t-test. Error bar: standard error of the mean (SEM). (C) Results of the dual-luciferase reporter assays when there is no stop codon for the uORF in the 5′-UTR.

**Human-Specific Exon Loss in SLC7A6 Increases mRNA Translational Efficiency**

To test whether human-specific exon loss in the 5′-UTR could affect mRNA translational efficiency, we selected the exon loss event in SLC7A6 for detailed analyses. This event is interesting for several reasons. SLC7A6 is a broadly expressed gene in human and nonhuman primate tissues and encodes an amino acid transporter involved in the cellular uptake of dibasic amino acids and certain neutral amino acids (Broer et al. 2000). The exon is constitutively spliced in all nonhuman primate tissues but completely lost in humans due to genomic deletion. In the SLC7A6 mRNA from nonhuman primates, there was a uORF of 88 codons in length with the stop codon of the uORF located 50 bp downstream of the start codon of the canonical ORF (fig. 4A). The human-specific exon loss event removed a 51 bp exon, causing an in-frame deletion within the uORF and shortened the uORF from 88 codons to 71 codons.

We hypothesized that this human-specific exon loss event increased the translational efficiency of the SLC7A6 mRNA by shortening the uORF. To test this hypothesis, we made different versions of the SLC7A6 5′-UTR and analyzed their translational efficiency using a dual luciferase reporter construct pIRES-Luc2 as described previously (fig. 4B) (Shen et al. 2011). For each 5′-UTR sequence, the resulting reporter construct expressed both the Firefly luciferase and the Renilla luciferase bicistronically. The Renilla luciferase was fused downstream of the cloned 5′-UTR sequence, whereas the Firefly luciferase translation was driven independently by an internal ribosome entry site (IRES) and was not regulated by the cloned 5′-UTR. Because the stop codon of the uORF in the endogenous gene is located downstream of the canonical translation start site, to test the effect of the uORF on translational efficiency using luciferase assays, we artificially introduced a stop codon in the 5′-UTR upstream of the Renilla luciferase translation start site (fig. 4B). The translational efficiency of each 5′-UTR construct was calculated as the ratio between the Renilla luciferase and the Firefly luciferase activities. As there was an independent human-specific nucleotide substitution in the uORF, we also made additional 5′-UTR

...
constructs to test the effect of this nucleotide substitution on mRNA translation.

As shown in figure 4B, when transfected into the HeLa cells the human version of the 5′-UTR had a significantly higher translational efficiency as compared with the chimpanzee version of the 5′-UTR (“Human skipping” vs. “Chimp inclusion”). Interestingly, the exon loss event and the single nucleotide change in the human 5′-UTR appeared to have synergistic effects and both increased mRNA translational efficiency (fig. 4B). For example, when we inserted the exon back into the human version of the 5′-UTR (“Human inclusion”), the mRNA translational efficiency was greatly reduced as compared with the human 5′-UTR without the exon (“Human skipping”). Similarly, the human-specific single nucleotide substitution within the uORF increased mRNA translational efficiency (“Human skipping” vs. “Chimp skipping”). This nucleotide substitution did not change the amino acid encoded by the codon, but replaced a rare serine codon (AGU) in the chimpanzee sequence with a more common (optimal) serine codon (AGC) in the human sequence. Previous studies show that the replacement of rare codons by common codons in the uORF can increase translation of the downstream protein-coding ORF (Meijer and Thomas 2003; Qiao et al. 2011). We obtained similar results when we repeated the luciferase reporter assays in the JEG-3 cell line (supplementary fig. S6, Supplementary Material online).

To confirm that the observed translational effect was mediated by the uORF, we mutated the stop codon of all tested 5′-UTR constructs. After the stop codon was removed, there was no significant difference in translational efficiency among different SLC7A6 5′-UTR constructs (fig. 4C).

Human-Specific Exon Loss in SLC7A6 Resulted from Alu–Alu-Recombination Mediated Genomic Deletion

Finally, we investigated the molecular mechanism for the human-specific exon loss of SLC7A6. This exon loss event was due to the deletion of an approximately 2-kb genomic region during human evolution (fig. 2). To elucidate the molecular mechanism for this genomic deletion, we aligned the chimpanzee genomic sequence with the orthologous human genomic sequence. We found that in the chimpanzee genome, the human-deleted region was flanked by two Alu elements (AluSq and AluSc) (fig. 5A), whereas in the human genome there was a single Alu element (AluSx) spanning the breakpoint. We then aligned the human Alu element to the two chimpanzee Alu elements located at the two ends of the deleted region. We found that the human Alu was a chimera of the two chimpanzee Alus, with the majority of the human Alu (at its 5′-end) matching the upstream chimpanzee Alu element and a small 3′-end segment of the human Alu matching the downstream chimpanzee Alu element (fig. 5B). The observed sequence pattern of the human Alu is consistent with the model of genomic deletion mediated by Alu–Alu recombination, a process known to be responsible for genomic deletion in evolution and disease (Sen et al. 2006). Our data suggest that the two Alus in the ancestral sequence recombined during human evolution, which deleted the approximately 2-kb genomic region and removed the exon from the human gene (fig. 5C).

Discussion

We carried out a transcriptome-wide search for exon loss events during human evolution, by comparing RNA-seq based de novo reconstructions of transcript structures across the primate phylogeny. We sought to identify exons spliced into the transcriptomes of chimpanzees and other nonhuman primates but absent in human RNAs. Using stringent computational and experimental filters, we identified six exons with complete loss of splicing activities in diverse human tissues, as well as additional exons with significant reduction in splicing levels during recent human evolution. Of these six experimentally confirmed exon loss events, three were due to human-specific deletion of genomic sequences, whereas others resulted from small-scale sequence changes that inactivated splicing signals. We identified a well-known human-specific exon loss event in the CMAH gene (Chou et al. 1998, 2002; Hayakawa et al. 2001), indicating that our computational strategy can effectively discover lineage-specific exon loss events from RNA-seq data.

Compared with exon creation which is widespread during human evolution (Zhang and Chasin 2006), human-specific exon loss events are markedly rare. Only one (in the CMAH gene) was well-documented in the literature (Chou et al. 1998, 2002; Hayakawa et al. 2001), and five additional events were identified from this study using extensive RNA-seq data. Of note, four out of the six human-specific exon loss events involve exons constitutively (100%) spliced in nonhuman primate tissues. This observation is consistent with a previous study of exon creation and loss events over much longer timescales of vertebrate evolution using ESTs, which reported that exon loss events were not biased toward lowly included exons (Aleksyenko et al. 2007). Thus, the difference in the frequencies of exon loss versus creation events may be attributed to the evolutionary barriers to these two types of splicing changes. It is well known that the vast majority of newly created exons in human genes are alternatively spliced at low splicing levels (Modrek and Lee 2003). Because the ancestral transcript isoform is still produced as the major isoform, a newly created exon may not be subject to strong purifying selection and could provide evolutionary intermediates for subsequent adaptive changes (Boue et al. 2003; Xing and Lee 2006). In contrast, the definition of human-specific exon loss events requires an exon to be detected (and conserved) in chimpanzee and other nonhuman primate species but absent only in human RNAs. Such events are more likely to involve exons with high splicing activities and functional importance and consequently undergo strong purifying selection. The deleterious effects of exon loss are evident in human genetics studies, in which mutations that disrupt exon splicing frequently cause human diseases (Sterne-Weiler et al. 2011). Taken together, these data suggest that during human evolution, almost all mutations that inactivated ancient functional exons were removed by purifying selection, while a small number survived and became fixed.
in humans, likely by acquiring adaptive benefits as in the case of CMAH (Varki 2001, 2010).

Although complete human-specific loss of exon splicing is rare, we should note that exon loss does not necessarily have to be viewed as a binary event, and a significant but incomplete human-specific reduction in exon splicing levels could still have important functional and regulatory consequences. In this study, of the 33 candidate exon loss events identified by the RNA-seq analysis, 27 had detectable RT-PCR signals in at least some of the tested human tissues. A subset of these exons were analyzed by semiquantitative RT-PCR in human and nonhuman primate tissues. The majority of these exons analyzed had a significant reduction in splicing levels in human tissues as compared with nonhuman primate tissues (supplementary table S4, Supplementary Material online). Moreover, for these 27 events, we reanalyzed the corresponding gene expression levels of 17 events that had normalized RNA-seq RPKM (Reads Per Kilobase per Million mapped

**Fig. 5.** Human-specific exon loss in SLC7A6 resulted from Alu–Alu-recombination mediated genomic deletion. (A) Alignment matrix of human and chimpanzee genomic sequences surrounding the exon loss event. The alignment of one human Alu and two chimpanzee Alus near the crossover site are highlighted in orange boxes. (B) The detailed sequence alignment of one human Alu and two chimpanzee Alus as highlighted in (A). The positions in the consensus Alu sequence are numbered, and only the second half of the Alu is shown. Only nucleotides different from the other two sequences are labeled. Blocks of sequences with high similarity in the alignment are represented by the same color. The ellipse indicates the putative crossover site where the Alu–Alu recombination event likely occurred during human evolution. (C) Schematic diagram illustrating the proposed molecular mechanism for human-specific exon loss in SLC7A6. The directions of arrows in the genome sequence indicate the orientations of Alu elements. Note that the position of the crossover site in this diagram is not drawn to the scale of SLC7A6.
reads) values provided in the original publication (Brawand et al. 2011). Most of these 17 genes were broadly and highly expressed. In human, chimpanzee, and rhesus macaque tissues, all 17 genes were expressed (RPKM > 1) in at least one tissue, and more than half of the 17 genes (ten in human, nine in chimpanzee and rhesus macaque) were expressed in all six tissues from the original study. Furthermore, the majority of them (13 in human and chimpanzee, 12 in rhesus macaque) were highly expressed (RPKM > 10) in at least one tissue. These results suggest that significant but incomplete reduction in exon splicing levels may be more widespread and target abundantly expressed and functionally important genes during human evolution. It is possible that some of these events may reflect an intermediate step to complete exon loss, analogous to newly created exons being spliced at low levels. Finally, we also investigated the potential molecular mechanisms of the generally reduced splicing levels of these exons in human tissues. For 7 of these 27 cases, the splice site scores of the human exons were significantly lower than those of the orthologous exons in nonhuman primates, as reflected by at least 4-fold reduction in the likelihood odd ratios of matching to the consensus MAXENT splice site model (Yeo and Burge 2004). This suggests that a minor subset of these 27 cases may be attributed to human-specific nucleotide changes that weakened the splice sites of the human exons. For other cases, the molecular mechanisms for their generally reduced splicing levels in human tissues are unclear, which is understandable given the complexity of splicing regulation in mammalian cells.

Our study raises the question about the functional and regulatory consequences of human-specific exon loss events. Changes in RNA splicing patterns can affect gene products and functions at both protein and RNA levels, either by producing protein isoforms with divergent structural and functional characteristics (Black 2000) or by generating mRNA isoforms with distinct regulatory properties such as mRNA decay or translational efficiency (Lewis et al. 2003; Sterne-Weiler et al. 2013). For all six human-specific exon loss events validated in this study, the effects appear to be regulatory rather than protein-coding. For example, in CMAH, the exon loss event disrupted the ORF of the mRNA, turning a functional gene into a pseudogene. In ZMYND11, the human-specific exon loss event removed a minor mRNA isoform subject to mRNA nonsense-mediated decay. Of note, in three of the six events (KANK1, RIIAD1, and SLC7A6), the exon lost during human evolution was located in the mRNA 5′-UTR and the exon loss event removed or shortened uORFs in the human mRNA (table 1). These events are expected to increase the translational efficiency of the human mRNA. In SLC7A6, using luciferase reporter assays, we experimentally confirmed the effect of the exon loss event on mRNA translational efficiency (fig. 4). Although the total number of exon loss events is small, the high frequency of events occurring in the 5′-UTR (three out of six) is reminiscent of the observations made about species-specific exon creation events, which are also enriched in the mRNA 5′-UTR (Zhang and Chasin 2006; Lin et al. 2008; Shen et al. 2011) and may modulate translational efficiency (Shen et al. 2011). Together, these data suggest that the mRNA 5′-UTR is a hotspot of evolutionary changes in gene splicing patterns, and such changes may have played a regulatory role during human evolution by fine-tuning mRNA translational efficiency and protein production in a lineage-specific manner.

We performed detailed analyses of a human-specific exon loss event in SLC7A6. SLC7A6 encodes an amino acid transporter involved in the uptake of arginine, leucine, and glutamine and the release of arginine in exchange for glutamine (Broer et al. 2000). It plays a role in the trafficking of amino acids between brain cells (Broer et al. 2000). It also regulates nitric oxide (NO) synthesis through transport of arginine which is a substrate for the production of NO (Arancibia-Garavilla et al. 2003). The exon of interest is present in the 5′-UTR of SLC7A6 in nonhuman primates but absent from the human mRNA due to Alu-mediated deletion of the corresponding genomic region. This exon loss event shortens the uORF in the 5′-UTR. Because of the lack of a suitable antibody, we were unable to directly compare protein abundance of SLC7A6 in human and nonhuman primate tissues. Nonetheless, luciferase reporter assays in multiple cell lines consistently suggest that this human-specific exon loss event increases the translational efficiency of the human mRNA via its effect on the uORF. Intriguingly, besides this human-specific exon loss event, there is also a human-specific single nucleotide substitution in the 5′-UTR of SLC7A6. This U to C single nucleotide change in the human mRNA also increases the translational efficiency according to luciferase assays, by replacing a rare serine codon (AGU) with an optimal serine codon (AGC) which may enhance the translation of the downstream protein-coding ORF (Méijer and Thomas 2003; Qiao et al. 2011). The observation that two independent human-specific sequence changes in the SLC7A6 5′-UTR have synergistic effects on mRNA translation implies that increased translational efficiency of SLC7A6 may confer adaptive benefits and have been selected for during recent human evolution. Future studies should investigate whether these evolutionary changes to the SLC7A6 5′-UTR functionally contributed to the differences in metabolic activities of human and chimpanzee cells.

**Conclusions**

Our study represents a systematic, unbiased search for species-specific exon loss events in human transcriptomes. Using stringent computational and experimental filters, we identified six exons with complete loss of splicing activities in human transcriptomes after the divergence of humans and chimpanzees. Four of these six exons are constitutively spliced in nonhuman primate tissues, and three of them are located in the mRNA 5′-UTR and may affect translational efficiency. The overall low incidence of human-specific exon loss events may reflect the evolutionary barriers toward removing ancient exons with high splicing activities from functional genes. Nonetheless, a small number of exon loss events may have acquired adaptive benefits and become fixed in humans. By affecting mRNA translation or degradation, these events may have contributed to the regulatory evolution of the human transcriptome and proteome.
Materials and Methods

Data Source
We analyzed published RNA-seq data of six tissues (brain, cerebellum, heart, kidney, liver, testis) of human, chimpanzee, orangutan, and rhesus macaque (Brawand et al. 2011), with a total of approximately 1.6 billion 76 bp single-end or 101 bp paired-end Illumina RNA-seq reads (supplementary table S5, Supplementary Material online). We also analyzed additional human RNA-seq data covering 19 tissues, including 16 tissues in the Human Body Map 2.0 data set (http://www.ensembl.info/blog/2011/05/24/human-bodymap-2-0-data-from-illumina/, last accessed November 28, 2014) and 3 tissues from different anatomical compartments of the human placenta (Kim et al. 2012; supplementary table S5, Supplementary Material online).

RNA-seq Read Mapping and De Novo Transcript Reconstruction
We mapped RNA-seq reads of human, chimpanzee, orangutan, and rhesus macaque tissues to corresponding genomes (hg19, panTro2, ponAbe2, and rheMac2) using TopHat (version 1.4.1) (Trapnell et al. 2009) with default parameters. The uniquely mapped reads were then used for de novo transcript reconstruction using Cufflinks (version 1.2.0; Trapnell et al. 2010) with default parameters.

Identification of Candidate Human-Specific Exon Loss Events from RNA-seq Data
The basic idea of discovering human-specific exon loss events is to identify chimpanzee exons that are absent in the human transcriptomes but present in the transcriptomes of outgroup species. To test the presence or absence of a particular chimpanzee exon in other species, we matched the orthologous exon regions via the UCSC pairwise alignments among these genomes (hg19, panTro2, ponAbe2, and rheMac2) using the python package Pygr (http://code.google.com/p/pygr/, last accessed November 28, 2014).

To identify candidate human-specific exon loss events, for each internal (spliced) exon in chimpanzee transcripts as identified by Cufflinks, we first required that the exon cannot overlap with any human exon in either known human transcriptomes (UCSC [downloaded in September 2011], Ensembl [release 61], RefSeq [release 48], and Vega [release 44]) or human transcripts reconstructed de novo by Cufflinks using RNA-seq data (supplementary table S5, Supplementary Material online). To distinguish exon loss events from gene loss events, we required that the flanking exons in chimpanzee transcripts must have matching exons in the orthologous human genes. Here, we defined a pair of chimpanzee and human exons as matching if they had identical exon start or end positions according to the UCSC pairwise alignment and the strands were consistent.

To distinguish exon loss events in humans versus exon creation events in chimpanzees, we analyzed transcript annotations of outgroup species. Specifically, we required that for any candidate human-specific exon loss event, the chimpanzee exon must be present in at least one Cufflinks transcript of orangutan or rhesus macaque, or known mouse transcripts from UCSC (downloaded in December 2011), Ensembl (release 61), RefSeq (release 50), and Vega (release 45).

We applied three additional filtering criteria to remove potential false positive findings of human-specific exon loss events. First, we examined the alignment of human RefSeq RNA sequences to the chimpanzee genome, available at the UCSC Genome Browser database via the "Non-Chimp RefSeq Genes" track (downloaded in December 2011). All chimpanzee exons aligned to human RefSeq RNA sequences were removed from candidate human-specific exon loss events. Second, to confirm the validity of the UCSC genome alignment, we used another alignment tool Blat (Kent 2002) to map chimpanzee transcripts to the genomes of other species, and removed candidate exon loss events from consideration after manual inspection of the alignment result if the UCSC genome alignment was inconsistent with the Blat alignment. Third, to address potential failures of Cufflinks in exon detection, we examined the TopHat mapping of human RNA-seq reads. We removed candidate events if TopHat suggested the existence of splice sites in the human genomic region orthologous to the chimpanzee exon.

Human and Primate Tissue Samples
RNA samples of 18 human (Homo sapiens) tissues were purchased from Clontech (Mountain View, CA). Fifteen postmortem tissue samples from adult chimpanzees (Pan troglodytes) and 25 postmortem tissue samples from adult rhesus macaques (Macaca mulatta) were generously provided by the Southwest National Primate Research Center (San Antonio, TX). Chimpanzee and rhesus monkey fibroblast cell line samples were acquired from the Coriell Institute for Medical Research (Camden, NJ). The genomic DNA samples of human, chimpanzee, and rhesus macaque were from HapMap lymphoblastoid cell lines or primate fibroblast cell lines purchased from the Coriell Institute for Medical Research (Camden, NJ). Details of these RNA and DNA samples are provided in supplementary table S1, Supplementary Material online.

RT-PCR and Sequencing
Total RNA was extracted using TRizol (Invitrogen, Carlsbad, CA). High-capacity cDNA Reverse Transcription Kit (Applied Biosystems, Foster city, CA) was used for cDNA synthesis. For each exon or genomic region analyzed, regular PCR was carried out for 29–40 cycles. Final PCR products were resolved on 5% Tris-Borate-EDTA (TBE) polyacrylamide gels. For DNA fragments with ambiguous sizes, we conducted Sanger sequencing (University of Iowa DNA facility, Iowa City, IA). PCR primer sequences are shown in supplementary table S6, Supplementary Material online.
SLC7A6 5'‐UTR Constructs and Luciferase Reporter Assay

The intact chimpanzee 5'‐UTR (i.e., the 5'‐UTR including the exon of interest) of SLC7A6 was obtained from IDT Gene Synthesis Service (Integrated DNA Technologies, Inc, Coralville, IA). The pIRES‐Luc2 vector described in our previous study (Shen et al. 2011) was used to measure the translational efficiency of SLC7A6 5'‐UTR. Site‐directed mutagenesis was carried out to make different versions of the SLC7A6 5'‐UTR, and all mutant constructs were confirmed by sequencing (University of Iowa DNA facility, Iowa City, IA). Hela and JEG‐3 cells were grown in either Gibco Dulbecco’s Modified Eagle’s Medium (DMEM) (Invitrogen, Carlsbad, CA) or Minimum Essential Media (MEM) supplemented with 10% Fetal Bovine Serum (FBS). Lipofectamine 2000 (Invitrogen) was used in transient transfection. Dual‐color luciferase assay was carried out approximately 24 h after transfection following the manufacturer’s protocol (Promega, Madison, WI). Final results were confirmed by at least three independent experiments.

Supplementary Material

Supplementary figures S1–S6 and tables S1–S6 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Lan Lin for discussions of experimental design. This work was supported by the National Institutes of Health grant P51RR013986.

References


