BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 1 2011, pages 14–21
doi:10.1093/bioinformatics/btq612

Genome analysis

Advance Access publication October 29, 2010

Identification of human-specific transcript variants induced by DNA insertions in the human genome

Dong Seon Kim and Yoonsoo Hahn*

Department of Life Science (BK21 Program) and Research Center for Biomolecules and Biosystems, Chung-Ang University, Seoul 156-756, Korea

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Many genes in the human genome produce a wide variety of transcript variants resulting from alternative exon splicing, differential promoter usage, or altered polyadenylation site utilization that may function differently in human cells. Here, we present a bioinformatics method for the systematic identification of human-specific novel transcript variants that might have arisen after the human–chimpanzee divergence.

Results: The procedure involved collecting genomic insertions that are unique to the human genome when compared with orthologous chimpanzee and rhesus macaque genomic regions, and that are expressed in the transcriptome as exons evidenced by mRNAs and/or expressed sequence tags (ESTs). Using this procedure, we identified 112 transcript variants that are specific to humans; 74 were associated with known genes and the remaining transcripts were located in unannotated genomic loci. The original source of inserts was mostly transposable elements including L1, Alu, SVA, and human endogenous retroviruses (HERVs). Interestingly, some non-repetitive genomic segments were also involved in the generation of novel transcript variants. Insert contributions to the transcripts included promoters, terminal exons and insertions in exons, splice donors and acceptors and complete exon cassettes. Comparison of personal genomes revealed that at least seven loci were polymorphic in humans. The exaptation of human-specific genomic inserts as novel transcript variants may have increased human gene versatility or affected gene regulation.

Contact: hahny@cau.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 30, 2010; revised on October 4, 2010; accepted on October 26, 2010

1 INTRODUCTION

The human genome acquired many genetic modifications after the human–chimpanzee divergence, some of which may have played a significant role in the evolution of human traits. The transcription factor *FOXP2*, a gene involved in human speech and language, shows accelerated evolution in humans, and alterations within this gene may have been responsible for changes in the transcription of central nervous system development genes (Enard *et al.*, 2002; Konopka *et al.*, 2009). The *HAR1F* non-coding gene, expressed during cortical development, was shown to have increased the rate of nucleotide sequence substitution in humans (Pollard *et al.*, 2006). In addition, at least three human protein-coding genes have been shown to originate from ancestrally non-coding DNA (Knowles and McLysaght, 2009). Interestingly, loss of gene function has also been suggested to play a role in the acquisition of human-specific phenotypes (Olson, 1999). For example, *MYH16* and *BASE* were inactivated by a frameshift mutation (Hahn and Lee, 2005; Stedman *et al.*, 2004); *KRTHAP1* and *SERPINA13* by a nonsense mutation (Hahn and Lee, 2006; Winter *et al.*, 2001); and *CMAH*, *MOXD2* and *S100A15A* by an exon deletion (Chou *et al.*, 1998; Hahn *et al.*, 2007). Human-specific sequence modifications in *cis*-regulatory elements that alter gene expression patterns and hence may have driven the developmental evolution of humans have also been identified (Noonan, 2009).

Insertion of DNA fragments into a genome plays an important role in the shaping of the genome, including the insertion of mobile elements such as SINEs and LINEs (Deininger *et al.*, 2003), retroviruses (Dewannieux *et al.*, 2006), retrogenes (Marques *et al.*, 2005) and mitochondrial DNAs (Hazkani-Covo and Graur, 2007). Most inserted DNAs accumulate sequence substitutions and are eventually inactivated. However, some integrated genetic elements develop interactions with proximal genes, altering the structure or regulation of the adjacent genes. In many cases the insertion of Alu sequences is deleterious due to strong activation of splice sites, which disrupts host gene transcript integrity and leads to many genetic diseases (Knebelmann *et al.*, 1995; Mitchell *et al.*, 1991). This type of destructive insertion, a mechanism known as exon- or gene-trapping, is eventually removed from the gene pool by strong negative selection. Therefore, most Alu exons that survive are involved in alternative splicing, allowing a substantial portion of the transcript to encode the original functional proteins (Sorek *et al.*, 2002).

Exonized Alu elements can evolve to acquire protein-coding potential, increasing protein versatility (Krull *et al.*, 2005). The LINE-1 (L1) retrotransposon intrinsically possesses a poly-adenylation signal and promoter activity. L1 elements inserted into introns often break the host gene into two separate transcripts by a mechanism called gene breaking (Wheelan *et al.*, 2005). Alu and L1 transposable elements have been found in many human protein-coding genes, implying that they have played important roles in gene evolution (Lorenc and Makałowski, 2003; Nekrutenko and Li, 2001).

In this study, we developed a bioinformatics method to systematically identify human-specific transcript variants that might

*To whom correspondence should be addressed.

have arisen in the human genome after the human–chimpanzee divergence. First, we collected human-specific genomic insertions by comparing the human, chimpanzee and rhesus macaque genomes. Next, we examined whether the inserts, in whole or part, were expressed in the human transcriptome as exons. When an insert aligned with spliced mRNA or expressed sequence tag (EST) sequences mapped at the same locus, we concluded that the transcript variant containing the insert was actually expressed in human cells. Using this procedure, we identified 112 transcript variants that are expressed uniquely in humans.

2 METHODS

2.1 Identification of human-specific genomic inserts

We downloaded the human, chimpanzee and rhesus macaque genome annotation databases, hg18, panTro2 and rheMac2, respectively, from the University of California Santa Cruz (UCSC) Genome Browser database (http://genome.ucsc.edu/) (Rhead *et al.*, 2010) in January 2009. We then collected putative human-only genomic fragments by filtering the top-level alignment gaps from the database tables for human versus chimpanzee genome alignments (hg18.netPanTro2) and for human versus rhesus macaque genome alignments (hg18.netRheMac2). Alignment gaps that met the following conditions were collected: (i) the human-only fragment did not align with its orthologous chromosome from either the chimpanzee or rhesus macaque genomes, (ii) the human unaligned (insert DNA) length was >50 bp, (iii) the chimpanzee or rhesus macaque unaligned length was <10 bp and (iv) the human sequence did not overlap with any duplicated regions listed in the database table hg18.genomicSuperDups. The Y chromosome and the mitochondrial genome were excluded. As the result, we obtained 10 447 human-specific genomic inserts.

2.2 Identification of human-specific transcript variants

Next, we identified the human-only fragments that would be expressed as exons in the human transcriptome by inspecting whether they overlapped with human exonic regions of the human genome. For the human exonic regions, we collected human genomic segments aligned with human RefSeq, mRNA and/or EST sequences. We used the database tables, hg18.refSeqAli, hg18.chrN_mrna and hg18.chrN_intronEst, where N was the chromosome number (from 1 to 22 and X), to deduce human exonic regions. We excluded single-exon mRNAs and single-exon ESTs to avoid possible genomic contamination during cDNA preparation. We removed transcripts mapped to multiple locations in the human genome because these transcripts were likely derived from the repetitive elements or duplicated segments. We also excluded transcripts if the whole transcriptional unit resided within the insert because these transcripts did not interact with nearby genes. We obtained 240 human-specific genomic insert candidates that overlapped with human exonic regions. We then excluded candidate inserts where the orthologous chimpanzee chromosomal region was close to a sequence gap (within 500 bp) or exhibited low quality. For the chimpanzee genome sequence gap and quality data, we used the database tables, panTro2.gap and panTro2.quality, respectively. After these exclusions, we identified 209 regions as novel exonic candidates for human-specific transcript variants.

As a final step, we manually scrutinized the candidate regions to collect highly probable cases of human-specific transcript variants. The genome alignment tracks of the UCSC Genome Browser were analyzed to make sure that the insertion was human specific. We excluded cases showing short or poor alignment between the genome sequences or uncertain orthology for members of multigene families between the two species. Finally, we collected 112 human-specific transcript variants.

2.3 Identification of the insert contribution

The effects of inserts on host genes were determined based on the contribution of the insert to the derived transcript (Supplementary Fig. S1). The effects were categorized as follows: (i) 'promoter', the transcript started within the insert and the splice donor of the first exon was either within the insert or a cryptic donor site was used in the nearby downstream region, (ii) 'terminal exon', the transcript ended within the insert and the splice donor of the last exon was either within the insert or a cryptic acceptor site was used in the nearby upstream region, (iii) 'insertion', the insertion occurred within a host exon and the splice donor and acceptor sites of the exon were located on the outside of the insert, (iv) 'splice donor', a splice donor was located within the insert and a cryptic splice acceptor or a cryptic promoter was present in the nearby upstream region, (v) 'splice acceptor', a splice acceptor was located within the insert and a cryptic splice donor or a cryptic polyadenylation site was present in the nearby downstream region and (vi) 'exon cassette', both splice donor and acceptor sites were located within the insert and hence a part of the insert provided an internal exon.

2.4 Identification of the inserts and classification of the transposable of elements

The original source of the insert was identified in each human variant. The inserts were either transposable elements or unique genomic fragments. The transposable element subfamily classification was conducted using RepeatMasker version open-3.2.9 (http://www.repeatmasker.org/) with the '-s' option and RepBase libraries as of June 4, 2009 (Jurka, 2000). L1 elements were further classified into Ta subfamilies by manually inspecting the diagnostic sequences (Brouha *et al.*, 2003). BLAT searches of the human genome assembly at the UCSC Genome Browser database were used to identify non-repetitive genomic fragments.

2.5 Identification of insertion/deletion polymorphisms

We analyzed some individual human genome assemblies and the NCBI dbSNP database to check whether any of the 112 cases are polymorphic in humans. The hg18 genome fragments spanning the inserts were extracted and BLASTed against the personal genome contigs generated by reference-independent (*de novo*) assembly, including J. Craig Venter (JCV) genome (Levy *et al.*, 2007), and the Chinese YH and Yoruba NA18507 genomes (Li *et al.*, 2010). Absence or presence of the insertion was determined (see Supplementary Table S1 for details).

3 RESULTS AND DISCUSSION

3.1 Identification of human-specific transcript variants

The goal of this analysis was to collect human-specific transcript variants that emerged in the human genome after the human–chimpanzee divergence. It is possible that non-exonic segments or insert elements existing before the human–chimpanzee divergence could acquire a novel splice site and become part of a transcript only in the human genome (Lev-Maor *et al.*, 2003; Sorek *et al.*, 2004). A method of comparing human and non-human primate species transcriptome data itself could be used for the discovery of novel transcript variants generated by this route. However, this method would be very difficult because it would demand extensive transcriptome data for non-human primates.

Instead, we focused on human-specific genomic segments that were inserted after the human–chimpanzee divergence. When a human-specific genomic insert was expressed as a part of the transcriptome in the human cells, we assumed it represented a novel human-specific transcript variant. An example case observed in this study is given in Figure 1. We found a human-specific L1HS

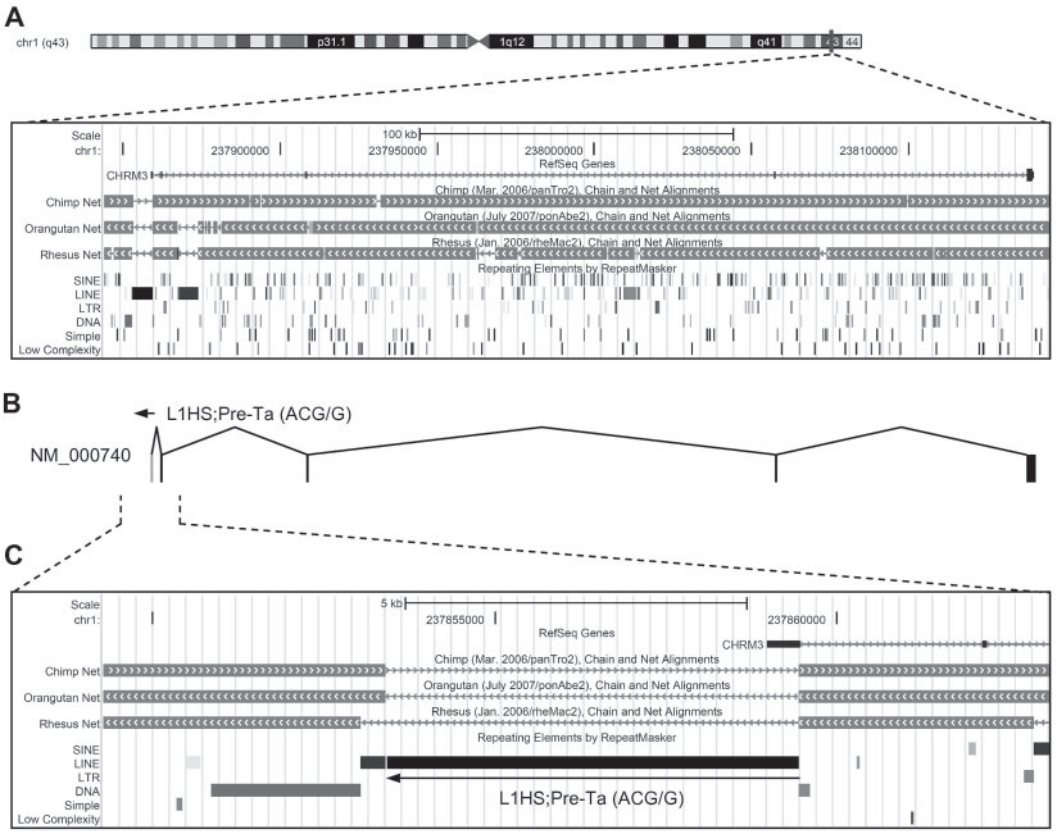

Fig. 1. Human-specific promoter of the *CHRM3* gene. (**A**) The *CHRM3* gene locus from the UCSC Genome Browser is presented. (**B**) Exon organization of the *CHRM3* RefSeq transcript NM_000740. The first exon (gray vertical line) is derived from the inserted L1 element. The black vertical lines are ancestral exons. (**C**) Detailed view of the *CHRM3* human-specific promoter. Direction of *CHRM3* transcription is from left to right. The annotation tracks labeled as 'Chimp Net', 'Orangutan Net' and 'Rhesus Net' depict sequence alignments between the human genome and the chimpanzee, orangutan and rhesus macaque genomes, respectively. Note that the common alignment gaps (thin lines on the 'Net' tracks) in the chimpanzee, the orangutan and the rhesus macaque genomes indicate the corresponding segment exists only in the human genome due to a specific insertion of the L1 element. The transcription start site is located at the 5′ end of the L1 element on the antisense strand, which is often referred to as the L1 ASP.

element insertion in the *CHRM3* gene locus. The RefSeq transcript (accession number NM_000740), an mRNA (AF279779) and three ESTs (CD654050, DA454603 and DA742231) were identified as derived from the antisense promoter (ASP) of the inserted L1 element, indicating that the *CHRM3* gene acquired a human-specific novel promoter.

To collect human-specific genomic inserts, we analyzed the human–chimpanzee and human–rhesus macaque genome alignment data available at the UCSC Genome Browser database (Rhead *et al.*, 2010). Human genomic fragments that did not align with either the chimpanzee or the rhesus macaque genomes were considered as human-specific genomic inserts. By parsing the genome alignment data with stringent conditions, we were able to collect 10 447 highly plausible instances of human-specific genomic insertions >50 bp. To examine whether the human-specific inserts were expressed as transcripts in the human cell, we utilized human transcriptome sequence data, including RefSeqs, mRNAs and ESTs. We excluded unspliced mRNA or EST sequence data to avoid possible genomic contamination during cDNA preparation. When a human-specific genomic insert aligned with a transcript, it was considered a human-specific transcript variant. After manual inspection of the candidates, we collected 112 human-specific transcript variants. The original source of each insert was identified, and the contribution of the insert to the derived transcript was determined (Supplementary Table S1).

3.2 Characteristics of human-specific transcripts

Out of the total 112 cases, 74 insertions were associated with known genes (Supplementary Table S1). The insert regions provided promoters, terminal exons, inserted segments or splice sites to the host genes, generating novel human-specific transcript variants of the genes. The remaining 38 cases were located in as of yet unannotated genomic loci. It has been reported that most human chromosomal regions that were thought not to harbor protein-coding genes were indeed actively transcribed (Johnson *et al.*, 2005; Wong *et al.*, 2001). A previous report has shown that intergenic transcripts show patterns of tissue-specific conservation of their expression, and about half of the expression differences in tissues between humans and chimpanzees are due to intergenic transcripts (Khaitovich *et al.*, 2006). Human-specific inserts in intergenic regions may drive novel cryptic transcription or reshape pre-existing

Table 1. Contribution of the inserts to the derived transcripts

Contribution	L1	Alu	SVA	HERV	genomic	Total
Promoter	35	4	12	1	0	52
Terminal exon	5	10	9	2	5	31
Insertion	1	11	1	0	2	15
Splice donor	3	4	2	0	0	9
Splice acceptor	0	3	0	0	1	4
Exon cassette	0	0	0	1	0	1

intergenic transcripts. They may also give rise to novel functional RNAs, such as large non-coding RNAs or serve as a molecular source for si- or miRNAs in humans. The novel inserts may promote antisense transcription in human genes (Conley *et al.*, 2008). Out of the total 112 cases identified in this study, 13 insertions were involved in generation or modification of antisense transcripts of known genes, which were described elsewhere in detail (Kim and Hahn, 2010).

The majority (104 out of 112 cases) of the original insert sources were transposable elements or retroviruses including L1 (44 cases), Alu (32 cases), SVA (24 cases) and HERVs (4 cases) (Supplementary Table S2). RepeatMasker classified the L1 elements observed in this study to L1HS, L1PA2 or L1PA3 subfamilies. Manual classification of L1 elements into Ta subfamilies revealed that the most abundant groups were the L1HS;Pre-Ta (ACG/G) and L1PA2;non-Ta subfamilies. These two L1 types have been shown to be the most common insertion sequence amplified during the period of ape evolution (Lee *et al.*, 2007; Mathews *et al.*, 2003), and are differentially present in humans and chimpanzees (Mills *et al.*, 2007). All Alu repeat elements observed in this study belonged to either the AluY or AluS subfamilies, which are known to be active in human cells (Mills *et al.*, 2007). SVA is a composite retrotransposon consisting of SINE-R, VNTR and Alu-like elements (Ostertag *et al.*, 2003). SVA elements are hominid specific and are actively mobilized in the human genome (Hancks *et al.*, 2009; Wang *et al.*, 2005). All SVA elements belong to one of four subfamilies; SVA/C, SVA/D, SVA/E and SVA/F. These subfamilies show differential insertion patterns between humans and chimpanzees, with SVA/D being most differentially present (Mills *et al.*, 2007). In this study, SVA/D insertion elements were found to be the most abundant SVA element that generated human-specific transcript variants. The four HERV elements involved in novel transcript formation were classified as HERVK, HERVK9 or HERVH. There is published evidence that HERVs can mobilize in the human genome, especially HERVK (Mills *et al.*, 2007).

We also found eight cases involved with non-repetitive genomic DNAs; three cases of the direct mobilization of non-genic (intron or intergenic) genomic segments, three cases of processed cDNA integration or retrogenes probably by exploiting LINE retrotransposition machinery, and two cases of tandem duplication caused by DNA replication slippage.

We determined the insert contribution to the derived transcripts on the basis of whether the insert contained a transcription start site, polyadenylation site or splice site (Table 1). The insert provided a promoter in 52 cases, a terminal exon in 31, an insertion in 15, a splice donor in 9, a splice acceptor in 4 and an exon cassette in 1.

3.3 Novel promoters

We identified 52 cases in which the insert may be responsible for driving transcription of the proximal genomic region (Table 1). Of them, 37 insertion events provided novel promoters for the annotated human genes. We assumed the insert to have promoter activity when the 5′ end of the transcript was positioned within the insert. We did not rule out the possibility that the cDNA product could be simply truncated in the insert. Most of the new promoters were involved with L1 elements; 35 cases were identified in this study. Some SVA and Alu elements were also identified to promote transcription. There is evidence in the literature that an Alu element can induce transcription of an adjacent protein-coding gene. For example, the first exon of a *p75TNFR* gene transcript variant was derived from an AluJo element (Singer *et al.*, 2004).

Some of the human-specific transcript variants identified in this study were assigned NCBI RefSeq transcripts: NM_000740, the only RefSeq record of *CHRM3*; NR_002140, a non-coding transcript of *OR6W1P*; and NM_001024647, one of five RefSeqs of the *RAB3IP* gene. The *CHRM3* gene was previously reported as an example of a L1 ASP-promoted transcription of cellular genes and the gene-breaking model (Speek, 2001; Wheelan *et al.*, 2005) although the L1 does not actually break the coding region of the *CHRM3* gene. The entire protein-coding region of *CHRM3* is in the last exon, and hence the L1-derived transcripts produce full-length proteins. Acquisition of a human-specific promoter may increase the expression level of full-length *CHRM3* in human cells. In the case of *RAB3IP*, the human-specific transcript originated from the ASP of a L1 element integrated in an intron, which is an excellent example of the gene-breaking model (Supplementary Fig. 2A) (Wheelan *et al.*, 2005). Other cases that complied with the gene-breaking model identified in this study included *CD96* and *NRXN3*.

It is known that L1 elements drive transcription of adjacent cellular genes when the L1 element is inserted into the 5′ upstream region of the gene (Speek, 2001). Transcripts starting from the L1 element could splice to the first or second exon of the host gene and would encode almost or fully functional full-length proteins. This is in contrast to the gene-breaking model in which L1 integration occurs within an intron and the novel downstream products usually lack a large portion of the host protein N-terminus. Notable cases of gaining new 5′ upstream promoters include *CHRM3*, *MAL2*, *TIGIT* and *RGS6* genes. The human-specific transcript variants of the *MAL2* and *RGS6* genes were derived from the ASP of the L1 elements. Interestingly, the novel *TIGIT* gene transcript variant was derived from the L1PA3;non-Ta element sense strand (Supplementary Fig. 2B). A splice donor in the 5′ untranslated region (UTR) of the L1 element was used to splice to the second exon of the *TIGIT* gene. The derived transcript would encode an almost full-length protein using a methionine codon in the second exon as a translation start codon.

We also found 12 cases where SVA elements appeared to drive transcription. An SVA element integrated in the 5′ upstream region provided a novel promoter for the *TBPL2* gene (Supplementary Fig. 2C). In the *WDR66* gene, the SVA/D element inserted in intron 19 seemed to promote the transcription of a human-specific

transcript variant that spliced to the last three exons of the *WDR66* gene.

3.4 Novel polyadenylation sites

We identified a total of 31 cases in which the newly inserted DNA introduced a polyadenylation site (Table 1). The splice donor site for the new terminal exon could be either within the insert or in the nearby upstream region possibly activated upon insertion. When the newly inserted polyadenylation site was very close to the upstream exon and induced premature cleavage of the transcript, the upstream exon splice donor would not be functional, retaining the following intronic segment. If the insert were situated in an intron, it would induce premature termination of host gene transcription, which is known as exon- or gene-trapping. Most of the mobile elements, including L1, Alu and SVA are known to cause exon-trapping (Chen *et al.*, 2009; Hancks *et al.*, 2009; Wheelan *et al.*, 2005).

The *LEPR* gene, which encodes the leptin receptor protein, contains a human-specific SVA element which serves as a terminal exon, encoding a variant with a different C-terminus (Damert *et al.*, 2004). We successfully identified this human-specific *LEPR* transcript variant in this study. The novel *ATP9B* gene transcript variant terminates within the L1HS element situated in intron 4 (Supplementary Fig. 3A). The L1 contains a splice acceptor site and a polyadenylation site, truncating the full-length *ATP9B* transcript by exon-trapping. Interestingly, the L1 element sense strand was incorporated into the transcript in contrast to the gene-breaking model where the polyadenylation site of the L1 element antisense strand was used (Wheelan *et al.*, 2005). In the case of the *FUT8* gene, the human-specific Alu in an intron shortened the transcript, a typical example of Alu-induced exon-trapping (Supplementary Fig. 3B).

Alu elements inserted in the 3′ flanking region close to a gene, as in the *RAB3B* and *KIAA0101* genes, appear to provide a novel polyadenylation site and extend the 3′ UTR of the host gene. The RefSeq transcript NM_014736 of the *KIAA0101* gene terminates within the human-specific AluYb8 element (Supplementary Fig. 3C). However, inspection of many mRNAs and ESTs of the gene revealed that there is an upstream polyadenylation site before the Alu. Most of the transcripts appear to use the upstream signal and only a small portion of the transcripts end within the Alu insert. It is likely that the extended read-through transcripts miss the original polyadenylation signal and use the alternative Alu polyadenylation signal instead.

3.5 Insertions within exons

Insertion within an exon results in a permanent change in the structure of the host gene (Knebelmann *et al.*, 1995; Makałowski *et al.*, 1994; Mitchell *et al.*, 1991). We identified 15 cases where the insert was situated within a host gene exon. Most of the insertions in annotated genes were found in the 3′ UTR. Examples include an Alu insertion in the *RAB21*, *KIAA0319L*, *UTP11L*, *LRRC58*, *SLC13A1*, *GNB5* and *DSG3* genes, and L1 in the *OPHN1* gene.

The *RAB21* gene case is notable in that the Alu insertion occurred within a highly conserved segment (Fig. 2). Although the region is untranslated, the nucleotide sequences were highly conserved in tetrapods, including chicken (*Gallus gallus*), lizard (*Anolis carolinensis*) and frog (*Xenopus tropicalis*). Alteration of the 3′ UTR sequence and structure have been implicated in human cancers, disease and distinct phenotypes (Abelson *et al.*, 2005; Mayr and Bartel, 2009). The evolutionarily conserved regions in the 3′ UTRs may harbor regulatory elements crucial for mRNA molecular function. Those regions may harbor post-transcriptional regulatory elements, which interact with *trans*-acting factors including miRNAs or RNA-binding proteins. The RAB21 protein is reported to control integrin trafficking, which is necessary for cell cytokinesis (Pellinen *et al.*, 2008). Alu disruption of the conserved *RAB21* 3′ UTR may have altered the post-transcriptional regulation of the human *RAB21* gene.

Human-specific disruption of conserved 3′ UTR elements by insertion would substantially modify regulation of mRNAs in the human cell. Altered mRNA regulation such as turnover rate, degradation or translatability may change the abundance of encoded proteins and subsequently develop novel human-specific phenotypes.

3.6 Novel splice donor and/or acceptor sites

A newly inserted segment could bring a splice acceptor and/or donor site within the insert. We identified 14 cases where a splice donor site (9 cases), a splice acceptor site (4 cases) or both (1 case) was positioned within the insert (Table 1). For example, the novel human-specific transcript variants of the *TRIM60* and *ALOX5* genes utilized a splice donor site in the newly inserted Alu and SVA elements, respectively.

It is common for an Alu element to introduce a splice acceptor site to the host gene transcript (Lev-Maor *et al.*, 2003). An example identified in this study was the *COPS7B* gene. A splice acceptor site within the Alu element and a cryptic donor site in the nearby downstream region were utilized to add an exon cassette to the *COPS7B* gene; 10 ESTs contain an Alu-derived exon. The Alu exon carries multiple inframe stop codons, so the novel transcript variant cannot encode a full-length COPS7B protein. Transcripts containing a premature stop codon (PTC) would be subject to nonsense-mediated mRNA decay (Brogna and Wen, 2009), minimizing the production of truncated proteins. Since a portion of the *COPS7B* transcripts would contain the PTC-containing Alu exon, the net effect would be a reduction of full-length COPS7B protein in the cell.

We identified an intriguing case of exon-trapping in which a mobile element was not used; instead, genomic DNA was involved. There is a human-specific insert in an intron of the *TBC1D8B* gene on X chromosome; the splice acceptor is within the insert and the shorter *TBC1D8B* (NM_198881) variant ends in the proximal downstream region of the insert (Fig. 3). Interestingly, the insert matched a portion of intron 6 of the *EBF1* gene on chromosome 5 in the reverse direction. The insert showed typical features for retrotransposition; a poly(A) tract and a direct repeat caused by target site duplication. Therefore, it is very likely that a reverse-transcribed product of a portion of the primary *EBF1* transcript was integrated into the *TBC1D8B* intron by exploiting LINE retrotransposition machinery.

3.7 Insertion/deletion polymorphisms

Analysis of individual human genome assemblies, including a Caucasian (JCV), an Asian (YH) and an African (NA18507) genome, and the NCBI dbSNP database revealed that at least seven loci out of 112 are polymorphic among humans (Supplementary Table S1). Six of these were known polymorphisms

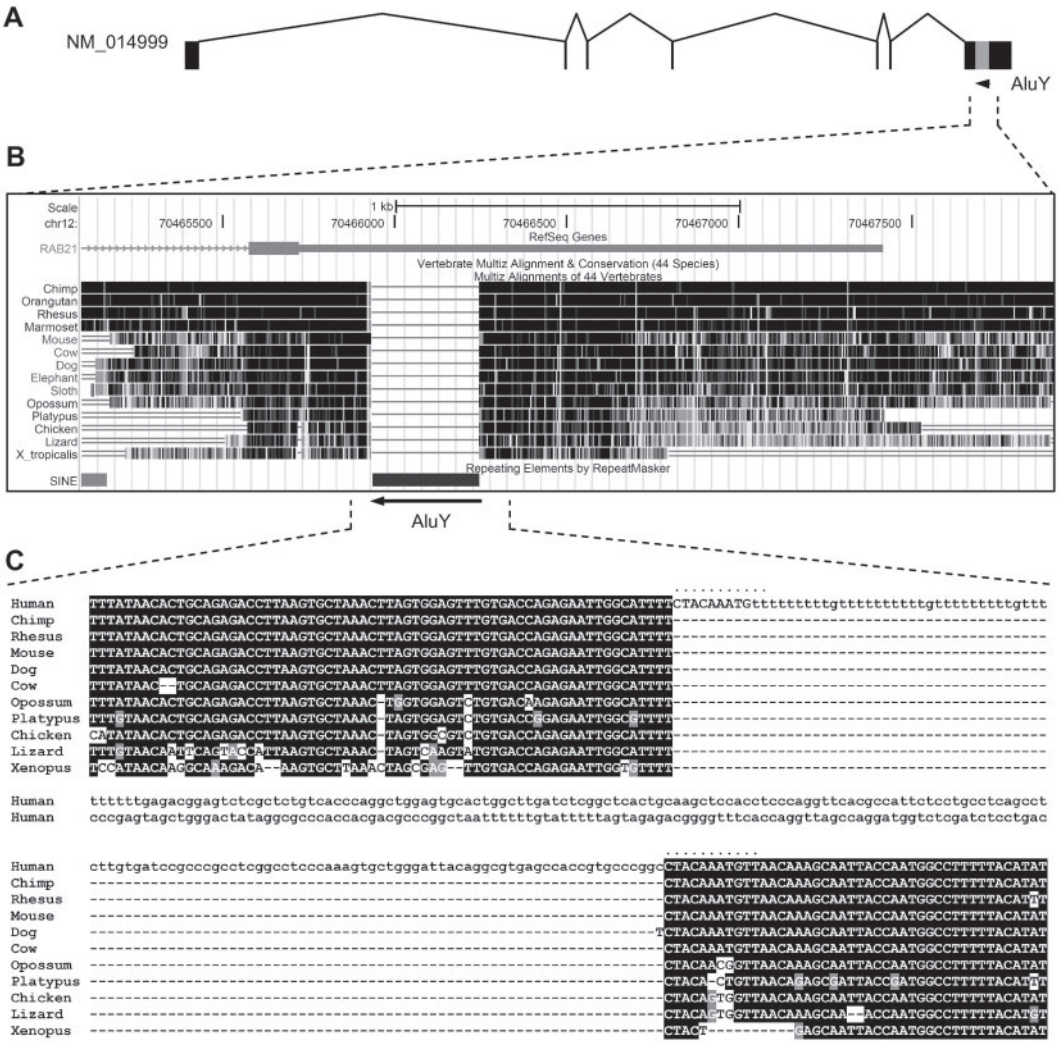

Fig. 2. Alu insertion in the conserved *RAB21* gene 3′ UTR. (**A**) The top line shows the full genomic structure of the human *RAB21* gene. The human-specific AluY element is marked in grey. (**B**) The genomic region surrounding the Alu insertion (an arrow) is depicted. The line labeled 'RAB21' indicates the RefSeq NM_014999; the thin, thick, and half thick lines represent the intron, coding exon, and 3′ UTR, respectively. The tracks labeled 'Multiz Alignments of 44 Vertebrates' show sequence conservation in tetrapod species. (**C**) Multiple sequence alignment of the genomic region surrounding the Alu element is shown. Note the high 3′ UTR sequence conservation, even in non-mammalian species including chicken, lizard, and X_tropicalis (*Xenopus tropicalis*). The Alu element identified by RepeatMasker program is in lowercase. The putative target site duplication is marked with dots. Genome sequences were retrieved from the UCSC Genome Browser database and the assembly versions are hg18 (human), panTro2 (chimpanzee), rheMac2 (rhesus macaque), mm9 (mouse), canFam2 (dog), bosTau4 (cow), monDom5 (opossum), ornAna1 (platypus), galGal3 (chicken), anoCar1 (lizard), and xenTro2 (Xenopus, *X. tropicalis*). The multiple sequence alignment was generated using MUSCLE (http://www.drive5.com/muscle/) and decorated by using BOXSHADE (http://www.ch.embnet.org/software/BOX_form.html).

recorded in dbSNP. For example, the L1HS;Pre-Ta (ACG/G) element insertion at position 132946863-132952889 on chromosome 5 of h18 genome is absent in YH and NA18507 genomes (dbSNP accession number rs66911382). These polymorphisms may potentially contribute to transcriptome and phenotypic diversity among humans.

4 CONCLUSIONS

In this analysis, we identified 112 cases in which human-specific genomic inserts mediated gene evolution in the human genome. The human-specific inserts gave rise to novel promoters, polyadenylation sites, insert segments and splice sites. Novel promoters generated intergenic transcripts or drove transcription of adjacent cellular genes. Novel polyadenylation sites and splice sites induced exon-trapping or insertion of exon cassettes. The new transcript variants may produce proteins with different N- or C-termini. Generation of prematurely terminated variants could lead to decreases in abundance of full-length functional proteins. Insertions within exons could disrupt evolutionarily conserved elements, resulting in substantial changes in the regulation of the genes involved. An altered cellular level of functional proteins

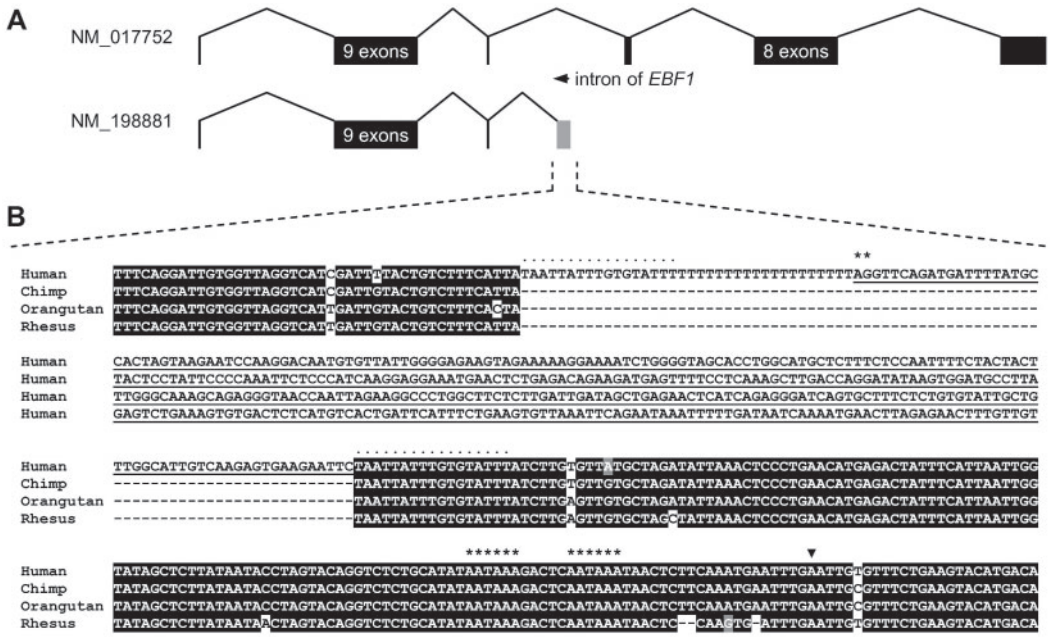

Fig. 3. A novel *TBC1D8B* gene terminal exon. (**A**) The last exon of RefSeq NM_198881 contains a part of the *EBF1* gene intron on chromosome 5, which was retrotransposed to an intron of the *TBC1D8B* gene. A splice acceptor is positioned within the insert and a cryptic polyadenylation signal in the downstream region. (**B**) A multiple sequence alignment of the nucleotide sequences surrounding the human-specific insert from human, chimpanzee, orangutan, and rhesus macaque genomes, is shown. Sequences derived from the *EBF1* gene intron are underlined. The target site duplication is indicated with dots. Note the poly(T) sequence, which is the poly(A) tail of the insert in the reverse direction. The novel splice acceptor site (AG), two putative polyadenylation signals (AATAAA), and polyadenylation site for the *TBC1D8B* transcript NM_198881 are marked by two asterisks, six asterisks, and an upside down triangle, respectively. Genome sequences were obtained from the UCSC Genome Browser database: hg18 (human), panTro2 (chimpanzee), ponAbe2 (orangutan) and rheMac2 (rhesus macaque).

may be accompanied by these gene modifications. We propose that a substantial portion of these human-specific modifications may be directly associated with the development of various human characteristics. Functional investigation of these human-specific transcripts may provide clues as to how humans acquired certain traits after divergence from the rest of the hominoids.

Funding: The Korea Research Foundation (MOEHRD, Basic Research Promotion Fund KRF-2008-331-C00263) and the National Research Foundation of Korea (2009-0071595).

Conflict of Interest: none declared.

REFERENCES

Abelson,J.F. *et al.* (2005) Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science*, **310**, 317–320.

Brogna,S. and Wen,J. (2009) Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat. Struct. Mol. Biol.*, **16**, 107–113.

Brouha,B. *et al.* (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl Acad. Sci. USA*, **100**, 5280–5285.

Chen,C. *et al.* (2009) Using Alu elements as polyadenylation sites: a case of retroposon exaptation. *Mol. Biol. Evol.*, **26**, 327–334.

Chou,H.H. *et al.* (1998) A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. *Proc. Natl Acad. Sci. USA*, **95**, 11751–11756.

Conley,A.B. *et al.* (2008) Human *cis* natural antisense transcripts initiated by transposable elements. *Trends Genet.*, **24**, 53–56.

Damert,A. *et al.* (2004) Leptin receptor isoform 219.1: an example of protein evolution by LINE-1-mediated human-specific retrotransposition of a coding SVA element. *Mol. Biol. Evol.*, **21**, 647–651.

Deininger,P.L. *et al.* (2003) Mobile elements and mammalian genome evolution. *Curr. Opin. Genet. Dev.*, **13**, 651–658.

Dewannieux,M. *et al.* (2006) Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res.*, **16**, 1548–1556.

Enard,W. *et al.* (2002) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, **418**, 869–872.

Hahn,Y. and Lee,B. (2005) Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. *Bioinformatics*, **21** (Suppl. 1), i186–i194.

Hahn,Y. and Lee,B. (2006) Human-specific nonsense mutations identified by genome sequence comparisons. *Hum. Genet.*, **119**, 169–178.

Hahn,Y. *et al.* (2007) Inactivation of MOXD2 and S100A15A by exon deletion during human evolution. *Mol. Biol. Evol.*, **24**, 2203–2212.

Hancks,D.C. *et al.* (2009) Exon-trapping mediated by the human retrotransposon SVA. *Genome Res.*, **19**, 1983–1991.

Hazkani-Covo,E. and Graur,D. (2007) A comparative analysis of numt evolution in human and chimpanzee. *Mol. Biol. Evol.*, **24**, 13–18.

Johnson,J.M. *et al.* (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.*, **21**, 93–102.

Jurka,J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.*, **16**, 418–420.

Khaitovich,P. *et al.* (2006) Functionality of intergenic transcription: an evolutionary comparison. *PLoS Genet.*, **2**, e171.

Kim,D.S. and Hahn,Y. (2010) Human-specific antisense transcripts induced by the insertion of transposable element. *Int. J. Mol. Med.*, **26**, 151–157.

Knebelmann,B. *et al.* (1995) Splice-mediated insertion of an Alu sequence in the COL4A3 mRNA causing autosomal recessive Alport syndrome. *Hum. Mol. Genet.*, **4**, 675–679.

Knowles,D.G. and McLysaght,A. (2009) Recent de novo origin of human protein-coding genes. *Genome Res.*, **19**, 1752–1759.

Konopka,G. *et al.* (2009) Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature*, **462**, 213–217.

Krull,M. *et al.* (2005) Alu-SINE exonization: en route to protein-coding function. *Mol. Biol. Evol.*, **22**, 1702–1711.
Lee,J. *et al.* (2007) Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene*, **390**, 18–27.
Lev-Maor,G. *et al.* (2003) The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science*, **300**, 1288–1291.
Levy,S. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
Li,R. *et al.* (2010) Building the sequence map of the human pan-genome. *Nat. Biotechnol.*, **28**, 57–63.
Lorenc,A. and Makałowski,W. (2003) Transposable elements and vertebrate protein diversity. *Genetica*, **118**, 183–191.
Makałowski,W. *et al.* (1994) Alu sequences in the coding regions of mRNA: a source of protein variability. *Trends Genet.*, **10**, 188–193.
Marques,A.C. *et al.* (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.*, **3**, e357.
Mathews,L.M. *et al.* (2003) Large differences between LINE-1 amplification rates in the human and chimpanzee lineages. *Am. J. Hum. Genet.*, **72**, 739–748.
Mayr,C. and Bartel,D.P. (2009) Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684.
Mills,R.E. *et al.* (2007) Which transposable elements are active in the human genome? *Trends Genet.*, **23**, 183–191.
Mitchell,G.A. *et al.* (1991) Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. *Proc. Natl Acad. Sci. USA*, **88**, 815–819.
Nekrutenko,A. and Li,W.H. (2001) Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.*, **17**, 619–621.
Noonan,J.P. (2009) Regulatory DNAs and the evolution of human development. *Curr. Opin. Genet. Dev.*, **19**, 557–564.
Olson,M.V. (1999) When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.*, **64**, 18–23.
Ostertag,E.M. *et al.* (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am. J. Hum. Genet.*, **73**, 1444–1451.
Pellinen,T. *et al.* (2008) Integrin trafficking regulated by Rab21 is necessary for cytokinesis. *Dev. Cell*, **15**, 371–385.
Pollard,K.S. *et al.* (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, **443**, 167–172.
Rhead,B. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
Singer,S.S. *et al.* (2004) From “junk” to gene: curriculum vitae of a primate receptor isoform gene. *J. Mol. Biol.*, **341**, 883–886.
Sorek,R. *et al.* (2002) Alu-containing exons are alternatively spliced. *Genome Res.*, **12**, 1060–1067.
Sorek,R. *et al.* (2004) Minimal conditions for exonization of intronic sequences: 5′ splice site formation in alu exons. *Mol. Cell*, **14**, 221–231.
Speek,M. (2001) Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol. Cell. Biol.*, **21**, 1973–1985.
Stedman,H.H. *et al.* (2004) Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature*, **428**, 415–418.
Wang,H. *et al.* (2005) SVA elements: a hominid-specific retroposon family. *J. Mol. Biol.*, **354**, 994–1007.
Wheelan,S.J. *et al.* (2005) Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome Res.*, **15**, 1073–1078.
Winter,H. *et al.* (2001) Human type I hair keratin pseudogene phihHaA has functional orthologs in the chimpanzee and gorilla: evidence for recent inactivation of the human gene after the Pan-Homo divergence. *Hum. Genet.*, **108**, 37–42.
Wong,G.K. *et al.* (2001) Most of the human genome is transcribed. *Genome Res.*, **11**, 1975–1977.