Alu-mediated inactivation of the human CMP-*N*-acetylneuraminic acid hydroxylase gene

Toshiyuki Hayakawa*, Yoko Satta*, Pascal Gagneux†, Ajit Varki†, and Naoyuki Takahata*‡

*Department of Biosystems Science, Graduate University for Advanced Studies (Sokendai), Hayama, Kanagawa 240-0193, Japan; and †Glycobiology Research and Training Center, Departments of Medicine and Cellular and Molecular Medicine, University of California at San Diego, La Jolla, CA 92093-0687

Edited by Henry C. Harpending, University of Utah, Salt Lake City, UT, July 17, 2001 (received for review May 30, 2001)

Inactivation of the CMP-*N*-acetylneuraminic acid hydroxylase gene has provided an example of human-specific genomic mutation that results in a widespread biochemical difference between human and nonhuman primates. We have found that, although a region containing a 92-bp exon and an *Alu*Sq element in the hydroxylase gene is intact in all nonhuman primates examined, the same region in the human genome is replaced by an *Alu*Y element that was disseminated at least one million years ago. We propose a mechanistic model for this *Alu*-mediated replacement event, which deleted the 92-bp exon and thus inactivated the human hydroxylase gene. It is suggested that *Alu* elements have played potentially important roles in genotypic and phenotypic evolution in the hominid lineage.

Human-specific traits (upright walking, language ability, etc.) have evolved over time through the emergence and extinction of several hominid species during human evolution (1). The acquisition of such traits accompanies marked changes in morphology and physiology. Humans are most closely related to the African great apes: the chimpanzee (*Pan troglodytes*), the bonobo (*Pan paniscus*), and the gorilla (*Gorilla gorilla*) (2, 3). Research to determine which genes differentiate humans from the great apes has been undertaken (reviewed in ref. 4), and comparative genomic analysis among primates is now underway. Three general classes of genetic differences have been proposed as factors separating humans from the great apes: chromosomal differences, small sequence differences that change gene expression, and biochemical changes resulting from gene inactivation (5). Thus far, three genes are known to have been altered in human-specific manners. One is loss of exon 34 in the tropoelastin gene, which was possibly facilitated by *Alu*-mediated recombination events (6). However, because exon 35 has already been deleted in catarrhines, loss of an additional exon in the hominid lineage may be of secondary significance. Another is a recent introduction of a premature termination codon into a member of the human type I hair keratin gene cluster (7). Unlike these gene changes, inactivation of the CMP-*N*-acetylneuraminic acid hydroxylase gene is unique in that it is a single gene and inactivated in the hominid lineage only (8–10).

CMP-*N*-acetylneuraminic acid (CMP-Neu5Ac) is a nucleotide sugar donor of Neu5Ac, the most common sialic acid in humans. The CMP-Neu5Ac hydroxylase converts CMP-Neu5Ac to the hydroxylated form, CMP-*N*-glycolylneuraminic acid (CMP-Neu5Gc) (11–14). Sialic acids such as Neu5Ac and Neu5Gc belong to a family of acidic sugars, and they are typically found on the cell surface in all mammals (15–17). These nine-carbon sugars function as ligands in recognition systems mediated by sialic acid-binding lectins, such as CD22, myelin-associated glycoprotein, sialoadhesin, and influenza A virus hemagglutinin (reviewed in ref. 17). Some of these lectins can discriminate between Neu5Ac and Neu5Gc (summarized in table 3 of ref. 18), and expression of the hydroxylase gene thus contributes to regulation of cell–cell interaction mediated by lectins of both endogenous and exogenous origin. It is therefore reasonable to think that the inactivation of the CMP-Neu5Ac hydroxylase gene in humans caused significant changes in several lectin-mediated interactions and possibly contributed to unique features of human evolution (discussed in refs. 8, 10, and 18)

It is known that the human-specific inactivation of the hydroxylase gene resulted from the deletion of a 92-bp exon (8, 9) and a subsequent frameshift in the coding sequence (8). The 92-bp exon encodes a part of the Reiske iron-sulfur-binding region that is essential for the enzyme's activity (8, 9, 19). The truncated hydroxylase therefore cannot convert CMP-Neu5Ac to CMP-Neu5Gc. The same deletion is found in all humans thus far examined, but not in the African apes, so that the 92-bp exon must have been deleted in the early evolution of the hominid lineage (8, 10). To gain insight into the genomic event that produced the human-specific inactivation of the hydroxylase gene, we have performed a comparative genomic analysis of the hydroxylase gene among six hominoids and two cercopithecoids: human, chimpanzee (*Pan troglodytes*), bonobo (*Pan paniscus*), gorilla (*G. gorilla*), orangutan (*Pongo pygmaeus*), gibbon (*Hylobates lar*), baboon (*Papio anubis*), and rhesus monkey (*Macaca mulatta*). We then propose a model to explain how *Alu* insertion can delete an exon and thus inactivate a gene.

Materials and Methods

DNA Samples. Chimpanzee, gorilla, orangutan, and gibbon genomic DNAs were generous gifts from Shintaroh Ueda (University of Tokyo) and Colm O'hUigin (Max Planck Institute, Tübingen, Germany). Human samples were from volunteers in the Varki laboratory or provided by Steven Warren's laboratory (Emory University School of Medicine, Atlanta). Additional great ape samples were from Epstein–Barr virus-transfected lymphoblastoid cell lines obtained from Peter Parham (Stanford University School of Medicine, Stanford, CA). Baboon DNA was kindly provided by Jeffrey Rogers (Southwest Foundation for Biomedical Research Genetics, San Antonio, TX). Rhesus monkey genomic DNA was purchased from CLONTECH.

PCR Products of Chimpanzee Genomic DNA. By genomic PCR, 10 fragments covering the ≈23-kb region, including the 92-bp exon, were obtained. The PCR primers were designed on the basis of the intron sequence of human CMP-Neu5Ac hydroxylase (GenBank accession no. AB009668). Fragment 1 was generated by using primers CH-18 (5′-TCGCAATAAGAGCACTG-

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: Neu5Ac, *N*-acetylneuraminic acid; Neu5Gc, *N*-glycolylneuraminic acid; sah*Alu*Sq, sialic acid hydroxylase *Alu*Sq; sah*Alu*Y, sialic acid hydroxylase *Alu*Y; ms*Alu*Y, most similar *Alu*Y; E-A region, exon to *Alu*Sq region.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database [accession nos. AB060157 (human genomic region around the 92-bp exon of CMP-Neu5Ac hydroxylase), AB060158 (chimpanzee genomic region around the 92-bp exon of CMP-Neu5Ac hydroxylase), AB060159 (gorilla genomic region around the 92-bp exon of CMP-Neu5Ac hydroxylase), and AB060160 (rhesus monkey genomic region around the 92-bp exon of CMP-Neu5Ac hydroxylase)].

‡To whom reprint requests should be addressed. E-mail: takahata@soken.ac.jp.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

GCAAAGAC-3′) and CH-25 (5′-ACAAACCAGAAAGC-CCAAGCATGTC-3′). Fragment 2 was generated with primers CH-32 (5′-ACATGCTTGGGCTTTCTGGTTTGTC-3′) and CH-28 (5′-GCTAAGAGGGGAGGACTAATGTGTC-3′). Fragment 3 was generated by using primers CH-9 (5′-TGACACATTAGTCCTCCCCTCTTAG-3′) and CH-13 (5′-CAAATGTTCCCTTCGTGGCAGTGTC-3′). Fragment 4 was generated by using primers CH-8 (5′-CCCTCTTAGCTCTCCT-GCCCATGAG-3′) and CH-12 (5′-GAGGGAGGACAG-CAACCACCAGAAC-3′). Fragment 5 was generated by using primers CH-34 (5′-TCTGGTGGTTGCTGTCCTCCCTCT-C-3′) and CH-36 (5′-AAGCAGGAACCAGACAAGCAGTT-TC-3′). Fragment 6 was generated by using primers CH-15 (5′-CTGCTTGTCTGGTTCCTGCTTTTAG-3′) and CH-19 (5′-TAAGTCCCAAGGGTTAGGAGGATTC-3′). Fragment 7 was generated by using primers CH-18 and CH-84 (5′-AGAAGCAAGAGCAGGATGGAGTCAG-3′). Fragment 8 was generated by using primers CH-92 (5′-GCAGAGGGTG-CAAGAGAAAGGAGAG-3′) and CH-53 (5′-CTAAAATC-CTTGACCCCTAGAATAG-3′). Fragment 9 was generated by using primers CH-10 (5′-TGTGTTGCCAGCATTCTC-CCAGTTC-3′) and CH-38 (5′-ACCATATAGCCCAGCAAT-TCCATTC-3′). Fragment 10 was generated by using primers CH-44 (5′-GTCTATCCTTCTGCCAGTTCCACAC-3′) and CH-106 (5′-AAGAAGGAAACCACATCATCATCTC-3′). The genomic PCR was performed with 20 pmol of each primer and 30 ng of chimpanzee genomic DNA in a total volume of 50 μl containing 200 μM dNTPs and 2.5 units of Ex*Taq* DNA polymerase (TaKaRa) in a TaKaRa Ex*Taq* buffer containing 2 mM $MgCl_2$. A RoboCycler Gradient 96 (Stratagene) was used to produce the following conditions: denaturation at 95°C for 5 min followed by 30 amplification cycles of 95°C for 1 min; 60–62°C for 1 min; 69°C for 7 min; and extension at 69°C for 10 min.

PCR Products of Gorilla Genomic DNA. The primers CH-54 (5′-CATGGTTCTGCCAATTTTCCCTTTC-3′) and CH-95 (5′-ACACACATGCCCACAACCTGATCTG-3′) were used for genomic PCR. PCR was performed as described in the second section.

PCR Products of Rhesus Monkey Genomic DNA. The primers MSA-1 (5′-GTCTGTTAGATGCACAAAGCATAAC-3′), MSA-3 (5′-GGTTGATATACTTCATGGTGCTCAC-3′), and CH-96 (5′-AGCTCAGCTCCCTTAACAGGTAATC-3′) were newly designed on the basis of the cDNA and genomic sequences of the human, chimpanzee, and rhesus monkey (GenBank accession nos. AB009668, AF074481, and AB013814). In addition to these primers, CH-9, CH-54, and CH-95 were used for genomic PCR. Twenty picomoles of each primer was used to amplify 100 ng of the rhesus monkey genomic DNA (CLONTECH) in a TaKaRa ExTaq buffer containing 2 mM $MgCl_2$. PCR was performed as described in the section on chimpanzee DNA.

PCR Products of Human, Bonobo, Orangutan, Gibbon, and Baboon Genomic DNAs. The primers CH-114 (5′-TGGGAAATCATT-AGGCATCCACCTG-3′) and CH-148 (5′-TCTTTATTCT-GCTGTCTCTGTTCTC-3′) were used for genomic PCR. The PCR conditions were as follows: denaturation at 95°C for 5 min followed by 30 cycles of 95°C for 1 min; 60°C for 1 min; 69°C for 1 min; and extension at 69°C for 10 min.

Sequencing of Genomic PCR Products. The PCR products were purified by using the QIAquick PCR Purification Kit (Qiagen, Chatsworth, CA) and sequenced directly with an ABI Prism BigDye Terminator Cycle Sequencing FS Ready Reaction Kit (Applied Biosystems). Sequencing primers were produced based on the human intronic sequence (GenBank accession no. AB009668). With use of a GeneAmp PCR system 9600 (Applied Biosystems), cycle sequencing reaction was performed according to the manufacture's instructions. Each reaction sample was analyzed on an ABI Prism 377 fluorescent automated DNA sequencer (Applied Biosystems).

Comparative Genomic Analysis. DNASIS software (Hitachi, Tokyo) was used for comparative analysis. Repetitive elements on the genomic sequence of each species were detected by using the REPEATMASKER program at the University of Washington Genomic Center web site.

Analysis of Human *Alu*Y Element. The target human *Alu*Y element was picked up by QBLASTQ search with the NR and HTGS databases at the National Center for Biotechnology Information web site. The primers 0Y-1 (5′-GACGATGCTGAAAAGAGCTGTTTG-3′) and 0Y-2 (5′-CCCTTAGCCCTCAGAAAGATACAC-3′) were de-signed on the basis of the flanking sequences of selected *Alu*Y element (GenBank accession no. AC005692). Twenty picomoles of 0Y-1 and 20 pmol of 0Y-2 were used to amplify each great ape genomic DNA in a 50-μl reaction with 200 μM dNTPs and a TaKaRa Ex*Taq* buffer containing 2 mM $MgCl_2$. The PCR conditions were as follows: denaturation at 95°C for 5 min followed by 30 cycles of 95°C for 1 min; 60°C for 1 min; 69°C for 1 min, and extension at 69°C for 10 min. By using the primers 0Y-1 and 0Y-2 as sequencing primers, sequencing of PCR products was performed as described above.

Results and Discussion

Comparison of Genomic Structure Around the 92-bp Exon. The 92-bp exon is intact in chimpanzees (six individuals), a bonobo (one individual), gorillas (four individuals), orangutans (three individuals), a gibbon (one individual), a baboon (one individual), and a rhesus monkey (one individual) (Figs. 1 and 2). The chimpanzee samples include representatives of two subspecies, Central and West African chimpanzees. These primates all have an *Alu*Sq element ≈350 bp downstream from the 92-bp exon (Figs. 1–3). This *Alu*Sq element, subsequently designated as sah*Alu*Sq after sialic acid hydroxylase *Alu*Sq, belongs to a relatively ancient *Alu*Sq subfamily (average age: 44 million years; ref. 22; see Fig. 3). The *Alu* repetitive family is a primate-specific nonautonomous retroposon and is one of the short interspersed elements (23). The *Alu* family occupies 10.6% of the human genome (24) and is found on average once every 3 kb (23). Insertion of new *Alu* elements into the genome seems to occur by means of target-primed reverse transcription of *Alu* RNA transcript, which is catalyzed by the reverse transcriptase of the L1 non-LTR (long terminal repeat) retroposon (23, 25–27). Such *Alu* insertions typically accompany insertion-site duplications, so that integrated *Alu*s are flanked by short direct repeats of a duplicated insertion site (23, 26). The 5′-TAAAG-3′ sequence immediately adjacent to both ends of the sah*Alu*Sq in the chimpanzee and rhesus monkey (Fig. 1) indicates that this sequence is a direct repeat of the sah*Alu*Sq. In the gorilla, the same repeat is found at the 5′ end of the sah*Alu*Sq, but the 3′ end repeat has an A → G transition. Comparison of sequences surrounding the sah*Alu*Sq element among nonhuman primates reveals that the original target site of sah*Alu*Sq insertion was 5′-TAAAGATTGNTTTTT(TTT)AA-3′. The reverse transcriptase encoded by the human L1 non-LTR retroposon is a key enzyme in *Alu* insertion (23, 26, 27). It contains a domain homologous to the apurinic/apyrimidinic (AP) endonuclease family that can nick DNA by recognizing runs of pyrimidines and purines in a very A+T-rich region (28, 29). This AP endonuclease activity of the reverse transcriptase is essential for target-primed insertion of *Alu*. The sequence deduced as the target site of the sah*Alu*Sq is consistent with these observations. In fact, the

Fig. 1. Comparison of genomic nucleotide sequences around the 92-bp exon of various primate CMP-Neu5Ac hydroxylase genes. Hs, Pt, Gg, and Mm refer to the human, chimpanzee, gorilla, and rhesus monkey, respectively. The shaded boxes represent the 92-bp exon deleted in the hominid lineage. The *Alu* element is represented by the open box. The direct repeats of the sialic acid hydroxylase *Alu*Sq (sah*Alu*Sq) are underlined. The arrowheads indicate replacement boundaries. The 5′-TAAAGATTAATTTTTATTTTT-3′ sequence, which would have a strong preference to the target-priming by the *Alu* poly(A) tail, is located in the 5′ region immediately adjacent to the upstream replacement boundary. Dots refer to identical nucleotides in the other primates; dashes indicate gaps used for sequence alignment. In the gap corresponding to the human deletion, the complete sequences of the other primate genes are shown.

sequence contains potential runs that can be recognized by the reverse transcriptase.

In humans, we found that an *Alu*Y element singly occupies the

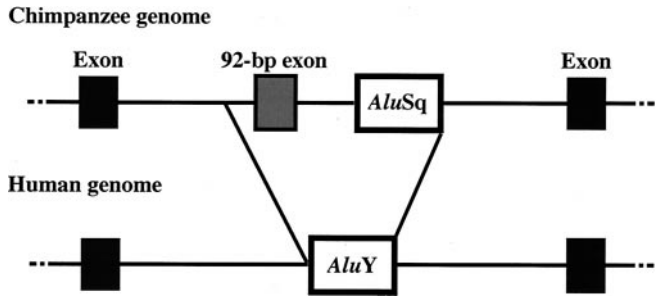

Fig. 2. Schematic comparison of chimpanzee and human CMP-Neu5Ac hydroxylase genomic DNA. In the human genome, the exon to *Alu*Sq (E-A) region in the chimpanzee genome is replaced by the sah*Alu*Y.

region of about 800 bp that, in nonhuman primates, still contains the 92-bp exon and the sah*Alu*Sq (Figs. 1–3). This *Alu*Y element [sialic acid hydroxylase *Alu*Y (sah*Alu*Y)] belongs to a relatively young *Alu* subfamily (average age: 19 million years; ref. 22; see Fig. 3). Because the sah*Alu*Y occurs at exactly the same chromosomal location for all human samples thus far examined (22 humans from Africa, Europe, and Asia), it is likely that the element has been fixed in the population. However, unlike other genomic regions where the human and the chimpanzee differ by 1–2% (30), the sah*Alu*Y in the human and the sah*Alu*Sq in the chimpanzee differ from each other by 17% overall. This discrepancy cannot be accounted for by the high nucleotide substitution rate in the *Alu* family owing largely to the high mutation rate in CpG doublets. It is much more likely that the original human sah*Alu*Sq was replaced by a newly disseminated sah*Alu*Y. This replacement was human-specific (Fig. 1) and accompanied deletion of a genomic region encompassing the 92-bp exon and the sah*Alu*Sq [exon to *Alu*Sq region (E-A region)].

The deletion of the 92-bp exon has resulted in fusion of two

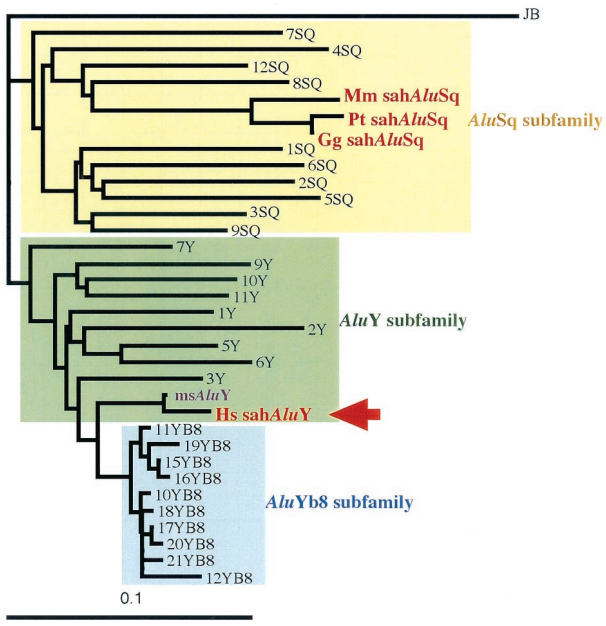

Fig. 3. Phylogenetic tree of *Alu* subfamilies. Human *Alu* elements having intact head and tail were randomly selected from both the GenBank database and the on-line database of *Alu* pairs (http://dir.niehs.nih.gov./ALU/). The tree was made by the neighbor-joining method (20). Distances were calculated with Kimura's two-parameter method (21). The poly(A) tails of sequences were not used in tree-making. The sequence of an *Alu*Jb element, which belongs to the old *Alu*Jb subfamily, was used as an outgroup. The average age of *Alu*Jb, *Alu*Sq, *Alu*Y, and *Alu*Yb8 subfamilies has been estimated at 81, 44, 19, and 3 million years, respectively (ref. 22). The *Alu* elements shown in Fig. 1 are represented by Hs sah*Alu*Y, Pt sah*Alu*Sq, Gg sah*Alu*Sq, and Mm sah*Alu*Sq. ms*Alu*Y indicates the sequence most similar to the one of sah*Alu*Y. Hs, *Homo sapiens*; Pt, *Pan troglodytes*; Gg, *Gorilla gorilla*; Mm, *Macaca mulatta*.

introns (refs. 8 and 9; Fig. 2). Within the fused region of 22.5 kb, there are six other *Alu* elements in addition to sah*Alu*Y(data not shown): one *Alu*Jb, two *Alu*Sxs, two *Alu*Sqs, and one free left *Alu* monomer. The *Alu* density of this region is not especially high and indeed is nearly at the standard level (23). Because all these additional *Alu* elements are shared by humans and chimpanzees (data not shown), they undoubtedly have not been involved in deleting the E-A region in the human. It seems to be the new sah*Alu*Y element that was responsible for the deletion of the E-A region.

***Alu*Y Insertion in the Hominid Lineage.** To find the most closely related *Alu* element to the sah*Alu*Y, we performed a BLAST search at the National Center for Biotechnology Information web site. The closest match is the human *Alu*Y, which is located at positions 20358–20638 of *Homo sapiens* PAC clone RP5–842K16 (GenBank accession no. AC005692; see Fig. 3). The sequence comparison shows that this *Alu*Y element [most similar *Alu*Y (ms*Alu*Y)] differs in four non-CpG sites and two CpG sites from the sah*Alu*Y. We examined the presence or absence of the orthologous ms*Alu*Y among the chimpanzee, gorilla, orangutan, and gibbon by genomic PCR and direct sequencing with primers of the 5′ and 3′ flanking sequences of the ms*Alu*Y. We found that these apes do not possess the ortholog (data not shown). Thus, in addition to sah*Alu*Y, there is another *Alu*Y specific to humans. Although the detailed chromosomal location is not yet identified, it is possible that insertion of ms*Alu*Y provides another instance similar to sah*Alu*Y. The individual members of different *Alu* subfamilies are thought to have arisen by amplification of a small subset of "source" genes, which allows subfamilies to evolve in a sequential order (ref. 31; see Fig. 3). One such is sah*Alu*Y and another is ms*Alu*Y, both of which have been inserted uniquely in the hominid lineage. This information would be useful to address the timing of sah*Alu*Y dissemination in the hominid genome. This problem will be considered elsewhere, but here we would like to point out that the dissemination could not be very recent. Because there are six nucleotide differences between the two human-specific *Alu*s, it is unlikely that the dissemination occurred within the past 1 million years or so even though the substitution rate in *Alu*s is generally high (30).

Model. No apparent human-specific sequence feature exists at the boundaries of the E-A region (Fig. 1). It is therefore reasonable to assume that the deletion was triggered by an accidental event that has no sequence preference. A likely possibility is a double-strand break that can be induced by a wide range of factors such as oxidative damage, ionizing radiation, mechanical stress, and action of DNA endonucleases. To model a series of molecular events, we first assume that the deletion was initiated by a double-strand break. We note that a particular sequence of 5′-TAAAGATTAATTTTATTTTT-3′ is found in the 5′ region immediately adjacent to the upstream deletion boundary in the human, chimpanzee, and gorilla (Fig. 1) and that this sequence is similar to the target site of the sah*Alu*Sq and may have a strong preference to the target-priming by the *Alu* poly(A) tail. The sah*Alu*Y in the human can be easily aligned with the sah*Alu*Sq in nonhuman primates, although the sequence similarity of the tail region is somewhat lower than that of the head region. Furthermore, both head ends of the sah*Alu*Y and sah*Alu*Sq elements are identical with the downstream deletion boundary. These findings suggest that free sah*Alu*Y RNA transcript interfered in the recombinational repair of a double-strand break because of the target-priming by its poly(A) tail and the annealing to the sah*Alu*Sq by its cDNA that resulted from target-primed reverse transcription. On the basis of these considerations, we propose possible molecular mechanisms that caused the *Alu*-mediated replacement event (Fig. 4).

As a first step, a double-strand break occurs at a position that provides the 5′ deletion boundary (5′ end of the E-A region; see Figs. 1 and 2) (Fig. 4*A*). Recombinational repair then starts with 5′-to-3′ exonucleolytic digestion of one DNA strand, which leads to the formation of 3′-overhanging single-stranded DNA tails (Fig. 4*B*). After homologous recombination between the injured and intact alleles, free sah*Alu*Y RNA transcript interferes with the repair process through two mechanisms. One mechanism is the target-priming to the 5′-TAAAGATTAATTTT-TATTTTT-3′ sequence immediately upstream from the double-strand breakpoint by its poly(A) tail (Fig. 4*C*). The second mechanism consists of the annealing to the original sah*Alu*Sq downstream from the 92-bp exon by its cDNA, which can result from the target-primed reverse transcription (Fig. 4 *D–F*). The target-priming occurs without enzymatic nicking by L1 reverse transcriptase because the double-strand break provides the nick of DNA (Fig. 4*C*). The reverse transcription follows the target-priming and produces the sah*Alu*Y cDNA from its RNA transcript (Fig. 4*D*). After elimination of sah*Alu*Y RNA transcript (Fig. 4*E*), the sah*Alu*Y cDNA anneals to the genomic sah*Alu*Sq (Fig. 4*F*). The annealing intensity of the *Alu* tail region is presumed to be somewhat lower than that of the *Alu* head region because of the regional variation of sequence similarity between the sah*Alu*Y and sah*Alu*Sq. These actions bring the sah*Alu*Sq close to the target site of the sah*Alu*Y and block the DNA polymerase extension from the double-strand breakpoint to the head of the sah*Alu*Sq (Fig. 4 *F* and *G*). DNA synthesis then starts from the head of the sah*Alu*Y cDNA (Fig. 4*G*). Finally, the rearranged allele emerges by DNA replication (Fig. 4*H*). Thus,

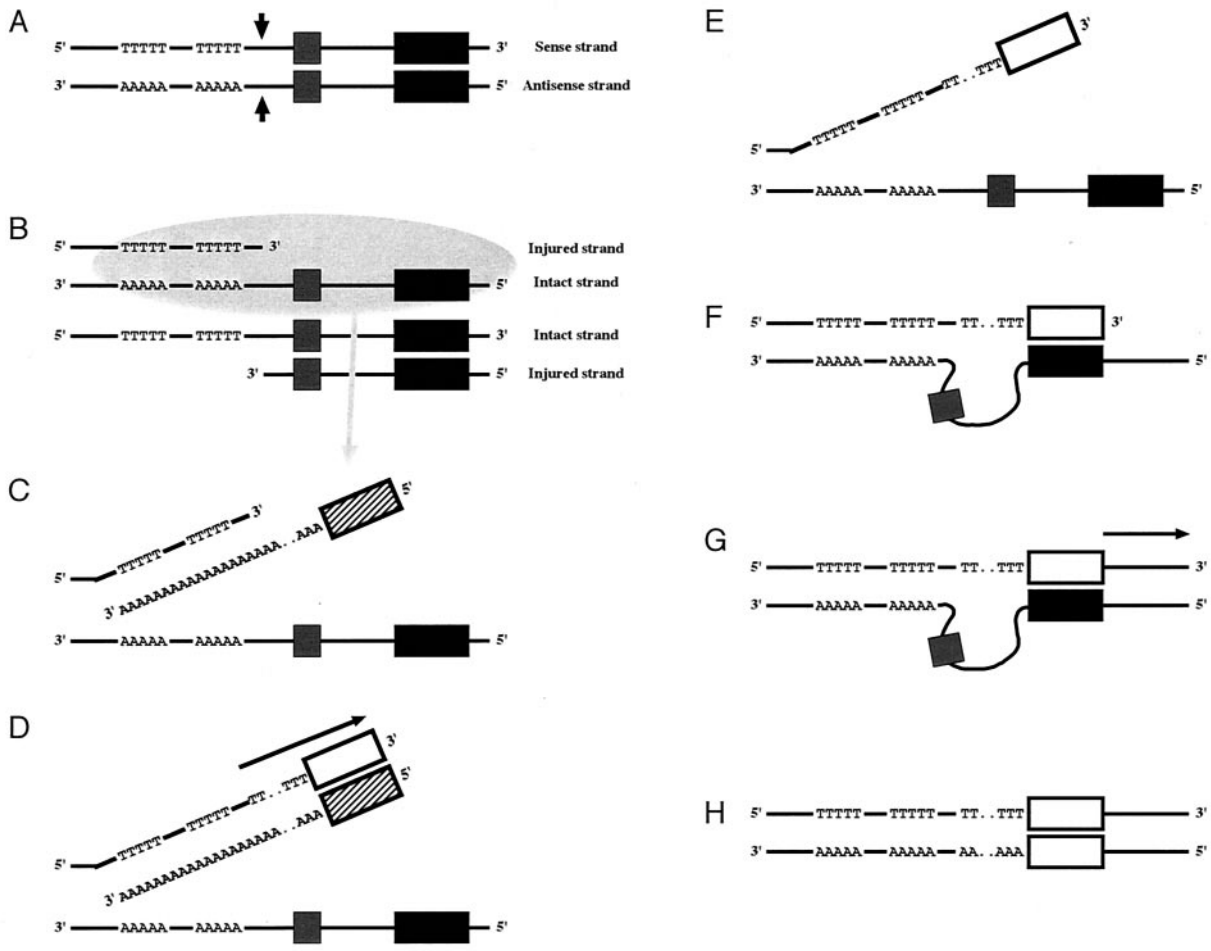

Fig. 4. Model of the *Alu*-mediated replacement event that occurred in the CMP-Neu5Ac hydroxylase gene in the hominid lineage. (*A*) Double-strand break indicated by vertical arrows. The solid boxes represent the sah*Alu*Sq elements; the shaded box represents the 92-bp exon. The *Alu* target region containing A/T stretch is located in the 5′ immediately adjacent region of the double-strand break point. (*B*) Homologous recombination between the injured and intact alleles, after 5′-to-3′ exonucleolytic digestion, which generates a 3′-single-stranded tail. (*C*) Target-priming to the target site by free sah*Alu*Y RNA transcript. Free sah*Alu*Y RNA transcript is indicated by both a cross-hatched box and the letter ''A,'' representing the poly(A) tail. (*D*) Reverse transcription. An arrow indicates reverse transcription. The open box represents the sah*Alu*Y cDNA. (*E*) Elimination of RNA. (*F*) Annealing between the sah*Alu*Y cDNA and genomic sah*Alu*Sq. (*G*) DNA synthesis. An arrow indicates DNA synthesis. (*H*) Production of the allele that lacks the E-A region by DNA replication. This allele is derived from the upper strand shown in *G*.

Alu-induced incomplete repair of the double-strand break replaces the E-A region by the sah*Alu*Y.

The model above (Fig. 4) proposes that a double-strand break repair and *Alu* target-primed reverse transcription cause an *Alu*-mediated replacement event. It requires the following structural condition for exon deletion: an exon exists between the tail of a genomic *Alu* and an *Alu* target site. The primate genome contains abundant *Alu* elements (24, 32, 33), and A+T-rich sequences, which can be regarded as potential *Alu* target sites, frequently occur in the genome. The sequence data from the Human Genome Project confirm that gene-rich regions are *Alu*-rich (24, 32, 33). Thus, the above-mentioned condition is actually met in the human genome.

The model can also predict *Alu* conversion (34) when a double-strand break occurs within an *Alu* element. An inserted *Alu* is sandwiched in between direct repeats that are derived from their target site (23, 26). In *Alu* conversion, the poly(A) tail of *Alu* RNA transcript primes to the adjacent flanking sequence of a genomic *Alu* element, and therefore *Alu* cDNA anneals to the genomic *Alu* element without loop structure (Fig. 4 *F* and *G*). This priming introduces a replacement of sequences within an *Alu* element, leading to *Alu* conversion. Such an example of *Alu* conversion has been reported in the low-density lipoprotein receptor gene (34). However, in that case an exon deletion did not follow (34).

Role of *Alu* Repetitive Family in Primate Evolution. Because *Alu*-related events (e.g., *Alu* insertion and *Alu*--*Alu* recombination) can be responsible for diseases caused by abnormal truncations and rearrangements of genes (23, 27), *Alu*s are generally regarded as a disadvantageous or at best neutral agents of organismal evolution. Accordingly, the sah*Alu*Y could be a ''destructive agent'' in primate evolution. However, the situation of inactivation of the hydroxylase gene might be different. As discussed below, inactivation could actually have been favored and fixed in the human population (8, 10).

Genomic *Alu*s are known to be capable of contributing to regulation of gene expression (35–40). Such *Alu*s are referred to as ''regulatory *Alu*s'' because cis-acting regulatory elements reside within *Alu*s and many *Alu* classes possess consensus sequences of such regulatory elements (35–39). Hamdi *et al.* (41) reported that some regulatory *Alu*s are differentially distributed among primates. This finding supports the notion that *Alu*-related events, such as *Alu* insertion, *Alu*–*Alu* recombination,

Alu conversion, and *Alu*-mediated replacement, might have played significant roles in diversification of primates. In agreement with this idea, the quantitative analyses of the *Alu*Ya5 and *Alu*Yb8 subfamilies of humans and great apes revealed that the rate of *Alu* insertion has increased specifically in the hominid lineage (23, 42). The hominid lineage could thus be unique in terms of a high frequency of *Alu*-related events.

We have provided evidence that *Alu* insertion caused an exon deletion. It is reported that, although the molecular mechanism has not yet been elucidated, L1 retroposons can transduce surrounding genomic sequences in retrotransposition and induce exon shuffling (43). It is thus possible that L1s have also had an impact on genome evolution as "editing agents." However, in the human genome, *Alu*s might have played more important roles in rearranging exons than L1s, because, in general, *Alu*s are inserted in gene-rich regions whereas L1s are inserted in gene-poor regions (24, 32, 33).

Pathogen-Mediated Selection of the *Alu*-Mediated Gene Inactivation. Many microbial pathogens initiate infection by binding to sialic acids, and some pathogens exert distinct preference for particular types of sialic acids. Influenza viruses show distinct species preference based on Neu5Ac or Neu5Gc expression (44–47), and enterotoxigenic bacteria *Escherichia coli* K99 adhere specifically to ganglioside GM3(Neu5Gc), but not to GM3(Neu5Ac) in intestinal epithelial cells (48). Furthermore, the amount of human *Alu* RNA transcripts increases during viral infection (49–51). It is possible that virus infection was somehow related to the hydroxylase inactivation. Thus, a lack of Neu5Gc expression may have conferred protection against infectious pathogens that prefer Neu5Gc. *Homo erectus* was the first species of *Homo* whose population expanded widely in mainland Eurasia and Africa (1). The range of *H. sapiens* expanded further until it included almost every corner of the globe. The genus *Homo* undoubtedly had to adapt to a wide range of new environments. It is tempting to speculate that the lack of Neu5Gc enabled our ancestors to expand their habitats, first, by evading various animal infectious agents in new environments of *H. erectus*, and second, by decreasing the infectious risk of *H. sapiens* from domestication of other vertebrates (some of whose current microbial pathogens are known to prefer Neu5Gc as a binding site) (44–47, 52, 53).

We thank Alain Silk for help with cell culture and sequencing, and two anonymous reviewers for their constructive criticisms. This research was supported in part by Japan Society for Promotion of Science Grant 12304046 (to N.T.).

1. Klein, R. G. (1999) *The Human Career: Human Biological and Cultural Origins* (Univ. of Chicago Press, Chicago).
2. Satta, Y., Klein, J. & Takahata, N. (2000) *Mol. Phylogenet. Evol.* **14,** 259–275.
3. Ruvolo, M. (1997) *Mol. Biol. Evol.* **14,** 248–265.
4. Gagneux, P. & Varki, A. (2001) *Mol. Phylogenet. Evol.* **18,** 2–13.
5. Gibbons, A. (1998) *Science* **281,** 1432–1434.
6. Szabó, Z., Levi-Minzi, S. A., Christiano, A. M., Struminger, C., Stoneking, M., Batzer, M. A. & Boyd, C. D. (1999) *J. Mol. Evol.* **49,** 664–671.
7. Winter, H., Langbein, L., Krawczak, M., Cooper, D. N., Jave-Suarez, L. F., Rogers, M. A., Praetzel, S., Heidt, P. J. & Schweizer, J. (2001) *Hum. Genet.* **108,** 37–42.
8. Chou, H.-H., Takematsu, H., Diaz, S., Iber, J., Nickerson, E., Wright, K. L., Muchmore, E. A., Nelson, D. L., Warren, S. T. & Varki, A. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 11751–11756.
9. Irie, A., Koyama, S., Kozutsumi, Y., Kawasaki, T. & Suzuki, A. (1998) *J. Biol. Chem.* **273,** 15866–15871.
10. Muchmore, E. A., Diaz, S. & Varki, A. (1998) *Am. J. Phys. Anthropol.* **107,** 187–198.
11. Shaw, L. & Schauer, R. (1988) *Biol. Chem. Hoppe-Seyler* **369,** 477–486.
12. Muchmore, E. A., Milewski, M., Varki, A. & Diaz, S. (1989) *J. Biol. Chem.* **264,** 20216–20223.
13. Kozutsumi, Y., Kawano, T., Yamakawa, T. & Suzuki, A. (1990) *J. Biochem. (Tokyo)* **108,** 704–706.
14. Kawano, T., Koyama, S., Takematsu, H., Kozutsumi, Y., Kawasaki, H., Kawashima, S., Kawasaki, T. & Suzuki, A. (1995) *J. Biol. Chem.* **270,** 16458–16463.
15. Schauer, R. (1982) *Sialic Acids: Chemistry, Metabolism and Function,* Cell Biology Monographs (Springer, New York), Vol. 10.
16. Kelm, S. & Schauer, R. (1997) *Int. Rev. Cytol.* **175,** 137–240.
17. Varki, A. (1997) *FASEB J.* **11,** 248–255.
18. Brinkman-Van der Linden, E. C. M., Sjoberg, E. R., Juneja, L. R., Crocker, P. R., Varki, N. & Varki, A. (2000) *J. Biol. Chem.* **275,** 8633–8640.
19. Schlenzka, W., Shaw, L., Kelm, S., Schmidt, C. L., Bill, E., Trautwein, A. X., Lottspeich, F. & Schauer, R. (1996) *FEBS Lett.* **385,** 197–200.
20. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
21. Kimura, M. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 454–458.
22. Kapitonov, V. & Jurka, J. (1996) *J. Mol. Evol.* **42,** 59–65.
23. Schmid, C. W. (1998) *Nucleic Acids Res.* **26,** 4541–4550.
24. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al. (2001) Nature (London)* **409,** 860–921.
25. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. (1993) *Cell* **72,** 595–605.
26. Jurka, J. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 1872–1877.
27. Kazazian, H. H., Jr. (1998) *Curr. Opin. Genet. Dev.* **8,** 343–350.
28. Feng, Q., Moran, J. V., Kazazian, H. H., Jr., & Boeke, J. D. (1996) *Cell* **87,** 905–916.
29. Martin, F., Oliveres, M., Lopez, M. C. & Alonso, C. (1996) *Trends Biochem. Sci.* **21,** 283–285.
30. Chen, F.-C. & Li, W.-H. (2001) *Am. J. Hum. Genet.* **68,** 444–456.
31. Schmid, C. W. & Maraia, R. (1992) *Curr. Opin. Genet. Dev.* **2,** 874–882.
32. Dunham, I., Hunt, A. R., Collins, J. E., Bruskiewich, R., Beare, D. M., Clamp, M., Smink, L. J., Ainscough, R., Almeida, J. P., Babbage, A., *et al*. (1999) *Nature (London)* **402,** 489–495.
33. Hattori, M., Fujiyama, A., Taylor, T. D., Watanabe, H., Yada, T., Park, H.-S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.-K., *et al*. (2000) *Nature (London)* **405,** 311–319.
34. Kass, D. H., Batzer, M. A. & Deininger, P. L. (1995) *Mol. Cell. Biol.* **15,** 19–25.
35. Brini, A. T., Lee, G. M. & Kinet, J.-P. (1993) *J. Biol. Chem.* **268,** 1355–1361.
36. Hambor, J. E., Mennone, J., Coon, M. E., Hanke, J. H. & Kavathas, P. (1993) *Mol. Cell. Biol.* **13,** 7056–7070.
37. McHaffie, G. S. & Ralston, S. H. (1995) *Bone* **17,** 11–14.
38. Norris, J., Fan, D., Aleman, C., Marks, J. R., Futreal, P. A., Wiseman, R. W., Iglehart, J. D., Deininger, P. L. & McDonnell, D. P. (1995) *J. Biol. Chem.* **270,** 22777–22782.
39. Vansant, G. & Reynolds, W. F. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 8229–8233.
40. Fornasari, D., Battaglioli, E., Flora, A., Terzano, S. & Clementi, F. (1997) *Mol. Pharmacol.* **51,** 250–261.
41. Hamdi, H. K., Nishio, H., Tavis, J., Zielinski, R. & Dugaiczyk, A. (2000) *J. Mol. Biol.* **299,** 931–939.
42. Zietkiewicz, E., Richer, C., Makalowski, W., Jurka, J. & Labuda, D. (1994) *Nucleic Acids Res.* **22,** 5608–5612.
43. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H., Jr. (1999) *Science* **283,** 1530–1534.
44. Higa, H. H., Rogers, G. N. & Paulson, J. C. (1985) *Virology* **144,** 279–282.
45. Ito, T., Suzuki, Y., Mitnaul, L., Vines, A., Kida, H. & Kawaoka, Y. (1997) *Virology* **227,** 493–499.
46. Suzuki, T., Horiike, G., Yamazaki, Y., Kawabe, K., Masuda, H., Miyamoto, D., Matsuda, M., Nishimura, S.-I., Yamagata, T., Ito, T., *et al*. (1997) *FEBS Lett.* **404,** 192–196.
47. Ito, T., Suzuki, Y., Suzuki, T., Takada, A., Horimoto, T., Wells, K., Kida, H., Otsuki, K., Kiso, M., Ishida, H., *et al.* (2000) *J. Virol.* **74,** 9300–9305.
48. Smit, H., Gaastra, W., Kamerling, J. P., Vliegenthart, J. F. G. & de Graaf, F. K. (1984) *Infect. Immun.* **46,** 578–584.
49. Panning, B. & Smiley, J. R. (1994) *Virology* **202,** 408–417.
50. Panning, B. & Smiley, J. R. (1995) *J. Mol. Biol.* **248,** 513–524.
51. Russanova, V. R., Driscoll, C. T. & Howard, B. H. (1995) *Mol. Cell. Biol.* **15,** 4282–4290.
52. Kyogashima, M., Ginsburg, V. & Krivan, H. C. (1989) *Arch. Biochem. Biophys.* **270,** 391–397.
53. Delorme, C., Brüssow, H., Sidoti, J., Roche, N., Karlsson, K.-A., Neeser, J.-R. & Teneberg, S. (2001) *J. Virol.* **75,** 2276–2287.