# Inactivation of CMP-*N*-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution

**Hsun-Hua Chou*†, Toshiyuki Hayakawa†‡, Sandra Diaz*, Matthias Krings§¶, Etty Indriati‖, Meave Leakey**, Svante Paabo§, Yoko Satta‡, Naoyuki Takahata‡, and Ajit Varki*††**

*Glycobiology Research and Training Center, Departments of Medicine and Cellular and Molecular Medicine, University of California at San Diego, La Jolla, CA 92093-0687; ‡Department of Biosystems Science, Graduate University for Advanced Studies (Sokendai), Hayama, Kanagawa 240-0193, Japan; §Max Planck Institute for Evolutionary Anthropology, D-04103 Liepzig, Germany; ‖Gadjah Mada University, Yogyakarta, Indonesia; and **Kenya National Museums, Nairobi, Kenya

Edited by Morris Goodman, Wayne State University School of Medicine, Detroit, MI, and approved July 18, 2002 (received for review April 30, 2002)

**Humans are genetically deficient in the common mammalian sialic acid *N*-glycolylneuraminic acid (Neu5Gc) because of an *Alu*-mediated inactivating mutation of the gene encoding the enzyme CMP-*N*-acetylneuraminic acid (CMP-Neu5Ac) hydroxylase (CMAH). This mutation occurred after our last common ancestor with bonobos and chimpanzees, and before the origin of present-day humans. Here, we take multiple approaches to estimate the timing of this mutation in relationship to human evolutionary history. First, we have developed a method to extract and identify sialic acids from bones and bony fossils. Two Neandertal fossils studied had clearly detectable Neu5Ac but no Neu5Gc, indicating that the CMAH mutation predated the common ancestor of humans and Neandertals, ≈0.5–0.6 million years ago (mya). Second, we date the insertion event of the inactivating human-specific sah*Alu*Y element that replaced the ancestral *Alu*Sq element found adjacent to exon 6 of the CMAH gene in the chimpanzee genome. Assuming *Alu* source genes based on a phylogenetic tree of human-specific *Alu* elements, we estimate the sah*Alu*Y insertion time at ≈2.7 mya. Third, we apply molecular clock analysis to chimpanzee and other great ape CMAH genes and the corresponding human pseudogene to estimate an inactivation time of ≈2.8 mya. Taken together, these studies indicate that the CMAH gene was inactivated shortly before the time when brain expansion began in humankind's ancestry, ≈2.1–2.2 mya. In this regard, it is of interest that although Neu5Gc is the major sialic acid in most organs of the chimpanzee, its expression is selectively down-regulated in the brain, for as yet unknown reasons.**

hominid evolution | sialic acids | *Alu* sequences

Sialic acid (Sia) is a generic term for a family of acidic monosaccharides found at the terminal ends of sugar chains attached to cell surfaces and to soluble glycoproteins (1–3). A major biochemical difference between humans and other mammals, including the closest living relatives of humans (chimpanzees and bonobos), is in the expression of the Sia *N*-glycolylneuraminic acid (Neu5Gc). Whereas human tissues and body fluids contain little or no detectable Neu5Gc, corresponding samples from chimpanzees and bonobos (and the other great apes, gorillas and orangutans) express high levels (4). Humans instead express an excess of the precursor Sia *N*-acetylneuraminic acid (Neu5Ac). Human deficiency of Neu5Gc is due to inactivation of the gene for CMP-*N*-acetylneuraminic acid (CMP-Neu5Ac) hydroxylase (CMAH), which converts CMP-Neu5Ac into CMP-Neu5Gc in other animals (3, 5, 6).

Neu5Gc expression in non-human mammals is developmentally regulated and tissue-specific (1, 2, 7–9), implying multiple biological roles. Some Sia-binding proteins can distinguish Neu5Gc from Neu5Ac. Thus, the CMAH mutation could have altered interactions involving endogenous human receptors such as sialoadhesin/Siglec-1 (10) and myelin-associated glycoprotein/Siglec-4a (11), as well as the binding of pathogenic microorganisms such as influenza A (12–14), rotaviruses (15), and *Escherichia coli* K99 (16). Such differences could have potentially affected human ontogeny, physiology, disease susceptibility, and/or the ability of humans to domesticate livestock.

Human Neu5Gc deficiency is due to a 92-bp frame-shifting exon deletion in the CMAH gene (5, 6) giving a markedly truncated protein (5), lacking amino acid residues necessary for enzyme activity (17). This mutation is homozygous in all human populations but absent in great apes (3, 5, 18)—i.e., it occurred after our last common ancestor with chimpanzees but before the diaspora of present-day humans. This mutation was apparently caused by a human-specific sah*Alu*Y element that replaced an ancestral *Alu*Sq element found adjacent to exon 6 of the CMAH gene in the genomes of great apes (18).

Regardless of the original selection process(es) involved in human loss of CMAH (e.g., a pathogen preferentially recognizing Neu5Gc as a host receptor), there could have been secondary biological consequences. Thus, determining when the mutation occurred would allow rational hypotheses regarding its potential relationship to changes during human evolution. Unfortunately, DNA sequence information is impossible to retrieve from fossils older than ≈100,000 years (19, 20). We therefore used three other independent methods to date the inactivation of the human CMAH gene. First, we reasoned that Sias are more likely to survive in fossils than DNA, as they are 9-carbon monosaccharide units rather than polymers of nucleic acids. Thus, we directly analyzed present-day human, great ape, and Neandertal and other fossil samples for the presence of Neu5Ac and Neu5Gc. Second, we estimated the timing of the *Alu* integration event that inactivated the human CMAH (18). Finally, because pseudogenes have different rates of nucleotide substitution as compared with active genes subject to selection, we used the molecular clock approach to date the mutation in the human CMAH gene (21).

## Materials and Methods

**Materials.** Epstein–Barr virus-transformed B cells from the great apes were kindly provided by Peter Parham (Stanford University, Stanford, CA). Contemporary primate bone samples were generously donated by the Natural History Museum in San

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

Abbreviations: Sia, sialic acid; Neu5Gc, *N*-glycolylneuraminic acid; Neu5Ac, *N*-acetylneuraminic acid; mya, million years ago; DMB, 1,2-diamino-4,5-methylene dioxybenzene; CMAH, CMP-*N*-acetylneuraminic acid hydroxylase.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF494221, AF494222, AF494223, AF494224, and AF494225).

†H.-H.C and T.H. contributed equally to this work.

¶Present address: Theron Business Consulting, Munich, Germany.

††To whom reprint requests should be addressed. E-mail: avarki@ucsd.edu.

Diego. Faunal fossil samples (cave bear, dugong, mammoth, etc.) were purchased from recognized dealers at the 1999 International Fossil Show in Tucson, AZ. Georgian Neandertal fossils were kindly provided by Ivan Nasidze (22). *Homo erectus* fossil samples (Solo skulls) found in Ngandong, East Java in the 1930s as well as fossil fauna from Ngandong excavated in 1976 (23) were obtained from Gadjah Mada University in Yogyakarta, with permission from Teuku Jacob. Kenyan fossil samples were collected from sites in the Turkana Basin dating between 4 and 1.5 million years, including Nariokotome, Nachukui, Kalachoro, Lomekwi, and Kanapoi to the west of the modern lake, and Allia Bay and Koobi Fora to the east (24–26).

**Release, Purification, and Analysis of Sias from Bones and Fossils.** The surface of the bone or fossil was filed off, and a powdered sample was collected by drilling into the center, which was then decalcified and solubilized in 0.5 M EDTA, pH 8, and 5% sarkosyl at 25°C overnight, followed by digestion with 10 mg/ml proteinase K at 37°C overnight. A Centricon-30 filtration unit was used to collect glycopeptides (in the run-through), which were hydrolyzed in 80 vol of 4 M acetic acid at 80°C for 3 h to release Sias. Cations were eliminated by using AG50W-X2 resin (H+ form) in water. After lyophilization, the sample was resuspended in 10 mM acetic acid, and the precipitate containing the acid form of EDTA was centrifuged out. The sample was then diluted to a reading of $<20$ mosM, applied to an AG1X8 anion exchange column (formate form), and Sias were eluted with 1 M formic acid. After passage through a SPICE C18 cartridge, the sample was finally lyophilized. Sia recovery was confirmed in pilot experiments using external sialoglycoprotein or internal [$^3$H]Neu5Ac standards. Purified samples were derivatized with 1,2-diamino-4,5-methylene dioxybenzene (DMB), and fluorescent DMB adducts ($E_x$ = 373 nm, $E_m$ = 448 nm) were resolved by reverse phase HPLC (27). The nature of the DMB derivatives of Sias was confirmed by mass spectrometry (28). Peak areas corresponding to the expected elution times of Sias were collected, concentrated, and analyzed by using a Finnigan MAT HPLC with online mass spectrometer model LCQ-mass spectrometer system A (28). A Varian C18 column was eluted isocratically at 0.9 ml/min with 8% acetonitrile/7% methanol/0.1% formic acid in water over 50 min, and the eluent was monitored by UV absorbance at 373 nm and by electrospray ionization mass spectrometry (capillary temperature 210°C, capillary voltage 31 V, and lens offset voltage 0 V). Spectra were acquired by scanning from $m/z$ 150-2000 in the positive-ion mode. In some instances, MS/MS spectra were acquired by selecting the parent mass and applying a 20% normalized collision energy. Data analysis was performed using the XCALIBUR data analysis program from the manufacturer.

**Collection and Comparison of *Alu* Sequences.** Eighty-six intact *Alu*Yb8 elements that are human-specific and fixed in human populations were picked up from a recently reported *Alu*Yb8 list (29) (see *Supporting Text*, which is published as supporting information on the PNAS web site, www.pnas.org, for a full list of sequences used for the analysis). The phylogenetic tree of each subfamily was made by the neighbor-joining method (30). The actual number of nucleotide substitutions was calculated by Kimura's two-parameter method (31). Poly(A) tails in the sequences were excluded from the analysis.

**Sequencing of Great Ape CMAH cDNAs.** Reverse transcription–PCR was performed on total RNA from Epstein–Barr virus-transformed great ape lymphocytes (kindly provided by Peter Parham). The mRNA was reverse-transcribed by using 50 pmol of random hexamer or CMAH-specific primer, 200 units of Superscript II, and 20 units of RNase inhibitor in a 40-$\mu$l reaction with first-strand buffer, 200 $\mu$M dNTP, and 10 mM DTT. A 2-$\mu$l portion of the reverse transcription reaction was then PCR-amplified with Boehringer Mannheim's Expand Long Template PCR system with 2.25 mM $MgCl_2$ and detergents; 20 pmol each of primers S5-Hom (GGCAGACGATGGGCAGCATCG) and A3-Hom (TGGATTCGTATCTACTACAG) were used, based on homologous sequences of known murine, human, and chimp CMAH genes. Some sequences were reamplified with a second round of PCR. PCR products were resolved on 1% agarose gels, gel-purified, and directly sequenced with 14 primers spanning both directions of the entire CMAH sequence, based on homologous sequences from the murine, human, and chimp. Data were verified visually and assembled into contiguous sequences by using SEQUENCHER 3.1. Final sequences are based on overlapping sequencing runs covering both strands.

## Results

**Sias Can Be Purified from Bone and Fossil Samples.** Sias are traditionally isolated from soft tissues. We established a modified protocol for Sia purification from highly mineralized bone tissue. Normally, after soft-tissue samples are mechanically disrupted and cells are lysed, the fraction of interest is subjected to mild acid hydrolysis to release Sias from their attachment to sugar chains. Released Sias are then purified by ion-exchange chromatography. In keeping with prior experience with bony fossils (19), we used 5% Sarkosyl and EDTA to demineralize and solubilize the powdered bone samples. However, the high EDTA concentration inhibited subsequent release of Sias by mild acid, which is normally performed using 2 M acetic acid. Using known sialoglycoprotein standards, we established that 80 vol of 4 M acetic acid were required to overcome the buffering effect of EDTA and obtain proper release of Sias. The high concentration of EDTA also inhibited subsequent binding of released Sias to the anion-exchange column. This problem was overcome by cation-exchange chromatography and subsequent precipitation of the resulting hydrogen form of EDTA under acidic conditions. This was followed by dilution to yield an electrolyte concentration of less than 20 mM, which no longer interferes with the anion-exchange step. In early experiments, adequate Sia recovery was monitored by adding tracer amounts of [$^3$H]Neu5Ac to the starting samples. As shown in Fig. 1*A*, this modified protocol allowed the detection of Sias in contemporary human and great ape bone samples. As expected, human bones gave only a single peak of Neu5Ac, whereas all great ape samples gave an easily detectable peak of Neu5Gc as well.

Sias could also be recovered from some, but not all, Pleistocene and Miocene fossils (Fig. 1*B*). Although older fossil samples showed increasing contaminating peaks in the region where Sia DMB adducts eluted, their presence could be confirmed by mass spectrometry, where the primary ion masses and electron-impact fragmentation patterns corresponded to those expected for Neu5Ac and Neu5Gc (Fig. 2).

**Neandertals, Like Humans, Lacked Neu5Gc Expression.** As shown in Fig. 3*A*, the HPLC profile of Sias purified from the Neandertal-type specimen contained Neu5Ac but no Neu5Gc. This particular sample was also originally processed with great care to avoid contamination during the PCR amplification of Neandertal mitochondrial DNA sequences (19). To confirm that handling did not cause contamination, a control for the processing steps was run alongside the test sample. This showed only a very minor peak in the area of Neu5Ac. Neu5Gc and Neu5Ac do not show any major differences in degradation rates under various conditions of pH and temperature. Based on the amount of Neu5Ac recovered from the Neandertal sample, Neu5Gc would have been detected if present at a level higher than 0.04% of the total Sias. Fig. 3*B* shows the DMB-HPLC profiles of Sias extracted from two additional Neandertal specimens. The Neandertal tooth root ID no. 486 (from Sakajia, Georgia) gave a finding very similar to the type specimen—i.e., clearly detectable Neu5Ac but no Neu5Gc. In this case, there was sufficient Sia recovered to

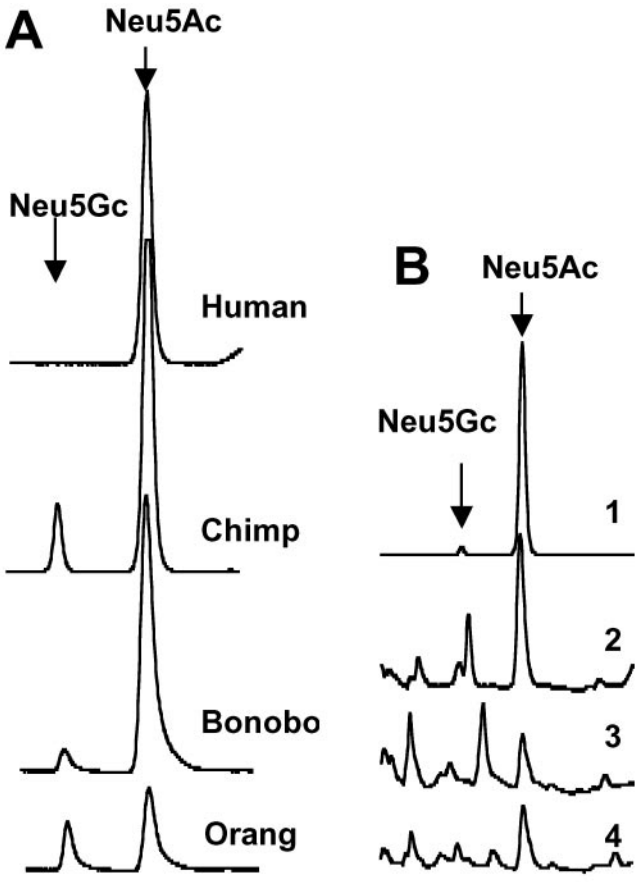


**Fig. 1.** Sias from bones and fossilized bones. Sias were released and purified, derivatized with DMB, and analyzed by HPLC. The elution positions of standard Neu5Ac and Neu5Gc are indicated. (*A*) Contemporary bones of humans and great apes. (*B*) Pleistocene and Miocene fossils. 1, mammoth tusk, Holocene, ≈1,000 years ago (kya); 2, cave bear jaw, Pleistocene, 40–80 kya; 3, deer leg bone, Pleistocene, 40–80 kya; and 4, dugong femur, Miocene, ≈20 mya.

confirm the Neu5Ac peak by mass spectrometry. We were unable to isolate any Sias from the other Neandertal specimen, the tooth root ID no. 2042 from Tsoutskhvati, Georgia.

It is very unlikely that positive samples were simply contaminated with exogenous Sias. First, samples were collected from the center of the specimen. Second, Sias are not present in plants, most invertebrates, or most microorganisms (2). Third, only some samples contained Sias. Fourth, blank controls incorporating all reagents and processing steps ruled out significant contamination during handling (Fig. 3*A*). Thus, the CMAH gene inactivation predated the human–Neandertal common ancestor ≈500,000–600,000 years ago (19).

Samples from *H. erectus* fossils found above the Solo River near Ngandong, Java in the mid-1900s, believed to be as recent as 50,000 years old (23), were examined, as were numerous faunal samples collected from sites (see *Materials and Methods*), where African bipedal hominid fossils had previously been found (24–26). Unfortunately, none of these samples yielded detectable Sias, indicating that the fossilization conditions at these sites were not conducive to Sias preservation. Based on these experiences, it is unlikely that current methods will allow us to recover and detect Sias from fossils found in tropical and subtropical regions. This is consistent with work by others indicating that the mean temperature in the area of fossilization is a major determinant of whether ancient DNA can be recovered (32).

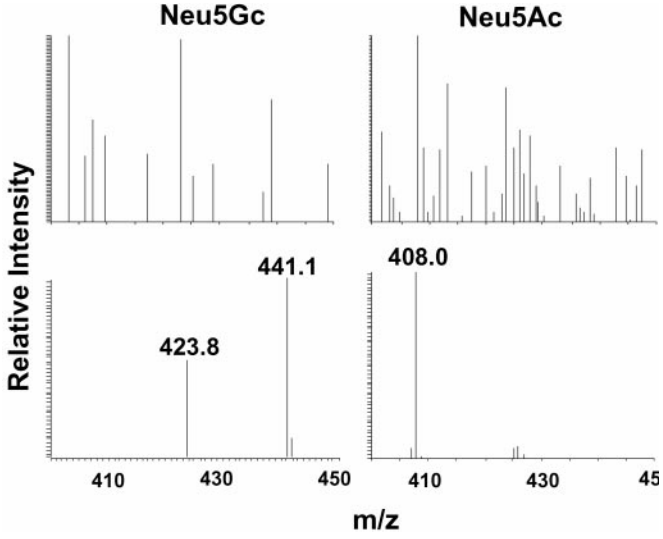


**Fig. 2.** Example of mass spectrometric proof of Sias isolated from fossils. DMB-derivatized Sias from a Pleistocene cave bear jaw (*B*2 in Fig. 1) were separated by C18 HPLC, peak areas corresponding to the expected elution times of Sias collected, concentrated, and analyzed by mass spectrometry. (*Upper*) $m/z$ profiles obtained at the elution times of the DMB derivatives of Neu5Ac and Neu5Gc. The ions corresponding to hydrated DMB derivatives of Neu5Ac and Neu5Gc (426 and 442) are not easily seen in the presence of contaminating ions of higher intensity. (*Lower*) The outcome of secondary mass spectrometry performed on the above ions, which each lose one molecule of water, giving DMB-Neu5Ac (408) and DMB-Neu5Gc (424). Variable amounts of the parent ion are also seen in the secondary spectrum.

**Analysis of *Alu* Sequences to Time the Inactivation of the CMAH Gene.** The *Alu* family of transposable elements includes primate-specific, nonautonomous retroposons comprising ≈10% of the human genome (33, 34). Inactivation of CMAH evidently involved replacement of an ancient *Alu*Sq present in the human–chimpanzee ancestral genome with a younger sah*Alu*Y (sialic acid hydroxylase *Alu*Y) that is related to the *Alu*Yb8 subfamily (18). We used the timing of human *Alu* integration events to calculate the inactivation time of CMAH.

Individual members of *Alu* subfamilies seem to have arisen by amplification of a small subset of "source" or "master" genes (33, 35). Thus, the actual number of nucleotide differences between an *Alu* member and its source gene is directly proportional to the age of the *Alu* member. Because the time ($t_1$) of the CMAH inactivation would be equal to the insertion time ($t_{sah}$) of sah*Alu*Y, it can be estimated as $t_1 = t_{sah} = d_{sah}/\lambda$, where $d_{sah}$ is the actual number of nucleotide substitutions between sah*Alu*Y and its source gene, and $\lambda$ is the nucleotide substitution rate of human *Alu*s per site per year. Kimura's two-parameter model (31) is suitable for calculating the actual number of nucleotide substitutions between two *Alu*s because the transition rate is approximately four times higher than the rate of transversion in typical *Alu* sequences (36). To calculate $\lambda$, we used members of the *Alu*Yb8 subfamily. Because >99% of *Alu*Yb8 subfamily members are human-specific (29), the dissemination of *Alu*Yb8 elements would have started around the divergence between humans and chimpanzees, and the oldest human-specific *Alu*Yb8s would have inserted into the human genome immediately after the divergence. It is therefore assumed that the age ($t_{yb8}$) of the oldest human-specific *Alu*Yb8 is equal to the divergence time ($t$) between human and chimpanzee. Thus, $\lambda$ is estimated by $\lambda = d_{yb8}/t_{yb8} = d_{yb8}/t$, where $d_{yb8}$ is the actual number of nucleotide substitutions that have accumulated in the oldest *Alu*Yb8 lineage. Using the two formulas above, we derive the following formula: $t_1 = (d_{sah}/d_{yb8})t$.

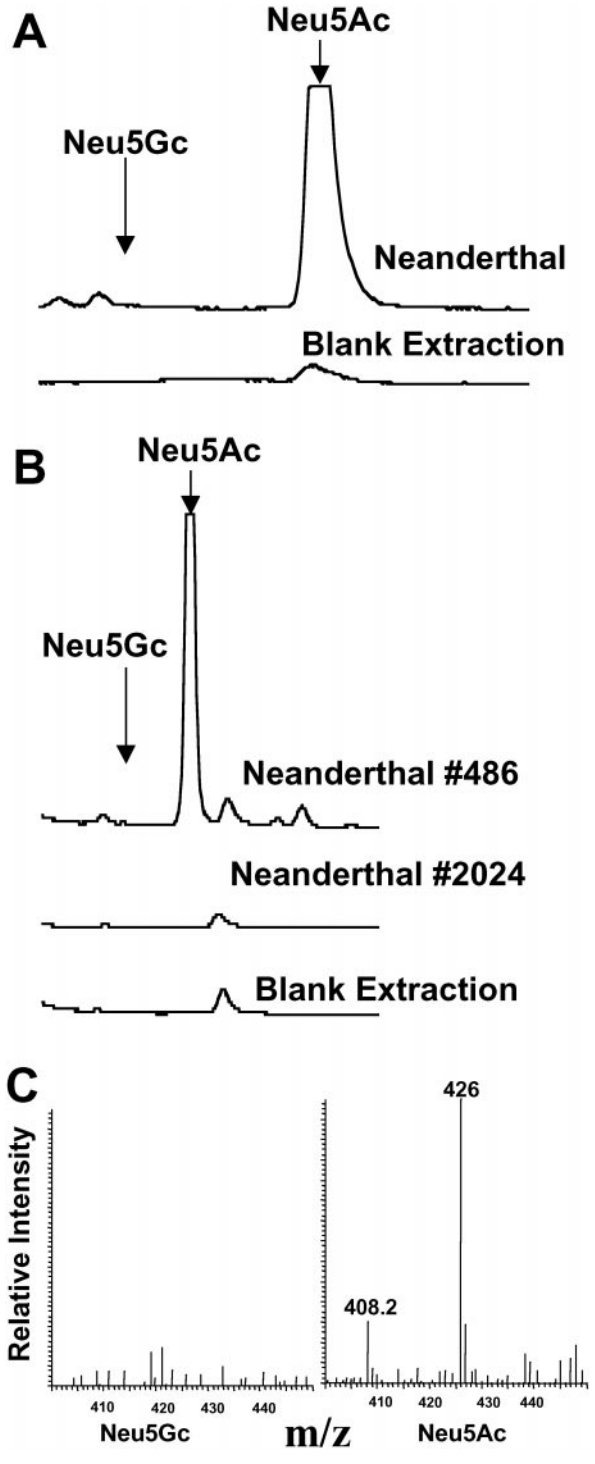


**Fig. 3.** Sias from Neandertal fossils. Sias were released, purified, and characterized as in Fig. 2. (*A*) HPLC profile of extract from the original Neandertal-type specimen. The wash through blank is a control for human handling. The column was deliberately overloaded to take the Neu5Ac peak off scale, highlighting the absence of Neu5Gc. (*B*) HPLC profile of Sias from Georgian Neandertals. The absence of Sias from sample ID no. 2024 was confirmed by MS analysis. (*C*) MS of Neu5Ac peak from sample ID no. 486 showing $m/z$ corresponding to the hydrated (426) and dehydrated (408) DMB derivatives of Neu5Ac. There were no detectable ions corresponding to Neu5Gc derivatives (442 or 424).

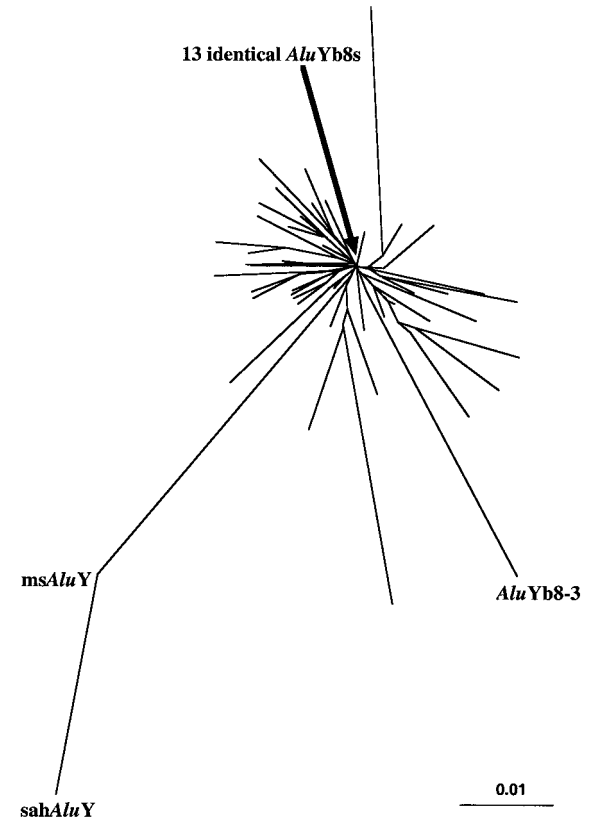


**Fig. 4.** Unrooted tree of *Alu*Yb8s. Eighty-six intact *Alu*Yb8 elements that are human-specific and fixed in human populations (29) were analyzed as described in *Materials and Methods*. sah*Alu*Y is the *Alu* that inactivated the CMAH gene, and ms*Alu*Y is most similar to it.

We have shown (18) that the CMAH inactivating sah*Alu*Y is most closely related to another human-specific *Alu*Y, ms*Alu*Y (most similar *Alu*Y). According to the *Alu* amplification model (33, 35, 37), it is likely that both sah*Alu*Y and ms*Alu*Y were disseminated from a single *Alu* source gene. We therefore prepared a phylogenetic tree with sah*Alu*Y, ms*Alu*Y, and 86 intact human-specific *Alu*Yb8s that are fixed in human populations (ref. 29 and Fig. 4). This unrooted tree shows that whereas sah*Alu*Y and ms*Alu*Y branch off from a common node, the branch length of ms*Alu*Y is almost nil. We also estimated the branch length, leading to sah*Alu*Y and ms*Alu*Y by selecting one *Alu*Yb8 member as an outgroup. These lengths are nearly the same, irrespective of the outgroup sequence (data not shown). It is therefore reasonable to assume that the sequence of ms*Alu*Y is identical to that of the source gene that generated sah*Alu*Y and ms*Alu*Y. By using sah*Alu*Y and ms*Alu*Y, we obtain $d_{\mathrm{sah}} = 0.022$. The tree also shows that most *Alu*Yb8s branch off from a single point that represents 13 identical *Alu*Yb8 sequences. This suggests that most members might have been derived from a single source gene whose sequence is identical with those of these 13 *Alu*Yb8 elements. Based on the above considerations, the *Alu*Yb8–3 is then chosen as the oldest human-specific *Alu*Yb8 because this is the most distant relative of the assumed *Alu*Yb8 source gene (see Fig. 4). Interestingly, the assumed *Alu*Yb8 source gene is identical with the YAP (Y *Alu* polymorphic) element (38). Because the YAP element is dimorphic and has been recently integrated into the human genome, it is possible that its sequence is identical with that of true *Alu*Yb8 source gene. These findings strongly support our assumption on the *Alu*Yb8 source gene. By using the assumed *Alu*Yb8 source gene and the *Alu*Yb8–3 element, we obtain $d_{\mathrm{yb8}} = 0.043$. We used the above data to derive an inactivation time of the CMAH gene of $(0.512 \pm 0.209)t$. Using the human–chimpanzee divergence time ($t$) estimated below in the CMAH molecular clock analysis, we calculate

EVOLUTION

that *Alu*-mediated inactivation of the CMAH gene occurred 2.7 ± 1.1 mya.

**Dating the Inactivation of the CMAH Gene by Molecular Clock Analysis of the CMAH (Pseudo)Gene.** Once the human CMAH gene suffered an inactivating mutation and became a pseudogene, there would no longer have been a selective pressure to maintain the appropriate functional sequence. Therefore, nonsynonymous substitutions (nucleotide substitutions that produce a change in amino acid sequence) would have begun accumulating at the neutral mutation rate, the same rate at which synonymous substitutions that do not change amino acid sequences accumulate (21). In contrast, the orthologous genes in other primates would have continued under selection pressure, favoring synonymous substitutions over nonsynonymous substitutions. We have used this expected difference in mutation rates to calculate when the human CMAH mutation might have occurred.

When we assume that the neutral mutation rate is $k$, the nonsynonymous substitution rate in a functional gene can be written as $f_N k$, where $f_N$ is the fraction of neutral substitutions. The value $f_N$ is less than or equal to unity and inversely relates to the degree of functional constraint as $1 - f_N$. This $f_N$ varies from gene to gene depending on the extent of functional importance. The more important the gene, the smaller its $f_N$. However, once it becomes a pseudogene, $f_N$ becomes unity and the rate increases to $k$. This increase in the substitution rate at nonsynonymous sites is thus used to estimate the time of inactivation of the CMAH gene.

For the human CMAH gene, we define $t_1$ as the time elapsed since the inactivation and $t$ as the time after the human and chimpanzee lineages diverged from each other. During $t - t_1$, the gene is functional so that the nonsynonymous sites evolve at a rate of $f_N k$, whereas during $t_1$ the rate is $k$ because of the loss of function. Therefore, the per-site number of nonsynonymous substitutions in the human lineage is expected to be $f_N k(t - t_1) + kt_1$. If we know $f_N$, $k$, and $t$, we can estimate the value of $t_1$.

The $f_N$ value for the functional CMAH gene is estimated as a ratio of the per site number of nonsynonymous substitutions to that of synonymous ones in non-human primate genes. To obtain an estimate of $f_N$, we directly sequenced multiple CMAH cDNAs from all of the great apes and determined the total number of nonsynonymous and synonymous substitutions ($n_N$ and $n_S$) along each lineage by the maximum parsimony method (Fig. 5). In all, nine sequences were used for this analysis (see *Supporting Text* for alignments of all of the sequences). In our original work comparing human and ape hydroxylase cDNAs, we noted some less common alternate forms of the human pseudogene transcript, including some insertions and deletions in the 3′ region. A similar diversity was not seen in cDNAs from chimpanzee cells. We speculated that these human transcript variations reflect random degeneration in splicing precision of a nonfunctional gene. We have not pursued this issue further. For this particular study, we chose to sequence only the longest dominant human transcript that could be fully aligned with the great ape cDNA sequences. The 5′ and 3′ untranslated regions were excluded from this analysis because of a large deletion in humans, bonobos, and gorillas. The 92-bp region deleted in humans was also excluded from the analysis, as was a short segment at the end of the human transcripts that showed marked differences between the two human samples. The average values become $n_S = 17$ and $n_N = 10$. The number of synonymous and nonsynonymous sites is $l_S = 393$ and $l_N = 1{,}236$, respectively. We thus have $f_N = n_N l_S / n_S l_N = 0.19$, suggesting rather strong selective constraint against the nonsynonymous substitutions in the functional CMAH gene (22).

The neutral mutation rate $k$ is estimated as the synonymous substitution rate under the assumption of no constraint on synonymous substitutions. To obtain $k$, we use the average number of synonymous substitutions along the lineage leading to the human and the great apes after their divergence of orangutans. The average number is 7.9. If we assume that the divergence time of orangutans is 13 mya, we have $k = 7.9/393/(13 \times 10^6) = 1.5 \times 10^{-9}$ per site per year. Using this $k$ and the average number (3.1) of synonymous substitutions in the human and chimpanzee lineages, we estimate the divergence time $t$ between the human and the chimpanzee as $t = 3.1/393/k = 5.3$ mya. Finally, the number of nonsynonymous substitutions in the human CMAH gene is given by $6/1{,}236 = 4.9 \times 10^{-3}$. Solving the equation of $f_N k\,(t - t_1) + kt_1 = 4.9 \times 10^{-3}$ for $t_1$ with the estimates of $f_N$, $k$, and $t$, we obtain $t_1 = [(4.9 \times 10^{-3})/k - f_N t]/(1 - f_N) = [(4.9 \times 10^{-3})/(1.5 \times 10^{-9}) - 0.19 \times 5.3 \times 10^6]/(1 - 0.19) = 2.8 \times 10^6$ years.

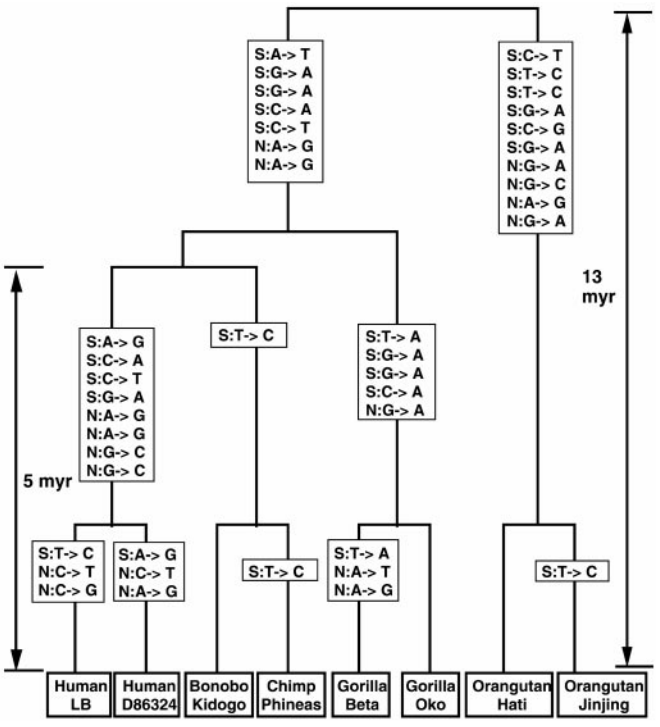


**Fig. 5.** Phylogenetic tree of hydroxylase sequences built by the maximum parsimony method. Individual nucleotide substitutions on each branch (based on ancestral sequence deduced from rhesus monkey) are indicated and designated S or N, depending on whether the nucleotide substitution is a synonymous or nonsynonymous change, respectively (chronological order is not implied in the order of each listing). The rhesus monkey sequence (GenBank accession no. AB013814) was used as an outgroup to deduce an ancestral sequence of orangutan, gorillas, chimpanzees, and humans. A change that induced a new stop codon at the C terminus of the gorilla Beta sequence was observed and excluded from the analysis.

## Discussion

We have shown that Sias can be purified from bones and from ancient fossils and specifically identified by DMB derivatization, HPLC separation, and mass spectrometry. This approach showed that Neandertals, like present-day humans, did not express Neu5Gc. This indicates that the hydroxylase was inactivated before the time of our common ancestor with Neandertals, likely about 500,000–600,000 years BP (19). Our failure to recover Sias from Javanese *H. erectus* samples and faunal fossils from Africa suggests that it may not be worthwhile to further study samples from tropical and semitropical areas, at least with presently available methods. Unfortunately, these are the regions where most of the more ancient bipedal hominids have been discovered. Thus, we used independent molecular approaches to estimate that the inactivating mutation in the

hydroxylase occurred just over 2 mya. No single one of these approaches is free of potential error. However, when taken together, they allow us to reasonably date the inactivation of the hydroxylase to the period just before the appearance of *Homo*.

Since our last common ancestor with the chimpanzee, significant changes in teeth and jaw structure, bone-frame size, locomotion style, ontogeny period, and brain size have occurred during human evolution (39). Fossil records indicate that australopiths had large teeth, small brains, and skeletal structures adapted to a more terrestrial lifestyle. Whereas their pattern of locomotion is likely to have included some arboreality, they were fully bipedal and did show many signs of advanced bipedal bone structure. The 3.5 million-year-old Laetoli footprints of *Australopithecus afarensis* show human-like proportions, arches, heel strike, and convergent big toes (40). Comparative anatomical analysis of human, apes, and fossil hominids indicate that *A. afarensis* had significant features of bipedality (39, 41). Therefore, upright locomotion was acquired at the early stage of bipedal hominid evolution, and the inactivation of the hydroxylase could not have contributed to its establishment.

Starting at ≈2.1–2.2 mya in the bipedal hominid clade (relatively soon after the approximated time of the CMAH mutation), one begins to see significant increases in brain size relative to body size (39, 41). Whereas the brains of early australopiths are in the size range of modern great apes, an increase in brain size accelerated episodically in later bipedal hominids, until the time of the first archaic *H. sapiens* about 400,000–500,000 years ago. This increase in brain mass relative to body size (encephalization) is believed to be at least partly related to secondary altriciality (incompletely developed and helpless state of the newborn), and the relative increase of brain size after birth (39, 42, 43). Unlike most primate brains that stop growing relatively soon after birth, human brains continue to grow for some time postnatally, at a rate similar to the body growth rate. Prenatal growth of the head is limited by the size of the mother's birth canal. Comparative anatomical analysis of humans and our earlier fossil relatives had previously suggested that secondary altriciality accompanied the appearance of brains of $>750$ cm$^3$ (39), although a recent study suggested a later origin of secondary altriciality (43).

It is premature to speculate much regarding possible roles of CMAH gene inactivation in the acquisition of human-specific features. It is intriguing to note that, in all mammals studied so far (1), including the chimpanzee (4), the amount of Neu5Gc in the brain is always very low, no matter what the levels are in other organs of the body. This seems to be explained by selective down-regulation of CMAH gene expression in the mammalian brain (44). A potentially testable hypothesis is that the low levels of residual brain Neu5Gc in other mammals somehow limited brain expansion and that the human CMAH mutation released our ancestors from such a constraint. We are therefore studying the effects of Neu5Gc overexpression in the mouse brain and exploring how Neu5Gc expression might affect the biology of neural cells and molecules.

We gratefully acknowledge Susan Anton (Rutgers University) and Teuku Jacob (Gadjah Mada University) for helping to arrange contacts among the investigators. We thank Pascal Gagneux for helpful discussions and Yen-Liang Chen for helpful technical suggestions. This work was supported by U.S. Public Health Service Grant R01-GM323373 and by the G. Harold and Leila Y. Mathers Charitable Foundation.

1. Schauer, R. (1982) *Sialic Acids: Chemistry, Metabolism and Function, Cell Biology Monographs* (Springer, New York), Vol. 10.
2. Angata, T. & Varki, A. (2002) *Chem. Rev.* **102,** 439–470.
3. Varki, A. (2002) *Yearbook Phys. Anthropol.* **44,** 54–69.
4. Muchmore, E. A., Diaz, S. & Varki, A. (1998) *Am. J. Phys. Anthropol.* **107,** 187–198.
5. 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.
6. Irie, A., Koyama, S., Kozutsumi, Y., Kawasaki, T. & Suzuki, A. (1998) *J. Biol. Chem.* **273,** 15866–15871.
7. Bouhours, D. & Bouhours, J. F. (1983) *J. Biol. Chem.* **258,** 299–304.
8. Muchmore, E., Varki, N., Fukuda, M. & Varki, A. (1987) *FASEB J.* **1,** 229–235.
9. Muchmore, E. A. (1992) *Glycobiology* **2,** 337–343.
10. 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.
11. Collins, B. E., Fralich, T. J., Itonori, S., Ichikawa, Y. & Schnaar, R. L. (2000) *Glycobiology* **10,** 11–20.
12. Ito, T., Suzuki, Y., Suzuki, T., Takda, A., Horimoto, T., Wells, K., Kida, H., Otsuki, K., Kiso, M., Ishida, H. & Kawaoka, Y. (2000) *J. Virol.* **74,** 9300–9305.
13. Ito, T., Suzuki, Y., Mitnaul, L., Vines, A., Kida, H. & Kawaoka, Y. (1997) *Virology* **227,** 493–499.
14. 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.
15. Delorme, C., Brüssow, H., Sidoti, J., Roche, N., Karlsson, K. A., Neeser, J. R. & Teneberg, S. (2001) *J. Virol.* **75,** 2276–2287.
16. Kyogashima, M., Ginsburg, V. & Krivan, H. C. (1989) *Arch. Biochem. Biophys.* **270,** 391–397.
17. Schlenzka, W., Shaw, L., Kelm, S., Schmidt, C. L., Bill, E., Trautwein, A. X., Lottspeich, F. & Schauer, R. (1996) *FEBS Lett.* **385,** 197–200.
18. Hayakawa, T., Satta, Y., Gagneux, P., Varki, A. & Takahata, N. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 11399–11404.
19. Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking, M. & Paabo, S. (1997) *Cell* **90,** 19–30.
20. Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M. & Paabo, S. (2001) *Nat. Rev. Genet.* **2,** 353–359.
21. Li, W.-H. (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
22. Gabunia, L. & Vekua, A. (1990) *L'Anthropologie (Paris)* **94,** 643–650.
23. Swisher, C. C., Rink, W. J., Anton, S. C., Schwarcz, H. P., Curtis, G. H., Suprijo, A. & Widiasmoro (1996) *Science* **274,** 1870–1874.
24. Ward, C. V., Leakey, M. G. & Walker, A. W. (2001) *J. Hum. Evol.* **41,** 255–376.
25. Brown, F., Harris, J., Leakey, R. & Walker, A. (1985) *Nature (London)* **316,** 788–792.
26. Leakey, R. E. & Walker, A. (1988) *Am. J. Phys. Anthropol.* **76,** 1–24.
27. Manzi, A. E., Diaz, S. & Varki, A. (1990) *Anal. Biochem.* **188,** 20–32.
28. Klein, A., Diaz, S., Ferreira, I., Lamblin, G., Roussel, P. & Manzi, A. E. (1997) *Glycobiology* **7,** 421–432.
29. Carroll, M. L., Roy-Engel, A. M., Nguyen, S. V., Salem, A. H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M., *et al.* (2001) *J. Mol. Biol.* **311,** 17–40.
30. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
31. Kimura, M. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 454–458.
32. Smith, C. I., Chamberlain, A. T., Riley, M. S., Cooper, A., Stringer, C. B. & Collins, M. J. (2001) *Nature (London)* **410,** 771–772.
33. Schmid, C. W. (1998) *Nucleic Acids Res.* **26,** 4541–4550.
34. 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.
35. Schmid, C. & Maraia, R. (1992) *Curr. Opin. Genet. Dev.* **2,** 874–882.
36. Kapitonov, V. & Jurka, J. (1996) *J. Mol. Evol.* **42,** 59–65.
37. Zietkiewicz, E., Richer, C., Makalowski, W., Jurka, J. & Labuda, D. (1994) *Nucleic Acids Res.* **22,** 5608–5612.
38. Hammer, M. F. (1994) *Mol. Biol. Evol.* **11,** 749–761.
39. Wood, B. & Collard, M. (1999) *Science* **284,** 65–66.
40. Leakey, M. D., Hay, R. L., Curtis, G. H., Drake, R. E. & Jackes, M. K. (1976) *Nature (London)* **262,** 460–466.
41. McHenry, H. M. (1994) *Proc. Natl. Acad. Sci. USA.* **91,** 6780–6786.
42. Jones, S., Martin, R. & Pilbeam, D. (1992) *The Cambridge Encyclopedia of Human Evolution* (Cambridge Univ. Press, New York).
43. Dean, C., Leakey, M. G., Reid, D., Schrenk, F., Schwartz, G. T., Stringer, C. & Walker, A. (2001) *Nature (London)* **414,** 628–631.
44. Kawano, T., Koyama, S., Takematsu, H., Kozutsumi, Y., Kawasaki, H., Kawashima, S., Kawasaki, T. & Suzuki, A. (1995) *J. Biol. Chem.* **270,** 16458–16463.