Evolutionary Origin and Human-Specific Expansion of a Cancer/Testis Antigen Gene Family

Qu Zhang*,‡,[1] and Bing Su*,[2]

[1]Department of Human Evolutionary Biology, Graduate School of Art and Science, Harvard University

[2]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

‡Present address: Pioneer Hi-Bred International, A DuPont Business, Johnston, IA

***Corresponding author:** E-mail: quzhang@post.harvard.edu; sub@mail.kiz.ac.cn.

Associate editor: Meredith Yeager

Abstract

Cancer/testis (CT) antigens are encoded by germline genes and are aberrantly expressed in a number of human cancers. Interestingly, CT antigens are frequently involved in gene families that are highly expressed in germ cells. Here, we presented an evolutionary analysis of the *CTAGE* (cutaneous T-cell-lymphoma-associated antigen) gene family to delineate its molecular history and functional significance during primate evolution. Comparisons among human, chimpanzee, gorilla, orangutan, macaque, marmoset, and other mammals show a rapid and primate specific expansion of *CTAGE* family, which starts with an ancestral retroposition in the haplorhini ancestor. Subsequent DNA-based duplications lead to the prosperity of single-exon *CTAGE* copies in catarrhines, especially in humans. Positive selection was identified on the single-exon copies in comparison with functional constraint on the multiexon copies. Further sequence analysis suggests that the newly derived *CTAGE* genes may obtain regulatory elements from long terminal repeats. Our result indicates the dynamic evolution of primate genomes, and the recent expansion of this CT antigen family in humans may confer advantageous phenotypic traits during early human evolution.

***Key words:* *CTAGE* gene family, cancer/testis antigen, positive selection, gene duplication, retroposition, human-specific expansion.**

Introduction

It is widely accepted that gene duplication during evolution is a major mechanism for new gene creation and gene family expansion (Vandepoele et al. 2005; Bosch et al. 2007). One of the main duplication forms is segmental duplications (SDs) with genomic regions larger than 1 kb and a sequence identity larger than 90% (Bailey et al. 2002; Bailey and Eichler 2006), derived via nonallelic homologous recombination (NAHR) (Lupski 1998) or nonhomologous end joining (Linardopoulou et al. 2005; Conrad and Hurles 2007; Kim et al. 2008). Moreover, a substantial number of studies have shown that new genes can also arise through retroposition, an RNA-based duplication mechanism (Kaessmann et al. 2009; Kaessmann 2010; Zhang 2013). Collectively, these genomic changes resulted in a large number of primate-specific genes or expansions of gene families (Vandepoele et al. 2005; Bosch et al. 2007; Andres et al. 2008; Wainszelbaum et al. 2008; Das et al. 2010; Arroyo et al. 2012; Dennis et al. 2012; Giannuzzi et al. 2013), some of which are reported to be related to phenotypic traits (Das et al. 2010), highlighting their important role in primate evolution.

The similarity between germ-cell development and cancer cell development has been long noticed (Simpson et al. 2005), and the discovery of the cancer/testis (CT) antigens is of particular interest, as they present only in germ cells, trophoblast, and tumor cells. Based on this observation, it has been proposed that the silenced gametogenic program in somatic cells is reactivated in tumors, which is considered as one of the major contributors of tumorigenesis (Simpson et al. 2005). Intriguingly, a majority of CT antigens are located on the X chromosome (CT-X antigens) to form recently expanded gene clusters (Simpson et al. 2005; Hofmann et al. 2008) and evolve rapidly (Kouprina et al. 2004; Stevenson et al. 2007). Although autosome-linked CT genes are mainly found as single-copy genes throughout the genome (Simpson et al. 2005; Fratta et al. 2011), and many may function in meiosis. Abnormal expression of autosome-linked genes in cancer cells might result in chromosome segregation problems and aneuploidy (Simpson et al. 2005). Despite that, limited attention has been paid on the evolution of these genes compared with their X-linked counterparts, though such information could shed light on their emergence and possible roles. Thus in this study, we evolutionarily characterized the human cutaneous T-cell-lymphoma-associated antigen (*CTAGE*) gene family, a CT antigen family not located on chromosome X. Expression analysis of tumor tissues and normal controls have found that the expression of *CTAGE1* can be detected in approximately 35% of the tested cutaneous T-cell lymphoma tumor specimens (Eichmuller et al. 2001), and *CTAGE4* and *CTAGE5* could also be detected in various tumor tissues and cell lines (Usener et al. 2003), suggesting their important role as putative tumor-specific immunotherapy targets. We analyzed the organization and gene structure of *CTAGE* copies in primates and other mammals

Article

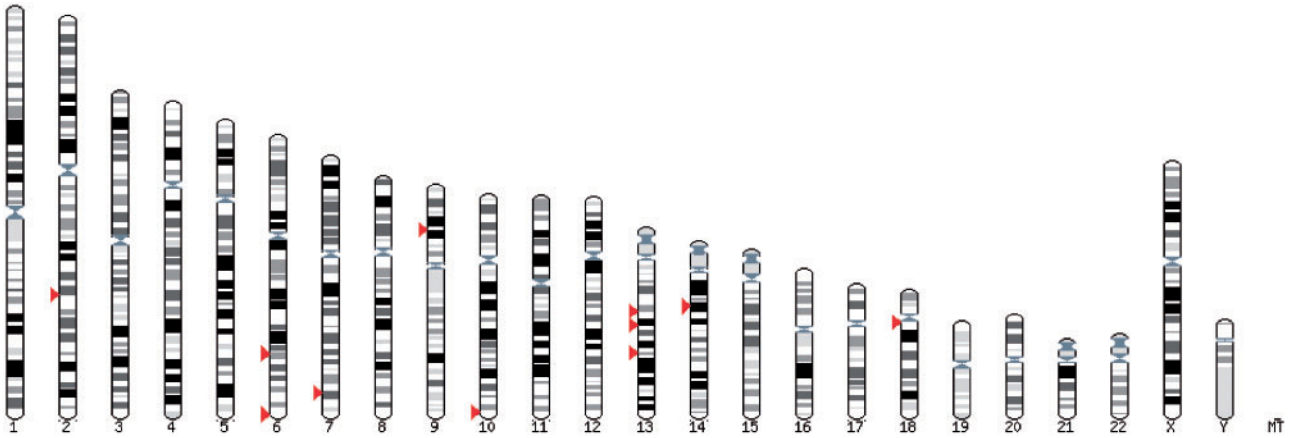

FIG. 1. Chromosomal distribution of human *CTAGE* copies. Each *CTAGE* copy is denoted by the triangle, and the multiexon *CTAGE5* gene is on chromosome 14. There are four copies in a small region on chromosome 7, and the triangles are mainly overlapped, so are the two copies on chromosome 13. See supplementary table S1, Supplementary Material online, for detail positions of each *CTAGE* copy.

and detected a rapid expansion in catarrhines via both DNA- and RNA-based duplications. Furthermore, several functional copies were found to be derived from human-specific expansions. We also identified signals of positive selection on recently derived coding *CTAGE* genes. The rapid evolution and their important role in tumor development of these CT antigen genes may reflect evolutionary trade-off between reproductive advantages and cancer risk.

Results

Genomic Organization of Primate *CTAGE* Family

CTAGE members were identified by searching the coding sequence of human *CTAGE9* gene against primate genomes via the BLAST-Like Alignment Tool (BLAT) (Kent 2002), requiring at least 70% nt identity. There are 15 distinct *CTAGE* copies (named as human 1–15 or H1–H15) in the human genome (GRCh37/hg19), including four copies clustered on chromosome 7, four copies on chromosome 13, two copies on chromosome 6, and single copies on chromosomes 2, 9, 10, 14, and 18, respectively (fig. 1 and supplementary table S1, Supplementary Material online). Among these copies, only one copy on chromosome 14 has multiple exons, and all others have a single exon. Five copies have intact open reading frames (ORFs) and Ensembl gene annotations, while the remaining contain disrupted reading frames. The chimpanzee genome (panTro4) has 11 *CTAGE* members (denoted as chimpanzee 1–11 or C1–C11), and all copies except the one on chromosome 14 have a single exon. The recently released gorilla genome (gorGor3) contains the smallest *CTAGE* family in apes, including eight single-exon copies (denoted as gorilla 1–8 or G1–G8) and one multiexon copy (G9) on chromosome 14. Because of sequencing gaps, two copies (G6 and G8) have incomplete sequences (supplementary table S1, Supplementary Material online). There are 12 copies (orangutan 1–12 or O1–O12) in the orangutan genome (ponAbe2) and 11 copies (rhesus 1–12 or R1–R12) in the rhesus macaque genome (rheMac3). In rhesus macaque, three copies (R8, R9, and R10) have sequence gaps. We also identified five distinct copies (marmoset 1–5 or M1–M5) in marmoset genome assembly (calJac3). In contrast, in nonprimate mammals including mouse (mm10), rat (rn5), dog (canFam3), and cow (bosTau7), only one multiexon copy can be identified. It should be noted that the individual multiexon copies in both primates and other mammals are resided on homologous chromosomes, indicating that the multiexon copy should be present in the common ancestor.

Evolutionary History of the *CTAGE* Family

To resolve the phylogenetic relationship among various *CTAGE* family members in primates, a maximum-likelihood tree was reconstructed using all primate copies except single-exon copies in marmosets, which are more diverse and cause problems in sequence alignment. Multiexon copies in additional primates (gibbon and bushbaby) and nonprimate mammals (mouse, rat, cow, and dog) were also included to depict a broad picture of the evolution of *CTAGE* family. Because all the nonprimate mammals have only one multiexon copy, this implies that the ancestral *CTAGE* copy in primates is a multiexon gene, and single-exon copies must be introduced by RNA-based retroposition. The topology of the phylogenetic tree revealed that a retroposition event of the multiexon copy occurred on the branch of the haplorhini ancestor (fig. 2), and further inter or intrachromosomal duplications led to the burst of the *CTAGE* family. Additionally, no further retropositions of the multiexon copy could be inferred, otherwise we expect certain single-exon copies will be clustered together with the multiexon copy. Notably, both DNA-based and RNA-based duplications could contribute to the prosperity of these single-exon copies, because the retroposed copy of an intronless gene is indistinguishable to its segmentally duplicated copies in genic regions. To obtain a thorough understanding of the *CTAGE* evolution, we further compared 500-bp flanking regions in each direction for single-exon *CTAGE* copies, and various degrees of homology could

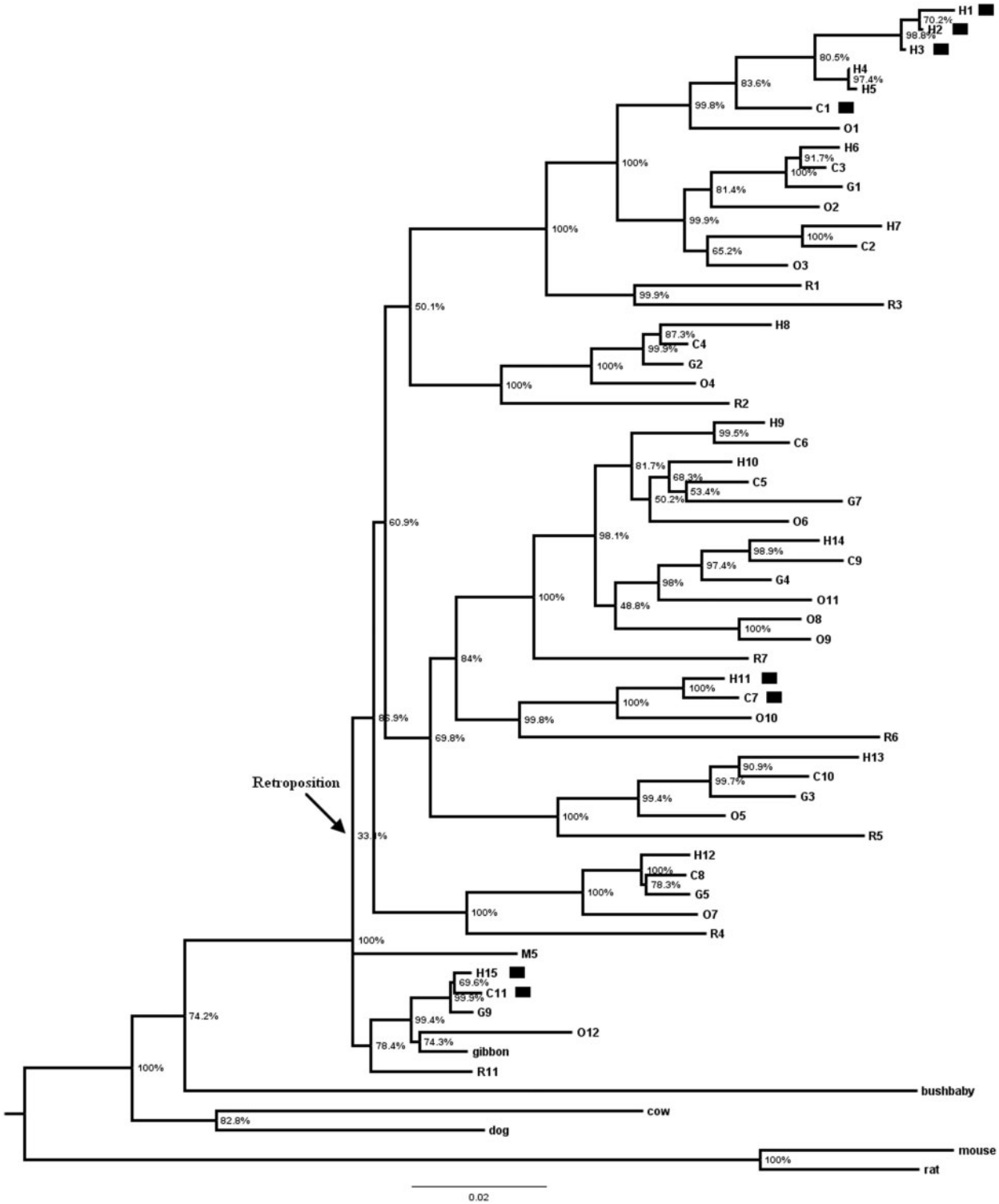

Fig. 2. *CTAGE* family phylogeny. Regions homologous to the coding sequence of human *CTAGE9* (H1) were aligned by MUSCLE. Phylogenetic tree was next constructed using maximum-likelihood method in MEGA 5. The branch length is scaled to the number of substitutions per site. Bootstrap values from 1,000 replicates were shown next to each node. The solid arrow denotes the branch that the retroposition event most likely occurred. Genes showed in table 1 were denoted by a black bar next to them. For *CTAGE* copy IDs, H is short for human, C for chimpanzee, G for gorilla, O for orangutan, R for rhesus monkey, and M for marmoset.

be found between different pairs, highlighting that single *CTAGE* copies may be derived from DNA-based duplication. Interestingly, we observed three rounds of duplications in the human lineage, resulting in three human-specific copies with intact ORFs and two pseudogenes.

Expression Profiling of *CTAGE* Members

Five out of 15 human loci have intact ORFs and are annotated as protein-coding genes. However, it has been argued that a large proportion of human annotated genes are false positives

by retaining an intact ORF by chance (Clamp et al. 2007). To test if these five *CTAGE* genes are annotation errors, we examined the expression pattern in different tissues by using RNA-Seq data. Expression evidence for all five genes was found in various tissues (table 1) and the ancestral multiexon gene *CTAGE5* (H15) showed a broader tissue distribution and higher expression level compared with the others. In contrast, a low and testis-biased expression was found for single-exon copies. This is consistent with previous findings that newly derived genes tend to be expressed in testis and have lower expression than their parental copies (Betran et al. 2002; Emerson et al. 2004; Levine et al. 2006; Vinckenbosch et al. 2006). The same pattern was also confirmed in chimpanzees, as the multiexon copy was expressed in all tissues investigated, whereas the single-exon copies were only expressed in testis and a few other tissues. For the remaining ten *CTAGE* pseudogenes, a pattern of strong testis-biased expression was also detected (supplementary table S2, Supplementary Material online), which is consistent with the known high transcriptional activity in testis.

Table 1. Tissue Expression of Human-Coding *CTAGE* Genes and Their Orthologous Genes in Other Primates.

Tissue	H15	C11	H1	H2	H3	C1	H11	C7
Brain	6.03	3.48	0.00	0.00	0.00	0.00	0.03	0.00
Cerebellum	0.00	2.21	0.00	0.00	0.00	0.04	0.00	0.05
Heart	0.00	4.44	0.00	0.00	0.00	0.00	0.00	0.00
Liver	24.60	12.63	0.00	0.02	0.00	0.00	0.00	0.00
Kidney	18.15	12.68	0.00	0.00	0.03	0.00	0.00	0.00
Testis	0.00	5.42	0.93	0.28	0.04	2.03	0.02	5.05

NOTE.—The number of FPKM estimated by TopHat are shown. H is short for human and C for chimpanzee. FPKM, fragments mapped.

Evolutionary Analysis of *CTAGE* Genes

We next sought to study the evolutionary pattern of coding *CTAGE* copies. To reconstruct an accurate phylogenetic relationship, multiexons coding *CTAGE* copies and single-exon coding copies were aligned separately to derive phylogenetic trees, respectively. Two trees were then concatenated, and the rodent group was manually positioned according to the species tree (fig. 3). Maximum-likelihood models were further applied to estimate evolutionary rates for branches and test different hypotheses (R6 was excluded from following analysis because it misses 58 bp starting from the start codon and may

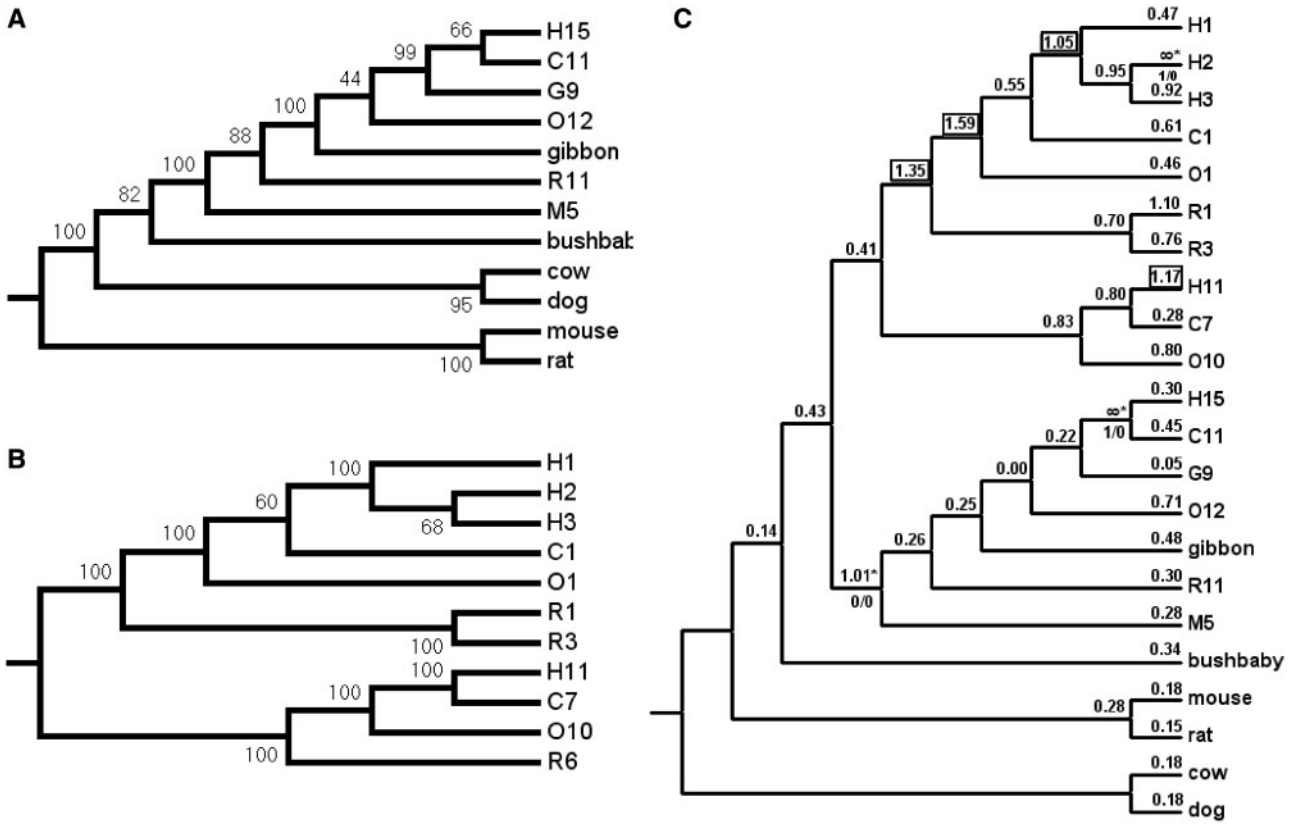

FIG. 3. Positive selection on *CTAGE* family during primate evolution. To obtain a clear phylogenetic relationship of coding *CTAGE* copies, maximum-likelihood trees were constructed using coding sequences of multiexon genes (*A*) and single-exon genes (*B*), respectively. The final phylogenetic tree for *CTAGE* coding genes (*C*) was created by combining (*A* and *B*) and placed the branch of cow and dog as the outgroup, as concordant with the species tree. The estimated $\omega = d_N/d_S$ based on free branch model is shown above each branch. $\omega > 1$ is a possible indicator of positive selection; however, several branches showed a large ω value due to no synonymous changes, even the number of nonsynonymous changes are small too. These branches were indicated by asterisk, with the number of nonsynonymous and synonymous changes shown below. Other branches with $\omega > 1$ were highlighted by red boxes, which suggest targets of positive selection. For *CTAGE* copy ids, H is short for human, C for chimpanzee, G for gorilla, O for orangutan, R for rhesus monkey, and M for marmoset.

not maintain the intact ORF, making the sequence alignment less informative). Using a "free-ratio" model in codeml in PAML, we estimated d_N/d_S (ω) ratios across all branches and found that the clade of multiexon copies tend to have a lower ω than the clade of single-exon copies (fig. 3C). To further test this, a "one-ratio" model was compared with a "two-ratio" model with different ω for single- and multiexon *CTAGE* genes. The result showed that the two-ratio model was significantly favored (P value = 7.8×10^{-21}, χ^2 test) with an elevation of ω in single-exon copies (0.793) compared with multiexon genes (0.261), suggesting multiexon copies may be under selective constraint. When we compared the two-ratio model with an alternative two-ratio model with a fixed $\omega = 1$ for multiexon copies, a strong signal of negative selection was found (P value $< 9.7 \times 10^{-78}$), confirming the strong functional constraint on multiexon *CTAGE* genes. It is also notable that several branches in the single-exon clade have $\omega > 1$, a sign of positive selection. We therefore performed the branch-site test and found the model allowing $\omega > 1$ for single-exon clade is significantly better than the model with $\omega = 1$ for that clade (P value = 0.0038), indicating that single-exon *CTAGE* genes evolve under positive selection. We noted that a new dynamic programing procedure was introduced to search for optimal models among a large number of possibilities (Zhang et al. 2011). However, it may not be relevant to the analysis in this study, as a priori hypothesis with clearly biological significance was used.

The branch-site model can also be used to infer the posterior probability that a codon is under positive selection by the Bayes empirical Bayes procedure (Yang et al. 2005, 2009; Nozawa et al. 2009). Here, a total of 37 positively selected codons were identified with a posterior probability over 0.95 (table 2). Among them, one is located in the transmembrane domain, nine in the coiled coil domain, and four in the proline-rich domain. As a number of studies suggest that the coiled coil motif has an important role in tumor development (Maeyama et al. 2011; Saito et al. 2011), positively selected sites in the coiled coil region are intriguing and may result in distinct functions in tumorgenesis. However, it should also be noticed that the false-positive predictions cannot be excluded, therefore further functional analysis is critical to reveal the true target site of positive selection.

Table 2. Positively Selected Codons Inferred by BEB Procedure.

Codon Position	Encoded Amino Acid	Posterior Probability	Domain
18	I	0.957	Transmembrane
61	I	0.956	
71	E	0.952	
101	D	0.952	Coiled coil
148	T	0.990*	Coiled coil
156	R	0.956	Coiled coil
172	T	0.989	Coiled coil
221	Q	0.971	
226	E	0.951	
248	P	0.956	
258	A	0.954	
263	V	0.955	
274	H	0.990*	
300	E	0.989	Coiled coil
309	E	0.954	Coiled coil
317	E	0.958	Coiled coil
341	I	0.955	Coiled coil
378	R	0.987	Coiled coil
390	N	0.986	
402	E	0.956	
406	C	0.986	
409	S	0.952	
425	Q	0.957	
433	R	0.955	
439	G	0.983	
479	P	0.964	
480	V	0.958	
484	R	0.988	
485	R	0.971	
490	P	0.954	
504	D	0.983	
511	E	0.987	
541	E	0.953	
558	S	0.985	Proline rich
560	P	0.957	Proline rich
611	I	0.967	Proline rich
630	N	0.988	Proline rich

NOTE.—BEB, Bayes empirical Bayes.
*Sites with a probability of 0.99 or large.

Human-Specific Expansion of *CTAGE* Family

The reconstruction of the phylogenetic tree showed a recent expansion of the *CTAGE* family in humans, resulting in two pseudogenes (*CTAGE15P* or H4, and *CTAGE6P* or H5) and three intact genes (*CTAGE9* or H1, *CTAGE4* or H2, and *CTAGE8* or H3). By comparing the flanking sequences to chimpanzee, we inferred *CTAGE15P* in the ancestor of human and chimpanzee. Further comparison of sequence similarity among *CTAGE4*, *CTAGE8*, and *CTAGE9* revealed a most likely evolutionary scenario: A 7-kb *CTAGE8* locus (chr7: 143961987–143968904) was first derived probably by SD, and then a 65-kb region involving *CTAGE15P* was duplicated and generated the *CTAGE6P* locus (chr7: 143439416–143504793). The *CTAGE9* locus (chr6: 132019328–132037070) was duplicated from at least an 18-kb genomic segment containing *CTAGE8* locus and the flanking *OR2A7* gene (fig. 4A). The *CTAGE4* locus (chr7: 143875684–143884033) was derived by a recent tandem duplication from at least a 9-kb region involving *CTAGE8*. Later a deletion with unknown size occurred in the 3′ downstream sequences of *CTAGE8*, leading to a 16-kb homologous region between *CTAGE8* and *CTAGE9* loci, a 7-kb homologous region between *CTAGE8* and *CTAGE4* loci, as well as a 9-kb homologous region between *CTAGE4* and *CTAGE9* loci (fig. 4A). Because gene duplications could also be due to assembly errors (Zhang and Backstrom 2014), we further investigated four clearly defined insertion/deletion sites (chr6: 132019328, chr6: 132037070, chr7: 143884033, and chr7: 143969444). First, a 201-bp human sequence

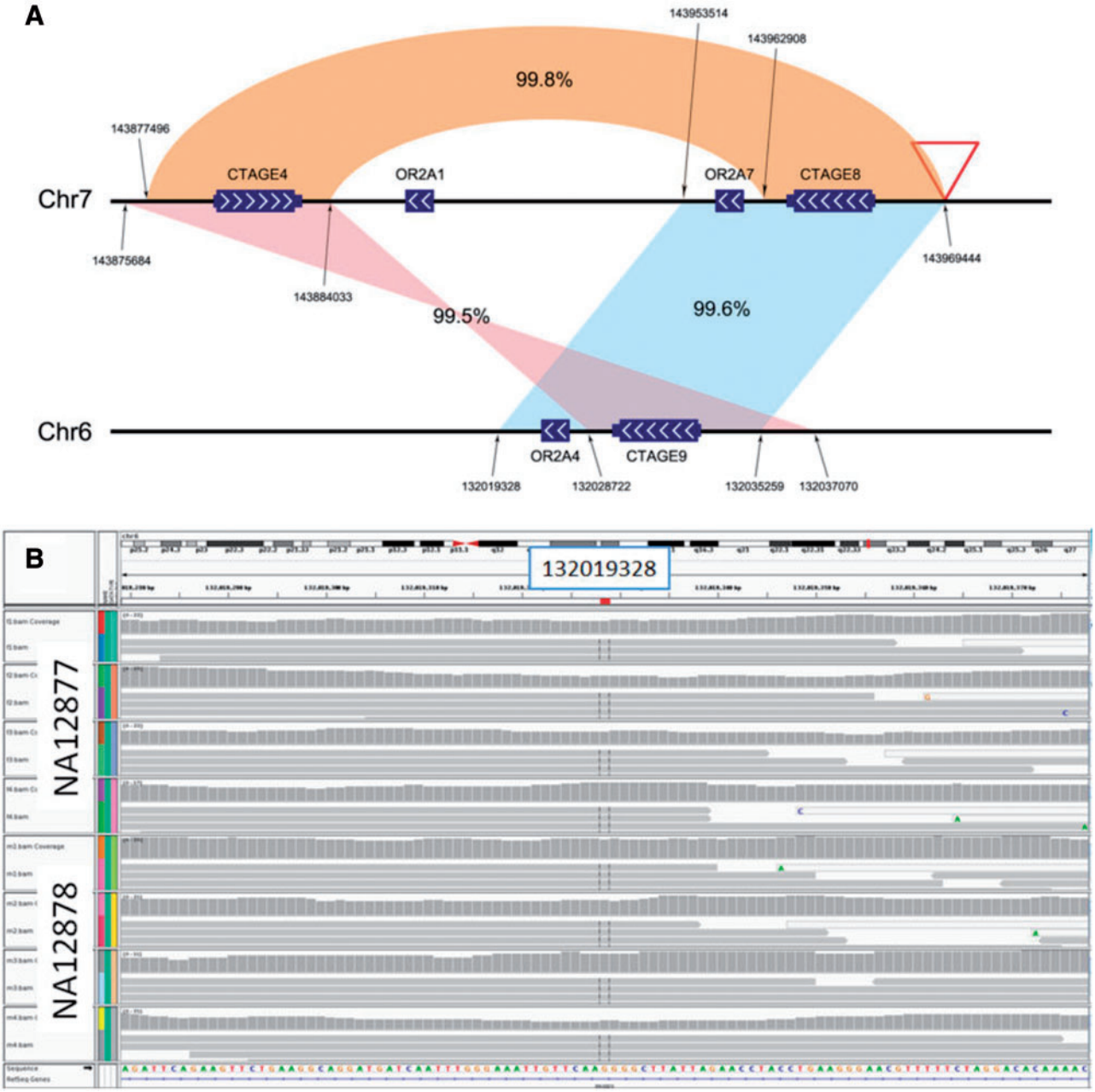

FIG. 4. Evolutionary scenario of human specific *CTAGE* members. (*A*) Genomic organization of human-specific *CTAGE* copies. Exons are indicated by solid blue boxes, whereas white arrows within exons denote transcription directions. Small solid blue boxes at either end of exons are untranslated regions. Paralogous sequences are linked by colored ribbons, with sequence identity shown. Numbers close to black arrows denotes ordinates for starts and ends of paralogous sequences. The red open triangle highlights the deletion (see main text for details). (*B*) Deep-sequencing reads confirm all junctions. Red rectangles highlight the junction position, gray horizontal bars denote short reads, and gray vertical bars denote the sequence coverage for each base. Mismatches between reads and the reference genome are showed by colored characters within reads. The reference genome sequence is showed at the bottom.

centering on each site was retrieved and mapped to genomes of nonhuman apes (chimpanzee, gorilla, and orangutan), but no hit was found to overlap these junction sequences, confirming their absence in other primates. Second, deep-sequencing data from two unrelated human individual genomes were mapped to the human genome reference, and abundant high-quality reads were found spanning the junction sites (fig. 4*B*), suggesting they are not human assembly errors. Third, deep-sequencing data from seven central chimpanzee (*Pan troglodytes troglodytes*) individuals (Prufer et al. 2012) were mapped to these four junction sequences, but no alignment was found. All these lines of evidence clearly indicate that *CTAGE4*, *CTAGE8*, and *CTAGE9* are human bona fide genes, and the deletion close to *CTAGE8* also occurred after the split of human and chimpanzee.

Molecular Mechanism of Human-Specific Expansion

One question of particular interest is what mechanism causes the human expansion of *CTAGE* family. In primate genomes,

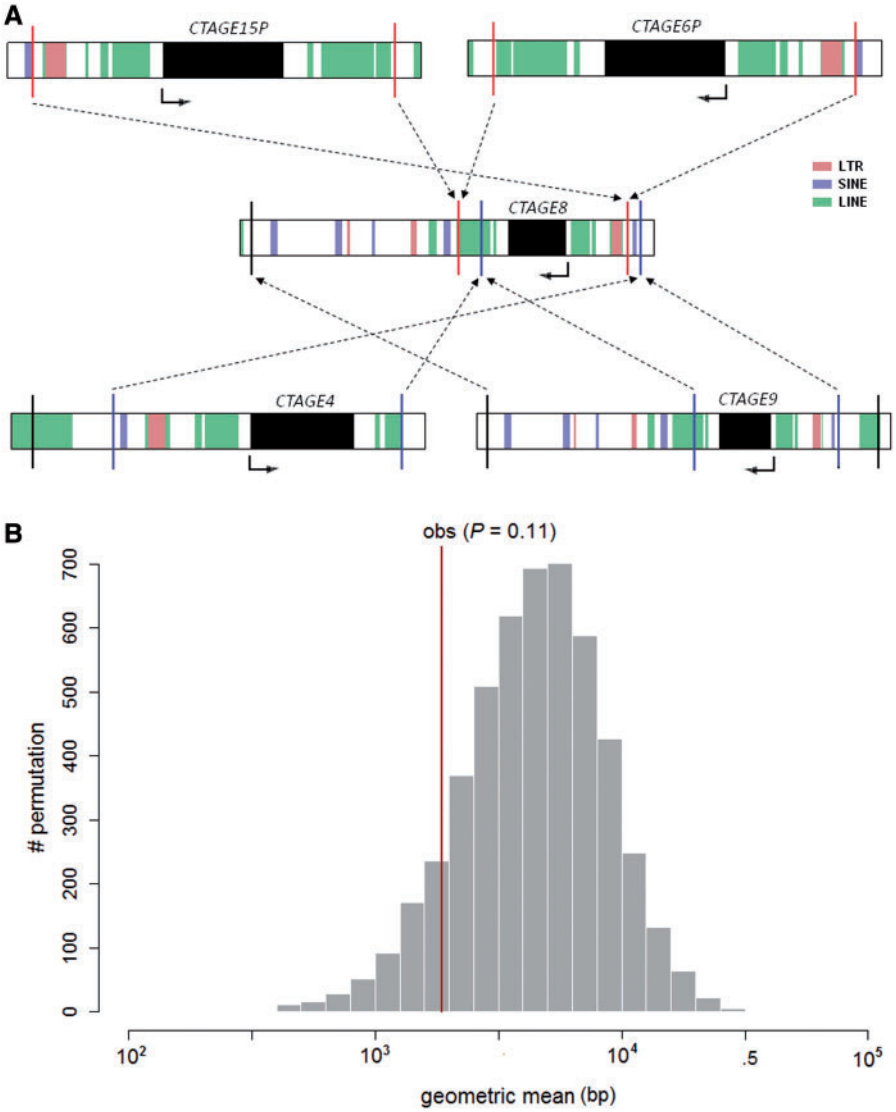

Fig. 5. Genomic characteristics of human specific *CTAGE* members. (*A*) Retro-elements are found in genic regions (black boxes) but are enriched in the boundary of orthologous regions. Colored lines denote break points of orthologous regions among *CTAGE* members. Dashed arrows denote the orthology of break points. Solid arrows indicate the transcriptional direction. Colored boxes indicate different transposable elements. (*B*) The geometric mean of the distance between five *CTAGE* members and the closest LTR element was calculated as denoted by the dark red vertical line. Permutation was performed using all protein-coding genes and pseudogenes for 5,000 times to generate the random distribution of the genomic mean.

retroposons including long terminal repeats (LTRs), short interspersed nucleotide elements (SINEs), and long interspersed nucleotide elements (LINEs) are abundant and make up approximately 40% of the human genome (Cordaux and Batzer 2009). Because of their high copy numbers, retroposon elements can also create duplications through recombination between nonallelic homologous elements (Yang et al. 2008; Cordaux and Batzer 2009). We thus studied the genomic composition of human-specific *CTAGE* duplicons identified above and found these duplicons were enriched for retroposons and almost all the break points adjacent to LINE, SINE, or LTR elements (fig. 5*A*). This finding suggests that the human-specific burst of *CTAGE* family may be derived by retroposon-mediated NAHR.

Because the *CTAGE* family is originally duplicated via retroposition, it is also interesting to understand how these human-specific copies acquired regulatory elements and get transcribed. There are several ways for new copies to acquire promoters, such as recruiting the promoter of a neighboring gene or utilizing promoter from LTR element (Romanish et al. 2007; Fablet et al. 2009). To study this, we first analyzed *CTAGE15P*, the ancestral copy that leads to the human-specific expansion. Although *CTAGE15P* is a pseudogene, it is transcribed exclusively in testis (table 3). The distance from the transcription start site of *CTAGE15P* to that of three closest 5′-upstream genes are 94 kb (*TAS2R41*), 128 kb (*TAS2R60*), and 184 kb (*ZYX*), which suggests that it is less likely to recruit promoters from these neighboring genes due to the long distance (table 3). Additionally, the tissue expression pattern of *CTAGE15P* is also different from flanking genes (table 3), further supporting that *CTAGE15P* may obtain distinct promoters. Interestingly, there is an LTR element 1,815 bp

Table 3. Tissue Expression of Human *CTAGE15P* and its Three 5′-Upstream Genes.

Gene	Distance to *CTAGE15P* (kb)	Brain	Cerebellum	Heart	Kidney	Liver	Testis
ZYX	184	—	Y	—	Y	Y	Y
TAS2R60	128	—	—	—	—	—	—
TAS2R41	94	Y	—	—	—	—	—
CTAGE15P	0	—	—	—	—	—	Y

NOTE.—"Y" denotes presence of the transcript and "—" denotes absence.

5′-upstream *CTAGE15P* (fig. 5*A*), which is a *THE1B* element belonging to ERVL-MaLR endogenous retrovirus. Similarly, all the other human-specific *CTAGE* copies are distant from promoters of their 5′-upstream neighboring genes but have the same LTR element within two 5′-upstream region (1,881 bp, 1,880 bp, 1,869 bp, and 1,902 bp for *CTAGE4*, *CTAGE8*, *CTAGE9*, and *CTAGE6P*, respectively). Recently, researchers have found that a *THE1B* member upstream the colony-stimulating factor receptor (*CSF1R*) was reactivated and led to the abnormal expression of *CSF1R* in Hodgkin's lymphoma cell lines and primary samples (Lamprecht et al. 2010), implying this LTR element near *CTAGE* members may serve as the promoter. We next sought to justify whether the observed vicinity to LTR elements in these *CTAGE* members is a random incidence. The median distance to the closet LTR element for each protein-coding gene or pseudogene is 4,904 bp, substantially larger than that for the *CTAGE* members. We also calculated the geometric mean of the distance in *CTAGE* members and compared it with the result from 5,000-time permutations, and a marginal significance was observed (fig. 5*B*, $P = 0.11$). Collectively, it is very likely that *CTAGE* copies become actively transcribed by recruiting the LTR promoter.

Discussion

In this study, we identified the *CTAGE* gene family members in primates including human, chimpanzee, gorilla, orangutan, macaque, and marmoset. By using comparative genomic approaches, we analyzed the genomic organization of *CTAGE* gene family and observed a rapid expansion of this family starting from the common ancestor of catarrhine primates (human, chimpanzee, gorilla, orangutan, and macaque). Among them, gorilla genome has the fewest *CTAGE* copies, which is probably due to its low sequence quality, for instance, multiple consecutive N's were found within G6 and G8. Evolutionary analyses revealed different evolutionary forces acting on *CTAGE* copies, as multiexon copies are under purifying selection, whereas positive selection is detected during the evolution of single-exon copies.

Both DNA- and RNA-based duplications are pivotal in genome evolution (Kaessmann et al. 2009). The gene structure and organization of the *CTAGE* family in various primates reveals a complicated dynamism, which includes both DNA-based duplications (tandem or SD) and RNA-based duplications. Based on the reconstructed phylogenetic tree, an intronless *CTAGE* copy was first generated by retroposition after the split of strepsirrhines and haplorrhines. Retroposed copies are normally nonfunctional due to the lack of regulatory elements (Long et al. 2003); however, we observed several single-exon *CTAGE* genes (e.g., *CTAGE1*, *CTAGE4*, *CTAGE8*, and *CTAGE9*) with intact ORFs. Hence, it is intriguing to find out whether the ancestral retrocopy gained regulatory elements and passed that to other duplicated copies. Because *CTAGE4*, *CTAGE8*, and *CTAGE9* are highly similar in their flanking regions, we only compared 1-kb upstream regions between *CTAGE1* and *CTAGE9* and no homology can be found, indicating that the ancestral single-exon copy is likely to be nonfunctional, but a few *CTAGE* copies obtained regulatory elements independently from their flanking regions. This finding also suggests that gaining promoters from flanking sequences may not be as rare as generally thought. It is also notable that a series of independent duplications probably occurred in the common ancestor of catarrhines, leading to the prosperity of the *CTAGE* family in apes and monkeys.

Lineage-specific genes are common targets of positive selection (Han et al. 2009) and may be important in adaptation (Wilson et al. 2005; Zhang et al. 2010; Long et al. 2013). Here, we found that the *CTAGE* family has undergone a human-specific expansion and led to four new copies. Among these five copies (H1–H5), three (H1–H3 or *CTAGE9*, *CTAGE8*, and *CTAGE4*) are functional protein-coding genes, and gene predictions using Genscan (Burge and Karlin 1997) showed that the corresponding chimpanzee copy also encodes a protein, suggesting the ancestral copy in the human–chimpanzee ancestral branch might be already functional. We also examined reads from the recently sequenced archaic Denisovan individual (Meyer et al. 2012) using UCSC genome browser and validated both the 5′ and 3′ insertion sites for *CTAGE9*, the 3′ insertion site for *CTAGE4*, and the 5′ deletion site for *CTAGE8*, indicating those genomic modifications occurred before the split of modern humans and Denisovans at least 740,000 and 820,000 years ago (Meyer et al. 2012).

Cancer is a major health threat in the modern human society, leading to about 7.6 million deaths in 2008 (Jemal et al. 2011), and CT antigens may contribute to neoplastic phenotypes and participate in oncogenesis through the germline gene-expression program (Simpson et al. 2005), it is thus intriguing to see *CTAGE* family is evolving under positive selection and experienced a human-specific expansion. Intriguingly, a number of studies have shown that nonhuman primates have low cancer incidence rate (Beniashvili 1989; Waters et al. 1998; Varki 2000; Puente et al. 2006), which is thought to be partially related to genetic differences. However, a sequence comparison of 333 human cancer genes with chimpanzee orthologs showed a high degree of conservation and limited genetic difference could be

identified to potentially account for different cancer susceptibility in these two species (Puente et al. 2006). Hence, the human-specific copy number gains of *CTAGE* genes, or in general gene copy number changes, may present a new avenue to understand the high cancer incidence rate in humans. Although the germline-biased expression could be due to widespread transcriptional activity in testis (Soumillon et al. 2013), genes expressed predominantly in gametogenesis or involved in reproduction are often under strong positive selection (Swanson and Vacquier 1995; Wyckoff et al. 2000; Bustamante et al. 2005; Stevenson et al. 2007; Zhang et al. 2007), and the emergence of novel genes could potentially create reproductive barriers (Singh and Kulathinal 2000; Wyckoff et al. 2000; Swanson and Vacquier 2002; Kouprina et al. 2004), therefore the human expansions may also have potential implications in speciation. We encourage further functional characterization of these human-specific *CTAGE* genes, which will be particularly helpful to test the above hypothesis.

Materials and Methods

Phylogenetic and Evolutionary Analysis

CTAGE family members were identified by comparing the coding sequence of *CTAGE9* against human (Lander et al. 2001), chimpanzee (Mikkelsen et al. 2005), gorilla (Scally et al. 2012), orangutan (Locke et al. 2011), macaque (Gibbs et al. 2007), marmoset, mouse (Waterston et al. 2002), rat (Gibbs et al. 2004), cow (Elsik et al. 2009), and dog (Lindblad-Toh et al. 2005) using UCSC genome browser (Kent et al. 2002). Comparison was done by BLAT (Kent 2002) and hits with $\geq$ 70% sequence similarity and 50% query coverage were retrieved. Subsequent manual inspection found that several hits with high sequence similarity ($\geq$90%) were overlapped with low-quality genomic regions (a long string of "N"s). Thus, they were excluded from further analyses as it is impossible to generate a high-quality sequence alignment. Multiexon *CTAGE* copies were also retrieved from gibbon and bushbaby using UCSC genome browser. Multiple sequence alignments were constructed by MUSCLE (Edgar 2004), which optimizes both alignment accuracy and speed. MEGA5 (Tamura et al. 2011) was used to infer phylogenetic relationships based on the maximum-likelihood method and Kimura two-parameter model (Kimura 1980). The reconstructed phylogeny was evaluated by the bootstrap test (Felsenstein 1985) with 1,000 replicates for each node.

Human-coding *CTAGE* genes and their homologs were aligned based on codons using MUSCLE. Phylogenetic relationships were reconstructed separately for single-exon copies and multiexon copies and were then concatenated. The positions of nonprimate outgroup were adjusted according to the species tree. To estimate d_N/d_S (ω) of each branch, free-ratio model in codeml embedded in PAML (Yang 1997) was applied. To test functional constraint in multiexon genes, we compared two two-ratio models: In the null model, single-exon genes and multiexon genes have two different ω values; in the alternative model, ω was fixed as 1 for multiexon genes. The statistical significance was measured by the likelihood ratio test . A branch-site model was further used to test positive selection on specific branches and sites (Yang and Nielsen 2002; Zhang et al. 2005). Positively selected sites (codons) were inferred by an empirical Bayes analysis (Yang et al. 2005).

Expression Analysis

RNAseq data were used to verify the expression of coding *CTAGE* copies. Data set with accession number GSE30352 was retrieved from NCBI-SRA, which includes brain, cerebellum, heart, liver, kidney, and testis in both humans and chimpanzees (Brawand et al. 2011). Short reads were mapped to the reference genome by TopHat (Trapnell et al. 2009), using the Ensembl (Flicek et al. 2013) protein-coding annotation file (release 69). Next, the result was fed into Cufflinks (Trapnell et al. 2010) to calculate the number of fragments per kilobase of transcript per million fragments mapped (Mortazavi et al. 2008; Trapnell et al. 2010).

Human Insert Site Confirmation by Deep-Sequencing Data

To confirm that human-specific expansions are not the result of assembly error, we downloaded whole-genome sequencing data for two human individuals (NA12877 and NA12878, NCBI-SRA accession ERP001228 and ERP000603) and aligned the short reads to human genome using BWA (Li and Durbin 2009, 2010) with default parameters. Whole-genome sequencing from seven central chimpanzee (*Pan troglodytes troglodytes*) individuals (Prufer et al. 2012) were also used to prove the absence of human-specific expansions in chimpanzee.

Human Retroposon Information

We used the Table browser in UCSC genome browser to download human retroposon positions defined by RepeatMasker (http://www.repeatmasker.org, last accessed June 17, 2014). The group is "Repeats" and the track is "RepeatMasker." Custom python scripts were used to calculate the closest distance between LTR elements and Ensembl genes.

Supplementary Material

Supplementary tables S1 and S2 and file S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors appreciate helpful comments from anonymous reviewers. This work was supported by the Department of Human Evolutionary Biology, Harvard University, by grants from the National 973 project of China (2011CBA00401), and the National Natural Science Foundation of China (31130051, 31301028, and 31321002).

References

Andres O, Kellermann T, Lopez-Giraldez F, Rozas J, Domingo-Roura X, Bosch M. 2008. RPS4Y gene family evolution in primates. *BMC Evol Biol.* 8:142.

Arroyo JI, Hoffmann FG, Opazo JC. 2012. Gene turnover and differential retention in the relaxin/insulin-like gene family in primates. *Mol Phylogenet Evol.* 63:768–776.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet.* 7:552–564.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* 297:1003–1007.

Beniashvili DS. 1989. An overview of the world literature on spontaneous tumors in nonhuman primates. *J Med Primatol.* 18:423–437.

Betran E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophila*. *Genome Res.* 12:1854–1859.

Bosch N, Caceres M, Cardone MF, Carreras A, Ballana E, Rocchi M, Armengol L, Estivill X. 2007. Characterization and evolution of the novel gene family FAM90A in primates originated by multiple duplication and rearrangement events. *Hum Mol Genet.* 16:2572–2582.

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478:343–348.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* 268:78–94.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437:1153–1157.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A.* 104:19428–19433.

Conrad DF, Hurles ME. 2007. The population genetics of structural variation. *Nat Genet.* 39:S30–S36.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet.* 10:691–703.

Das S, Nikolaidis N, Goto H, McCallister C, Li J, Hirano M, Cooper MD. 2010. Comparative genomics and evolution of the alpha-defensin multigene family in primates. *Mol Biol Evol.* 27:2333–2343.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* 149:912–922.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Eichmuller S, Usener D, Dummer R, Stein A, Thiel D, Schadendorf D. 2001. Serological detection of cutaneous T-cell lymphoma-associated antigens. *Proc Natl Acad Sci U S A.* 98:629–634.

Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, Adelson DL, Eichler EE, Elnitski L, Guigo R, et al. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324:522–528.

Emerson JJ, Kaessmann H, Betran E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. *Science* 303:537–540.

Fablet M, Bueno M, Potrzebowski L, Kaessmann H. 2009. Evolutionary origin and functions of retrogene introns. *Mol Biol Evol.* 26:2147–2156.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. *Nucleic Acids Res.* 41:D48–D55.

Fratta E, Coral S, Covre A, Parisi G, Colizzi F, Danielli R, Nicolay HJ, Sigalotti L, Maio M. 2011. The biology of cancer testis antigens: putative function, regulation and therapeutic potential. *Mol Oncol.* 5:164–182.

Giannuzzi G, Siswara P, Malig M, Marques-Bonet T, Mullikin JC, Ventura M, Eichler EE. 2013. Evolutionary dynamism of the primate LRRC37 gene family. *Genome Res.* 23:46–59.

Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316:222–234.

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493–521.

Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW. 2009. Adaptive evolution of young gene duplicates in mammals. *Genome Res.* 19:859–867.

Hofmann O, Caballero OL, Stevenson BJ, Chen YT, Cohen T, Chua R, Maher CA, Panji S, Schaefer U, Kruger A, et al. 2008. Genome-wide analysis of cancer/testis gene expression. *Proc Natl Acad Sci U S A.* 105:20422–20427.

Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. 2011. Global cancer statistics. *CA Cancer J Clin.* 61:69–90.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20:1313–1326.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet.* 10:19–31.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* 12:656–664.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res.* 12:996–1006.

Kim PM, Lam HY, Urban AE, Korbel JO, Affourtit J, Grubert F, Chen X, Weissman S, Snyder M, Gerstein MB. 2008. Analysis of copy number variants and segmental duplications in the human genome: evidence for a change in the process of formation in recent evolutionary history. *Genome Res.* 18:1865–1874.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol.* 16:111–120.

Kouprina N, Mullokandov M, Rogozin IB, Collins NK, Solomon G, Otstot J, Risinger JI, Koonin EV, Barrett JC, Larionov V. 2004. The SPANX gene family of cancer/testis-specific antigens: rapid evolution and amplification in African great apes and hominids. *Proc Natl Acad Sci U S A.* 101:3077–3082.

Lamprecht B, Walter K, Kreher S, Kumar R, Hummel M, Lenze D, Kochert K, Bouhlel MA, Richter J, Soler E, et al. 2010. Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat Med.* 16:571–579.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A.* 103:9935–9939.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589–595.

Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* 437:94–100.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ 3rd, Zody MC, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438:803–819.

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang SP, Wang Z, Chinwalla AT, Minx P, et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* 469:529–533.

Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet.* 4:865–875.
Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we know. *Annu Rev Genet.* 47:307–333.
Lupski JR. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* 14:417–422.
Maeyama Y, Otsu M, Kubo S, Yamano T, Iimura Y, Onodera M, Kondo S, Sakiyama Y, Ariga T. 2011. Intracellular estrogen receptor-binding fragment-associated antigen 9 exerts in vivo tumor-promoting effects via its coiled-coil region. *Int J Oncol.* 39:41–49.
Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338: 222–226.
Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang F, Enard W, Hellmann I, Lindblad-Toh K. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437: 69–87.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 5:621–628.
Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci U S A.* 106:6700–6705.
Prufer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486:527–531.
Puente XS, Velasco G, Gutierrez-Fernandez A, Bertranpetit J, King MC, Lopez-Otin C. 2006. Comparative analysis of cancer genes in the human and chimpanzee genomes. *BMC Genomics* 7:15.
Romanish MT, Lock WM, van de Lagemaat LN, Dunn CA, Mager DL. 2007. Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. *PLoS Genet.* 3:e10.
Saito K, Yamashiro K, Ichikawa Y, Erlmann P, Kontani K, Malhotra V, Katada T. 2011. cTAGE5 mediates collagen secretion through interaction with TANGO1 at endoplasmic reticulum exit sites. *Mol Biol Cell.* 22:2301–2308.
Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483:169–175.
Simpson AJ, Caballero OL, Jungbluth A, Chen YT, Old LJ. 2005. Cancer/testis antigens, gametogenesis and cancer. *Nat Rev Cancer.* 5: 615–625.
Singh RS, Kulathinal RJ. 2000. Sex gene pool evolution and speciation: a new paradigm. *Genes Genet Syst.* 75:119–130.
Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthes P, Kokkinaki M, Nef S, Gnirke A, et al. 2013. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* 3:2179–2190.
Stevenson BJ, Iseli C, Panji S, Zahn-Zabal M, Hide W, Old LJ, Simpson AJ, Jongeneel CV. 2007. Rapid evolution of cancer/testis genes on the X chromosome. *BMC Genomics* 8:129.
Swanson WJ, Vacquier VD. 1995. Extraordinary divergence and positive Darwinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa. *Proc Natl Acad Sci U S A.* 92: 4957–4961.
Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. *Nat Rev Genet.* 3:137–144.
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.
Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105–1111.
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 28:511–515.
Usener D, Schadendorf D, Koch J, Dubel S, Eichmuller S. 2003. cTAGE: a cutaneous T cell lymphoma associated antigen family with tumor-specific splicing. *J Invest Dermatol.* 121:198–206.
Vandepoele K, Van Roy N, Staes K, Speleman F, van Roy F. 2005. A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. *Mol Biol Evol.* 22:2265–2274.
Varki A. 2000. A chimpanzee genome project is a biomedical imperative. *Genome Res.* 10:1065–1070.
Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A.* 103:3220–3225.
Wainszelbaum MJ, Charron AJ, Kong C, Kirkpatrick DS, Srikanth P, Barbieri MA, Gygi SP, Stahl PD. 2008. The hominoid-specific oncogene TBC1D3 activates Ras and modulates epidermal growth factor receptor signaling and trafficking. *J Biol Chem.* 283:13233–13242.
Waters DJ, Sakr WA, Hayden DW, Lang CM, McKinney L, Murphy GP, Radinsky R, Ramoner R, Richardson RC, Tindall DJ. 1998. Workgroup 4: spontaneous prostate carcinoma in dogs and nonhuman primates. *Prostate* 36:64–67.
Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.
Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D. 2005. Orphans as taxonomically restricted and ecologically important genes. *Microbiology* 151:2499–2501.
Wyckoff GJ, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304–309.
Yang S, Arguello JR, Li X, Ding Y, Zhou Q, Chen Y, Zhang Y, Zhao R, Brunet F, Peng L, et al. 2008. Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*. *PLoS Genet.* 4:e3.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13:555–556.
Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol.* 19:908–917.
Yang Z, Nielsen R, Goldman N. 2009. In defense of statistical methods for detecting positive selection. *Proc Natl Acad Sci U S A.* 106:E95; author reply E96.
Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 22: 1107–1118.
Zhang C, Wang J, Xie W, Zhou G, Long M, Zhang Q. 2011. Dynamic programming procedure for searching optimal models to estimate substitution rates based on the maximum-likelihood method. *Proc Natl Acad Sci U S A.* 108:7860–7865.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.* 22:2472–2479.
Zhang Q. 2013. The role of mRNA-based duplication in the evolution of the primate genome. *FEBS Lett.* 587:3500–3507.
Zhang Q, Backstrom N. 2014. Assembly errors cause false tandem duplicate regions in the chicken (*Gallus gallus*) genome sequence. *Chromosoma* 123:165–168.
Zhang Q, Zhang F, Chen XH, Wang YQ, Wang WQ, Lin AA, Cavalli-Sforza LL, Jin L, Huo R, Sha JH, et al. 2007. Rapid evolution, genetic variations, and functional association of the human spermatogenesis-related gene NYD-SP12. *J Mol Evol.* 65:154–161.
Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol.* 8:pii:e1000494.