Available online at www.sciencedirect.com
ScienceDirect
Genomics 88 (2006) 745–751
GENOMICS
www.elsevier.com/locate/ygeno

# Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup

Xiao-Jing Yu [a,b,c], Hong-Kun Zheng [d,e], Jun Wang [d,e,f], Wen Wang [a], Bing Su [a,b,*]

[a] Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China
[b] Kunming Primate Research Center, Chinese Academy of Sciences, Kunming, Yunnan 650223, China
[c] Graduate School, Chinese Academy of Sciences, Beijing, China
[d] Beijing Genomics Institute, Chinese Academy of Sciences, Beijing, China
[e] The Institute of Human Genetics, University of Aarhus, DK-8000 Aarhus C, Denmark
[f] Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230, Odense M, Denmark

Received 12 April 2006; accepted 23 May 2006
Available online 18 July 2006

**Abstract**

Comparative genetic analysis between human and chimpanzee may detect genetic divergences responsible for human-specific characteristics. Previous studies have identified a series of genes that potentially underwent Darwinian positive selection during human evolution. However, without a closely related species as outgroup, it is difficult to identify human-lineage-specific changes, which is critical in delineating the biological uniqueness of humans. In this study, we conducted phylogeny-based analyses of 2633 human brain-expressed genes using rhesus macaque as the outgroup. We identified 47 candidate genes showing strong evidence of positive selection in the human lineage. Genes with maximal expression in the brain showed a higher evolutionary rate in human than in chimpanzee. We observed that many immune-defense-related genes were under strong positive selection, and this trend was more prominent in chimpanzee than in human. We also demonstrated that rhesus macaque performed much better than mouse as an outgroup in identifying lineage-specific selection in humans.


*Keywords:* Positive selection; Adaptive evolution; Brain-expressed gene; Hominoids

Although nearly 99% of genomic sequences are identical, the biological divergence between human and chimpanzee is distinctive, especially in view of cognitive abilities [1,2]. Comparative genetic analysis of human and chimpanzee may detect genetic divergences responsible for the human-specific characteristics. Indeed, in the past several years, scientists have devoted great effort toward understanding the evolutionary changes that have occurred in the human lineage after the divergence of human and chimpanzee about 5–6 million years ago [3–7]. Clark et al. conducted a genome-wide sequence comparison of 7645 orthologous genes between human and chimpanzee (with mouse as outgroup) to identify genes that underwent Darwinian positive selection in humans [8]. They concluded that the putative positively selected genes in human were responsible mainly for several biological functions, including olfaction, sensory perception, and transportation. By analyzing 13,731 human–chimp orthologs, Nielsen et al. also showed that genes involved in sensory perception and immune system tend to evolve rapidly due to positive selection and genes with maximal expression within the brain show little or no evidence of positive selection [9]. Recently, Bustamante et al. reported that genes involved in apoptosis, gametogenesis, and defense/immunity tend to evolve under positive selection in human. In their study, human population data were used to confirm the suggested positive selection of the rapid-evolving genes [10]. Another study by Dorus et al. investigated the evolutionary patterns of 214 nervous system genes, and they observed a higher nonsynonymous versus synonymous substitution rate ($K_a/K_s$) in primates (human and macaque) than in

* Corresponding author. Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China. Fax: +86 871 5193137.
*E-mail address:* sub@mail.kiz.ac.cn (B. Su).

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.05.008

rodents (mouse and rat). They also found that the $K_a/K_s$ ratio of the human lineage was considerably higher than that of the chimpanzee lineage [11].

The suggested positively selected candidate genes in human from different studies are, however, different from one another due to different sets of genes being analyzed and different methods and outgroup species used in inferring lineage-specific changes. In the study of Clark et al. [8], mouse was used as the outgroup species since it was the closest species to primates with a sequenced genome at that time. The divergence time separating human/chimpanzee and mouse is about 91 million years [12]. With such a deep divergence, the estimation of lineage-specific substitution patterns (nonsynonymous vs synonymous) might be subject to potential bias, e.g., the uncertainty of inferring human/chimpanzee ancestral sequences at rapidly evolving sites. Hence, a phylogenetically closely related species is preferred in choosing an outgroup, which allows more accurate and sensitive estimation of lineage-specific substitution patterns, and is critical to identify human-specific changes. Fortunately, the rapid growth of the genomic sequence of rhesus macaque (http://hgdownload.cse.ucsc.edu/goldenPath/rheMac1/) provided an opportunity to detect genes showing adaptive changes in the human lineage. In this study, using rhesus macaque as the outgroup species, we conducted a genome-wide comparison of the brain-expressed genes between human and chimpanzee to identify genes showing distinctive evolutionary changes along the human lineage. We also evaluated the effect of using different outgroups (rhesus macaque vs mouse) in detecting lineage-specific positively selected genes.

## Results

By combining reciprocal best matches and coincident location evidence between 10,184 human brain-expressed gene sequences and the genome sequences of three other species (chimpanzee, rhesus macaque, and mouse), we obtained the orthologous coding sequences for 2633 genes in human, chimpanzee, rhesus macaque, and mouse with rigorous criteria (see Methods for detailed description). Among these genes, 809 were included in Clark et al. [8], and 1821 and 1806 genes were previously analyzed by Nilesen et al. and Bustamante et al. [9,10].

The selective pressure at the molecular level can be measured by the ratio of nonsynonymous ($K_a$) and synonymous ($K_s$) substitution rates [13]. Using rhesus macaque as outgroup, we inferred the ancestral sequences of human and chimpanzee and then estimated the $K_a/K_s$ ratios of the 2633 genes in the human and chimpanzee lineages, respectively [14,15]. The average $K_a/K_s$ ratio was 0.22 for the human lineage and 0.25 for the chimpanzee lineage, indicating that overall the selective pressures on the brain-expressed genes are similar between human and chimpanzee.

There were 319 human genes and 321 chimpanzee genes showing $K_a/K_s>1$, among which 75 were shared by the two species. To test whether $K_a$ is significantly higher than $K_s$ in these genes, we then applied the maximum likelihood ratio test based on the branch-site model at the codon level [16]. In this test, we evaluated whether positive selection was performed in a specific lineage by comparing twice the difference in log-likelihood values of the null hypothesis, $K_a/K_s=1$, and alternative hypothesis, $K_a/K_s>1$. Our results showed that the null hypothesis was rejected for 47 of the 319 human genes and 54 of the 321 chimpanzee genes ($p<0.05$), implying that the number of brain-expressed genes under positive selection was similar between human and chimpanzee. However, only 4 genes were shared between them, indicating that selection acted on different sets of genes in human and chimpanzee, which might lead to different biological consequences. The results of all genes can be found in the Supplementary materials (Dataset S1).

### *Evolutionary pattern of brain-maximal-expressed genes*

The brain-maximal-expressed (BME) genes are likely to be involved in brain function and may show different evolutionary patterns between human and chimpanzee. We obtained the expression profiles of the 2633 genes from Gene Expression Atlas [17], and genes with maximal expression within the brain were selected. After eliminating genes with no substitutions ($K_a=0$, $K_s=0$), a total of 129 human genes and 134 chimpanzee genes were selected with maximal expression in the brain (Dataset S2; see Methods for detailed description).

Using the Mann–Whitney *U* test (MWU test), we compared the $K_a/K_s$ ratios of the BME genes with the non-BME genes in the human and chimpanzee lineages, respectively. Our results indicated that the $K_a/K_s$ ratios of the BME genes were significantly lower than those of the non-BME genes in the chimp lineage, a signature of strong functional constraint (negative selection) on the BME genes in chimpanzee ($p = 0.0057$; Table 1). However, in the human lineage, no such difference was observed. This difference implied more rapid evolution of the BME genes in the human lineage than in the chimpanzee lineage (Table 1), which could be caused by either Darwinian positive selection or relaxation of functional constraint. We further conducted the Wilcoxon signed-rank test to compare the $K_a/K_s$ ratios of the BME genes between human and chimpanzee. The human lineage showed a marginally larger value than the chimpanzee lineage (116 genes, $p=0.08$), further supporting that BME genes might evolve rapidly in the human lineage. For the BME genes tested, a total of five (*KLHDC3*, *EPS15*, *RPS6KL1*, *MARK1*, *C20orf46*) and two (*NAP1L2*, *ULK2*) genes were detected with significant $K_a/K_s>1$ (with $p<0.05$) in the human and chimpanzee lineages, respectively.

Table 1
The results of statistical tests for the BME genes

| Lineage | Number of genes | *p* value |
|---|---|---|
| Human | 129 | 0.1672 |
| Chimpanzee | 134 | 0.0057 |

The *p* values indicate the statistical significance of the BME–non-BME $K_a/K_s$ ratio difference (MWU right-tailed test).

### *Functional categories of overrepresentation of rapidly evolving genes*

We obtained the inferred functional categories of the 2633 genes based on the Panther classification system [18,19]. To identify certain functional categories with an excess of rapidly evolving genes compared to other functional categories, we used a MWU test based on the $K_a/K_s$ ratios (Dataset S3; Tables 2 and 3). We eliminated 360 human genes and 448 chimp genes with no substitutions from the analysis.

In the human lineage, the categories "biological process unclassified" and "molecular function unclassified" were significantly enriched in rapidly evolving genes ($p=0.0001$). Interestingly, many of the genes with unknown biological process were involved in several known molecular functional categories such as KRAB box transcription factors and receptors. Similarly, several genes grouped under "molecular function unclassified" were classified in known biological process categories, including intracellular protein traffic and spermatogenesis and motility. Nielsen et al. [9] had also identified that putative positively selected genes were significantly overrepresented in "biological process unclassified." Therefore, further studies delineating the functional roles of these genes will shed light on the evolutionary significance of these genes.

In both the human and the chimpanzee lineages, genes related to immune defense evolved rapidly. There have been numerous studies suggesting that immune-defense-related genes tend to be under positive selection [9,20–23], which was explained by the presence of genetic conflict between host and pathogen. Nielsen et al. [9] also gave another explanation, which had previously been used to explain the presence of positive selection in the human major histocompatibility complex [24], i.e., positive selection on immune- and defense-related genes was caused by overdominant selection to diversify the spectrum of immune responses.

In addition to the functional categories mentioned above, in the human lineage, other categories that had been identified as under positive selection previously [8,10] were also observed in our study, including cell adhesion, extracellular matrix, and zinc finger transcription factor.

Table 2
The biological process gene categories with an excess of rapidly evolving genes

| Lineage | Biological process | Number of genes | *p* value |
|---|---|---|---|
| Human | Biological process unclassified | 592 | 0.0001 |
| | Cytokine- and chemokine-mediated signaling pathway | 25 | 0.0094 |
| | Cell adhesion | 66 | 0.0274 |
| Chimpanzee | T-cell-mediated immunity | 22 | 0.0010 |
| | Immunity and defense | 135 | 0.0013 |
| | DNA metabolism | 221 | 0.0031 |
| | Exocytosis | 19 | 0.0137 |
| | Cell adhesion | 67 | 0.0160 |
| | Cell cycle | 87 | 0.0476 |

The *p* values indicate the statistical significance of one category vs other category disparities (MWU right-tailed test).

Table 3
The molecular function categories with an excess of rapidly evolving genes

| Lineage | Molecular function | Number of genes | *p* value |
|---|---|---|---|
| Human | Molecular function unclassified | 570 | 0.0001 |
| | Defense/immunity protein | 33 | 0.0138 |
| | Extracellular matrix | 51 | 0.0204 |
| | Zinc finger transcription factor | 83 | 0.0385 |
| | Ligase | 46 | 0.0441 |
| Chimpanzee | Defense/immunity protein | 31 | 0.0001 |
| | Cell adhesion molecule | 42 | 0.0255 |
| | Membrane-bound signaling molecule | 20 | 0.0428 |
| | Signaling molecule | 91 | 0.0499 |

The *p* values indicate the statistical significance of one category vs other category disparities (MWU right-tailed test).

We also identified categories including cytokine- and chemokine-mediated signaling pathways and ligases with an excess of rapidly evolving genes. As many of the cytokines and chemokines are involved in the immune system [25–27], again the rapid evolution of these genes confirmed the proposed positive selection on immune-function-related genes.

### *Analysis of the 47 putative positively selected genes in the human lineage*

Based on the likelihood ratio test, we detected 47 genes with significant $K_a/K_s>1$ ratios in the human lineage (Table 4). Thirty-nine of the 47 genes were also analyzed by Nielsen et al. [9], but none of them was listed among the 50 candidates with $K_a/K_s>1$ in their study. For the 50 candidate genes listed by Nielsen et al., only 3 were included in our dataset, and we also identified them with $K_a/K_s>1$, but the likelihood ratio test did not show significance. Therefore, the use of macaque as the outgroup species is more sensitive to detect lineage-specific positively selected genes, which is critical to delineate the mechanism of human evolution.

For the 47 candidate genes, 6 are involved in gametogenesis and developmental process (*CCNA1*, *KLHDC3*, *PMS2L2*, *OSR2*, *SEPP1*, *HSF4*), 6 in nucleic acid metabolism (*PMS2L2*, *ARID4A*, *HSF4*, *OSR2*, *POLR2G*, *ZNRD1*), 8 in signal transduction (*STAM2*, *PAQR6*, *RPS6KL1*, *SIGLEC5*, *DAG1*, *NPY*, *TSHR*, *NKIR*), and 4 (*LGALS3*, *NKIR*, *SIGLEC5*, *MTCP1*) in immunity. These functional categories had also been identified as undergoing positive selection in humans in previous studies [8–10]. Among these positively selected genes, *TSHR* (thyroid-stimulating hormone receptor) is a member of the glycoprotein hormone family[28]. *TSHR* has been suggested to have coevolved with its ligands, a novel glycoprotein hormone subunit family of genes, α2 and β5 [29]. Knudsen et al. reported that the divergence rate of residues or domains of *TSHR* is rapid in mammals and they also identified several functional sites specific to humans [30]. Only 2 nervous-system-related genes are in the candidate list (*NPY, EPS15*), and *NPY* is also involved in signal transduction (Table 4). This pattern suggests that the number of neurofunction-related genes contributing to the human cognitive ability during evolution might be very small.

Table 4
The 47 candidate genes that underwent positive selection in the human lineage

| Gene symbol | Function | *p* value |
|---|---|---|
| *STAM2* | Signal transducing adaptor molecule | 0.0000 |
| *OSR2* | Odd-skipped-related 2 | 0.0000 |
| *TNIP1* | Tumor necrosis factor-interacting protein | 0.0000 |
| *TLOC1* | Translocation protein | 0.0000 |
| *SEPP1* | Extracellular matrix glycoprotein; sperm development | 0.0000 |
| *EPS15* | Epidermal growth factor receptor pathway substrate | 0.0000 |
| *C20orf46* | Brain-maximal-expressed | 0.0000 |
| *FLJ12700* | Unknown | 0.0000 |
| *LGALS3* | Lectin, galactoside-binding soluble | 0.0000 |
| *RPS6KL1* | Nonreceptor serine/threonine protein kinase | 0.0000 |
| *CGI-115* | Unknown | 0.0002 |
| *CLMN* | Calponin-like, transmembrane | 0.0002 |
| *C6orf47* | Unknown | 0.0003 |
| *POLR2G* | Transcription factor | 0.0004 |
| *LOC123872* | Unknown | 0.0008 |
| *FLJ31795* | Unknown | 0.0009 |
| *ZNRD1* | DNA-directed RNA polymerase | 0.0009 |
| *ACAD8* | Dehydrogenase | 0.0013 |
| *UBL3* | Ubiquitin-like 3 | 0.0019 |
| *ARID4A* | Transcription cofactor | 0.0019 |
| *BNIP1* | BCL2/adenovirus E1B 19-kDa interacting protein 1 | 0.0019 |
| *DKFZP434P0316* | Unknown | 0.0025 |
| *NKIR* | Defense/immunity protein | 0.0025 |
| *CCNA1* | Kinase activator; spermatogenesis and motility | 0.0027 |
| *NPY* | Neuropeptide Y | 0.0035 |
| *UMPS* | Uridine monophosphate synthetase | 0.0041 |
| *CKLFSF3* | Unknown | 0.0052 |
| *MARK1** | MAP/microtubule affinity-regulating kinase 1 | 0.0052 |
| *HSF4* | Heat shock transcription factor 4 | 0.0053 |
| *MTCP1* | Mature T cell proliferation | 0.0058 |
| *TSHR** | Thyroid-stimulating hormone receptor | 0.0058 |
| *AQP4** | Transporter | 0.0059 |
| *KLHDC3* | Chromatin-binding protein; spermatogenesis and motility | 0.0075 |
| *FHOD1* | Nonmotor actin-binding protein | 0.0082 |
| *SPOCK* | Cysteine protease inhibitor | 0.0086 |
| *FLJ11046* | Unknown | 0.0102 |
| *PAQR6* | Receptor | 0.0103 |
| *EFO1* | Acetyltransferase | 0.0104 |
| *PMS2L2* | DNA binding | 0.0135 |
| *SIGLEC5* | Sialic acid binding Ig-like lectin 5 | 0.0194 |
| *FLJ30990* | Unknown | 0.0282 |
| *C10orf7* | Unknown | 0.0289 |
| *FLJ23263* | Unknown | 0.0294 |
| *DAG1* | Receptor | 0.0303 |
| *FLJ22794* | Unknown | 0.0440 |
| *D2LIC* | Dynein, cytoplasmic 2, light intermediate chain 1 | 0.0451 |
| *C1orf27* | Dehydrogenase | 0.0474 |

The asterisks indicate positively selected genes reported by Clark et al. [8]. The *p* values indicate statistical significance of $K_a/K_s > 1$ (likelihood ratio test).

### *Comparison of effects of using a different outgroup*

Of the 47 candidate genes identified in this study, 17 of them were included by Clark et al., and only 3 genes were shown to be positively selected in the human lineage (*MARK1*, *TSHR*, and *AQP4*) [8]. The discrepancy could be caused by using a different dataset (e.g., the length of the genes might be different) and/or using different outgroups. Hence, we conducted a comparative analysis using different outgroups, e.g., mouse or rhesus macaque, to see if the use of different outgroups has any effect on detecting positively selected genes in humans.

The sequence data of the 2633 genes were edited so that all the genes have the same length in both the human–chimp–macaque (HCR) group and the human–chimp–mouse (HCM) group (Dataset S4). When the mouse orthologous sequences were considered, there were some sequence length differences in the new HCR dataset compared with the former HCR dataset. Totals of 333 and 276 genes having $K_a/K_s > 1$ were detected in the HCR and HCM, respectively; 28 and 26 genes were statistically significant ($p < 0.05$, likelihood ratio test). However,

Table 5
The positively selected genes identified in HCR and HCM grouping analyses

| Group | Gene symbol | Function | *p* value |
|---|---|---|---|
| HCR only | *WDR20* | Unknown | 0.0000 |
| | *C20orf46* | Unknown | 0.0000 |
| | *CLMN* | Calponin-like, transmembrane | 0.0001 |
| | *TLOC1* | Translocation protein | 0.0004 |
| | *F2R* | Coagulation factor II (thrombin) receptor | 0.0014 |
| | *ARID4A* | Transcription cofactor | 0.0018 |
| | *UMPS* | Uridine monophosphate synthetase | 0.0023 |
| | *MARK1* | MAP/microtubule affinity-regulating kinase 1 | 0.0053 |
| | *TSHR* | Thyroid-stimulating hormone receptor | 0.0057 |
| | *PMS2L2* | DNA binding | 0.0195 |
| | *DAG1* | Receptor | 0.0299 |
| | *CASP4* | Apoptosis-related cysteine protease | 0.0309 |
| HCM only | *PNMT* | Methyltransferase | 0.0005 |
| | *OTUB2* | OTU domain, ubiquitin aldehyde binding 2 | 0.0010 |
| | *NOL5A* | Ribonucleoprotein; rRNA metabolism | 0.0020 |
| | *CHCHD5* | Unknown | 0.0026 |
| | *GGPS1* | Synthase; acyltransferase | 0.0030 |
| | *WFDC1* | Serine protease inhibitor; tumor suppressor | 0.0039 |
| | *LRRTM4* | Receptor | 0.0093 |
| | *SLC4A1AP* | Other RNA-binding protein | 0.0356 |
| | *C5orf16* | Unknown | 0.0377 |
| | *FLJ20635* | Unknown | 0.0387 |
| HCR and HCM | *OSR2* | Odd-skipped-related 2 | 0.0000 |
| | *RPS6KL1* | Nonreceptor serine/threonine protein kinase | 0.0000 |
| | *SIAT8C* | Glycosyltransferase | 0.0000 |
| | *FLJ31795* | Unknown | 0.0001 |
| | *POLR2G* | Transcription factor | 0.0004 |
| | *BNIP1* | BCL2/adenovirus E1B 19-kDa-interacting protein 1 | 0.0005 |
| | *KCTD10* | Unknown | 0.0007 |
| | *ZNF148* | Zinc finger transcription factor | 0.0015 |
| | *TIAM1* | Guanyl-nucleotide exchange factor | 0.0017 |
| | *NPY* | Neuropeptide Y | 0.0020 |
| | *UBL3* | Ubiquitin-like 3 | 0.0027 |
| | *C10orf7* | Unknown | 0.0029 |
| | *CKLFSF3* | Unknown | 0.0060 |
| | *MTCP1* | Mature T cell proliferation | 0.0097 |
| | *FLJ23263* | Unknown | 0.0140 |
| | *C6orf47* | Unknown | 0.0141 |

The *p* values indicate statistical significance of $K_a/K_s > 1$ (likelihood ratio test).

of these genes, only 16 were shared between HCR and HCM (Table 5), implying that the use of different outgroups could result in different inference of genes under selection.

We argued that this difference might be due to different outgroups in inferring ancestral sequences of human and chimpanzee. Therefore, we conducted a comparison of pairs of the ancestral sequences of 12 and 10 genes with significant $K_a/K_s > 1$ only in HCR and HCM, respectively. For these genes, we observed sequence differences in the human/chimp ancestral sequences inferred by HCR and HCM, respectively, and 80% of the sequence differences were caused by the sequence difference between macaque and mouse at the corresponding sites, and the human/chimp ancestral sequences were incorrectly inferred in the HCM dataset due to the deep divergence between human/chimp and mouse. For the other 20%, the sequences of the corresponding sites are identical in macaque and mouse. The difference in inferring ancestral sequences at these sites was likely caused by the sampling process because the ancestral sequence inference was based on the statistical evaluation of all sites, and the deep divergence between human/chimp and mouse again would introduce errors.

We noticed that there were 7 genes showing significant positive selection that were not included in the 47 genes analyzed above (*CASP4*, *F2R*, *WDR20*, *ZNF148*, *KCTD10*, *TIAM1*, and *SIAT8C*). This discrepancy was likely due to the sequence length variation between the new HCR dataset and the former HCR dataset since some of the genes in the new HCR dataset have shorter sequences compared with the former HCR dataset to keep the same sequence length between the HCR and the HCM datasets. This pattern suggested that selection could act on a certain region of a gene that is hard to detect when the complete coding region is under scrutiny. Therefore, a gene domain-based analysis might be informative in identifying genes under positive selection that could be neglected in the analyses considering only the complete coding regions.

## Discussion

Previously, Nielsen et al. [9] found that BME genes had undergone strong selective constraint and genes involved in several functional categories had undergone positive selection, but no outgroup was used, and they could not determine whether the effect was specific to the human lineage or the chimpanzee lineage. In our study, we showed that the selective constraint is very strong for BME genes in the chimpanzee lineage with the use of rhesus macaque as outgroup (Table 1). Dorus et al. [11] also found that genes expressed in the nervous system showed a relatively increased evolutionary rate in human compared to chimpanzee.

As indicated in Nielsen et al. [9], we also noted that the immune-defense genes evolve rapidly in both humans and chimps. However, there were three immune-function-related categories (T-cell-mediated immunity, immunity and defense, and defense protein) showing rapid evolution in the chimpanzee lineage, but only one (defense protein) in the human lineage (Tables 2 and 3), implying that there might be more immunity-related genes in chimpanzee undergoing adaptive evolution than in human, which can explain the relatively more intense virus challenge in chimpanzee than in human [31].

In this study, we did not detect several functional categories such as sensory perception, developmental processes that had been noted to undergo positive selection in humans by Clark et al. [8]. It could be due to the relatively small number of genes analyzed (2633) compared to Clark et al. (7645) [8].

We identified 47 genes that had undergone positive selection in the human lineage. These genes are involved in gametogenesis, development, signal transduction, and immune defense and are consistent with the previous studies [8–10]. Additionally, evidence for positive selection on *NPY* (neuropeptide Y) and *EPS15* (related to neurotransmitter release), two nervous-system-related genes, was detected in our study, which was not previously reported.

The observation that different lists of genes under positive selection were obtained when using different outgroups (macaque vs mouse) prompted us to investigate the reasons. We found that the difference was caused mainly by the different outgroup sequences (macaque or mouse) when inferring the ancestral sequences. A closely related outgroup (macaque) has more power and performs better in inferring lineage-specific molecular evolution patterns and identifying genes undergoing adaptive evolution.

## Methods

### *Orthologous coding sequence acquisition*

All the human sequences analyzed were obtained from the NCBI Reference Sequence (RefSeq) database (ftp.ncbi.nih.gov/refseq/Homo–sapiens/); the expression information in the NCBI UniGene database (ftp.ncbi.nih.gov/repository/UniGene/Homo–sapiens/) was checked and then 10,184 RefSeqs expressed in brain were selected. By using the TBLASTN program available in the BLAST v2.2.8 (*E* value cutoff 1E−5), the 10,184 human protein sequences (ftp.ncbi.nih.gov/refseq/Homo–sapiens/) were Blasted against the chimpanzee genome (ftp.ensembl.org/pub/chimp-22.1/). Then SOLAR (Sorting Out Local Alignment Results) v0.0.19, a dynamic program algorithm to link putative exons together, was employed to analyzed the TBLASTN results. There were cases in which the human protein sequences matched more than one chimp genomic DNA segment. The matched results with cutoff <50 (cutoff is the matched protein length/total protein length) were regarded as bad matches, and 1042 genes (10.2%) without eligible matched results were eliminated. After the TBLASTN cutoff selection, each human protein sequence was compared with its chimpanzee matched DNA segment sequences separately using GENEWISE [32].

The GENEWISE score is an additional quality check for similarity of a single DNA sequence and a single protein sequence. First, only GENEWISE results with a score >35 were used [33]. Second, the cutoff (matched protein length/total protein length) and identity (exact matched protein length/total matched length) of the human protein and its one specific corresponding chimpanzee DNA sequence were multiplied and the results with most products were selected. Third, the one DNA sequence matching its corresponding protein with the highest product was selected as the homolog of the protein. In addition to the similarity evaluation, another function of GENEWISE is that the structure (which part is coding and which part is noncoding) of DNA sequence can be inferred according to its matched protein sequence. So after implementation of GENEWISE, 9134 chimpanzee homolog coding sequences (89.6%) were obtained.

Each chimpanzee coding sequence was translated to protein and Blasted against the human genome (ftp.ensembl.org/pub/human-18.34/) by using

TBLASTN again. In this step, the best match between chimp sequence and human genome was selected, then another criterion was employed, i.e., if the genome location of the matched human genome segment coincided with that of the correspondent human gene, that chimpanzee coding sequence was then regarded as the orthologous coding sequence of the human gene. A total of 7070 genes (69.4%) were obtained after this. The remaining genes (7070 genes) were used to search the corresponding orthologs of mouse (ftp.ensembl.org/pub/mouse-31.33g) with the same strategy (BLASTN; GENEWISE) and 4019 orthologs were obtained. Finally, of the 4019 genes, 2633 orthologous coding sequences were obtained after a search in rhesus macaque (ftp.hgsc.bcm.tmc.edu/pub/data/Rmacaque/fasta/Mmul-0.1). The work started before the latest genome data were available, so the number of orthologs seems a little small.

### *Orthologous coding sequence alignment*

In the GENEWISE program, the sequence structure of chimpanzee/macaque/mouse sequence was inferred according to the human protein sequence. There were human genes that did not match in full length, and only the matched parts were used in the analysis. For the comparison of the HCR group and HCM group, to estimate $K_a/K_s$ under the same criterion, the same part and same length of the human–chimpanzee–macaque and human–chimpanzee–mouse orthologous coding sequence of each gene was aligned. About 99.7% of these alignments had a length $\geq 150$ bp.

### *Inference of ancestral sequences and $K_a/K_s$*

The human–chimpanzee ancestral sequences were inferred using the baseml program [14] available in PAML 3.13 [34]. The orthologous sequences from human, chimpanzee, macaque, or mouse were used to infer the human–chimp ancestral sequences. Those sequence sites with different nucleotides in all three species compared (HCM or HCR) were eliminated from the analysis. The yn00 program [15] was employed to estimate $K_s$ and $K_a$ substitution rates corrected for transition/transversion rate bias and codon usage bias. The inferred sequences at the human–chimpanzee ancestral node were compared with sequences at the human–chimpanzee node to calculate lineage-specific $K_a/K_s$ ratios.

To test whether the $K_a/K_s$ ratio was significantly larger than 1, the branch-site model A [16] was applied, the omega ($K_a/K_s$) ratio of each gene in the human-specific or chimp-specific lineage was estimated using the codeml program available in PAML 3.13. The significance tests for the human-specific and chimpanzee-specific adaptive evolution were performed by a maximum likelihood ratio test. Under the alternative hypothesis, the human lineage or chimpanzee lineage was marked as the foreground branch, the omega in the human lineage or chimpanzee lineage was estimated (used parameters were fix_omega=0 and omega=1.5). Under the null hypothesis, the omega in the human lineage or chimpanzee lineage was fixed (used parameters fix_omega=1 and omega=1). Twice the log-likelihood ratio calculated under the two models difference was compared with a $\chi^2$ distribution with $df = 1$ to test whether the omega ($K_a/K_s$) of a gene in the human lineage or the chimpanzee lineage was significantly greater than the background omega and also significantly greater than 1. The branch-site model had been demonstrated to be unreliable in identifying positive selection if used alone by Zhang et al. [35]. Therefore, in our analysis, we combined the results of yn00 with the results of the branch-site model.

### *Functional categories acquisition and analysis*

Functional categories were obtained from http://www.pantherdb.org [18,19]. Genes were selected only if the PANTHER score was better than E−3 and functional classifications were retained for analysis only if the category consisted of at least 20 genes. A total of 71 biological process categories and 61 molecular function categories were identified.

### *Brain-maximal-expressed genes acquisition*

Based on the same gene symbol, the human expression data of the 2633 genes were taken from the Gene Expression Atlas (http://symatlas.gnf.org) [17], which contains Affymetrix chip expression data for 79 human tissues. Many of the expression experiments had replicates; the average expression was taken for each tissue among the replicates. The Affymetrix data had been analyzed by applying the MAS5 condensation algorithm, which reports an average difference (AD) value for each gene [36,37]. Those genes that did not reach an Affymetrix AD value of at least 200, in all 79 tissues, were eliminated from the dataset [38]. Those genes with the maximal AD values in prefrontal cortex, whole adult brain, and fetal brain were regarded as the brain-maximal-expressed genes.

### *Tests of statistical significance*

For identifying an excess of rapidly evolving genes in one functional category, the right-tailed MWU test was used to compare the distribution of $K_a/K_s$ ratios of genes included in the category to the distribution of such ratios in genes not included in the category. To assess the significance that $K_a/K_s$ ratios of BME genes in the human lineage were higher than that in chimp lineage, the right-tailed nonparametric Wilcoxon signed-rank test was used. The test evaluated the likelihood of the null hypothesis that two groups of paired data were drawn from the same underlying distribution [39].

## Acknowledgments

We thank Xue-bin Qi and Hui-feng Jiang from KIZ and Jun Li and Heng Li from BGI for their help in data analysis. We also thank Xiao-na Fan and Yichuan Yu for their technical assistance in this study. This study was supported by grants from the Chinese Academy of Sciences (KSCX2-SW-121), the National Natural Science Foundation of China (30370755, 30440018, 30525028), the Natural Science Foundation of Yunnan Province of China, and the National 973 Project of China (2006CB701506).

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2006.05.008.

## References

[1] R.W. Byrne, A. Whiten, Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans, Clarendon Press, Oxford, 1988.
[2] T. Matsuzawa, Primate Origins of Human Cognition and Behavior, Springer-Verlag, Tokyo, 2001.
[3] W. Enard, et al., Molecular evolution of FOXP2, a gene involved in speech and language, Nature 418 (2002) 869–872.
[4] P.D. Evans, et al., Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans, Hum. Mol. Genet. 13 (2004) 489–494.
[5] J. Zhang, D.M. Webb, O. Podlaha, Accelerated protein evolution and origins of human-specific features: FOXP2 as an example, Genetics 162 (2002) 1825–1835.
[6] J. Zhang, Evolution of the human ASPM gene, a major determinant of brain size, Genetics 165 (2003) 2063–2070.
[7] Y.Q. Wang, B. Su, Molecular evolution of microcephalin, a gene determining human brain size, Hum. Mol. Genet. 13 (2004) 1131–1137.
[8] A.G. Clark, et al., Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios, Science 302 (2003) 1960–1963.
[9] R. Nielsen, et al., A scan for positively selected genes in the genomes of humans and chimpanzees, PLoS Biol. 3 (2005) e170.
[10] C.D. Bustamante, et al., Natural selection on protein-coding genes in the human genome, Nature 437 (2005) 1153–1157.
[11] S. Dorus, et al., Accelerated evolution of nervous system genes in the origin of Homo sapiens, Cell 119 (2004) 1027–1040.

[12] A. Ureta-Vidal, L. Ettwiller, E. Birney, Comparative genomics: genome-wide analysis in metazoan eukaryotes, Nat. Rev., Genet. 4 (2003) 251–262.
[13] W.H. Li, Molecular Evolution, Sinauer, Sunderland, MA, 1997.
[14] Z. Yang, S. Kumar, M. Nei, A new method of inference of ancestral nucleotide and amino acid sequences, Genetics 141 (1995) 1641–1650.
[15] Z. Yang, R. Nielsen, Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models, Mol. Biol. Evol. 17 (2000) 32–43.
[16] Z. Yang, R. Nielsen, Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages, Mol. Biol. Evol. 19 (2002) 908–917.
[17] A.I. Su, et al., A gene atlas of the mouse and human protein-encoding transcriptomes, Proc. Natl. Acad. Sci. USA 101 (2004) 6062–6067.
[18] P.D. Thomas, et al., PANTHER: a library of protein families and subfamilies indexed by function, Genome Res. 13 (2003) 2129–2141.
[19] P.D. Thomas, PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification, Nucleic Acids Res. 31 (2003) 334–341.
[20] C.I. Castillo-Davis, F.A. Kondrashov, D.L. Hartl, R.J. Kulathinal, The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint, Genome Res. 14 (2004) 802–811.
[21] T. Endo, K. Ikeo, T. Gojobori, Large-scale search for genes on which positive selection may operate, Mol. Biol. Evol. 13 (1996) 685–690.
[22] A.L. Hughes, Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells, Mol. Biol. Evol. 14 (1997) 1–5.
[23] S.L. Sawyer, M. Emerman, H.S. Malik, Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G, PLoS Biol. 2 (2004) E275.
[24] N. Takahata, M. Nei, Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci, Genetics 124 (1990) 967–978.
[25] M. Baggiolini, Chemokines and leukocyte traffic, Nature 392 (1998) 565–568.
[26] G. Trinchieri, Interleukin-12 and the regulation of innate resistance and adaptive immunity, Nat. Rev., Immunol. 3 (2003) 133–146.
[27] G. Trinchieri, S. Pflanz, R.A. Kastelein, The IL-12 family of heterodimeric cytokines: new players in the regulation of T cell responses, Immunity 19 (2003) 641–644.
[28] S.Y. Hsu, A.J. Hsueh, Discovering new hormones, receptors, and signaling mediators in the genomic era, Mol. Endocrinol. 14 (2000) 594–604.
[29] S.Y. Hsu, Bioinformatics in reproductive biology—Functional annotation based on comparative sequence analysis, J. Reprod. Immunol. 63 (2004) 75–83.
[30] B. Knudsen, N.R. Farid, Evolutionary divergence of thyrotropin receptor structure, Mol. Genet. Metab. 81 (2004) 322–334.
[31] S.L. Sawyer, L.I. Wu, J.M. Akey, M. Emerman, H.S. Malik, High-frequency persistence of an impaired allele of the retroviral defense gene TRIM5alpha in humans, Curr. Biol. 16 (2006) 95–100.
[32] E. Birney, R. Durbin, Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison, Proc. Int. Conf. Intell. Syst. Mol. Biol. 5 (1997) 56–64.
[33] D. Torrents, M. Suyama, E. Zdobnov, P. Bork, A genome-wide survey of human pseudogenes, Genome Res. 13 (2003) 2559–2567.
[34] Z. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood, Comput. Appl. Biosci. 13 (1997) 555–556.
[35] J. Zhang, Frequent false detection of positive selection by the likelihood method with branch-site models, Mol. Biol. Evol. 21 (2004) 1332–1339.
[36] E. Hubbell, W.M. Liu, R. Mei, Robust estimators for expression analysis, Bioinformatics 18 (2002) 1585–1592.
[37] W.M. Liu, et al., Analysis of high density expression microarrays with signed-rank call algorithms, Bioinformatics 18 (2002) 1593–1599.
[38] G.A. Singer, A.T. Lloyd, L.B. Huminiecki, K.H. Wolfe, Clusters of co-expressed genes in mammalian genomes are conserved by natural selection, Mol. Biol. Evol. 22 (2005) 767–775.
[39] M. Hollander, D.A. Wolfe, Nonparametric Statistical Methods, Wiley, New York, 1999.