Human Molecular Genetics, 2006, Vol. 15, No. 11 **1870–1875**
doi:10.1093/hmg/ddl109
Advance Access published on April 27, 2006

Origination and evolution of a human-specific transmembrane protein gene, *c1orf37-dup*

Haijing Yu[1,2], Huifeng Jiang[2,4], Qi Zhou[2,4], Jufen Yang[1], Yina Cun[1], Bing Su[2,3], Chunjie Xiao[1,*] and Wen Wang[2,3,*]

[1]Key Laboratory of Bioresources Conservation and Utilization and Human Genetics Center of Yunnan University, Kunming 650091, PR China , [2]CAS-Max Planck Junior Research Group and Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology and [3]Kunming Primate Research Center, Chinese Academy of Sciences, Kunming 650223, PR China and [4]Graduate School of Chinese Academy Sciences, Beijing 100039, PR China

Received January 26, 2006; Revised and Accepted April 14, 2006 GenBank (http://www.ncbi.nlm.nih.gov/Genbank) accession nos[†]

A transmembrane protein gene, *c1orf37-dup*, was identified as a young gene specific to humans. It was derived from the conserved *c1orf37* gene through retroposition after the divergence of human and chimpanzee. This gene has evolved rapidly driven by positive Darwinian selection as evident from a significantly high ratio of non-synonymous substitution rate to synonymous substitution rate ($K_a/K_s = 2.08$) between the new *c1orf37-dup* and the parental *c1orf37* genes. Population genetics analysis disclosed a very low level of polymorphism in the *c1orf37-dup* gene and its neighboring regions, thus providing support for the occurrence of a recent selective sweep. The GFP experiments revealed that it encodes a transmembrane protein associated with cell membranes. Non-random distribution of amino acid changes indicates the C1ORF37-DUP protein may have evolved diverged functions in the presumably functionally important N-terminal region in the cytoplasm and the extracellular loop. These lines of evidence support that the functional adaptation of *c1orf37-dup* has occurred in humans. Unlike its ubiquitously expressed parental gene, *c1orf37-dup* expresses selectively in several human tissues including brain. It is suggested that *c1orf37-dup* encodes a novel transmembrane protein in humans which potentially endows new properties to cell surface interactions.

INTRODUCTION

New genes with novel functions may have significantly contributed to the evolution of new phenotypes specific to species. The birth of a new gene comprises duplication, initial mutation events yielding a particular gene structure and subsequent evolutionary procedure, in which it is fixed in the species and improved for novel functions (1). Among the several molecular mechanisms that have been observed in the creation of new gene structures (2), retroposition, usually recruiting a new regulatory sequence to survive, contributes tremendous structural and functional novelties to extant genomes (3). Moreover, positive-Darwinian selection, the important evolutionary driving force, helps newly created genes to become established in the population by increasing the frequency of advantageous mutations (4,5), which has often been observed in gene origination cases (2,6,7).

The novel genes created in the human genome are of special interest because of the unique evolutionary status of human beings and our unique phenotypic traits, such as bipedalism and higher cognitive abilities (8,9). Despite the phenotypic diversities, the genetic divergence between human and chimpanzee at the level of DNA sequence is 1.23% (10–12); therefore, it has been proposed that there may be few novel genes associated with human-specific traits (9). Besides the differences found in gene expression regulation between humans and other primates (9,13,14), human unique

*To whom the correspondence should be addressed at: Kunming Institute of Zoology, No. 32 E. Jiao Chang Road, Kunming, Yunnan 650223, PR China. Tel: +86 8715192979; Fax: +86 8715193137; Email: wwang@mail.kiz.ac.cn or Yunnan University, No. 2 N. Cuihui Road, Kunming, Yunnan 650091, PR China. Tel and Fax: +86 8715034636; Email: cjxiao2005@yahoo.com.cn

[†]The *c1orf37-dup* gene and its flanking region (region A in Fig. 4B): DQ186292–DQ186413. The neighboring region D: DQ186414–DQ186427, region C: DQ186428–DQ186447, region B: DQ186448–DQ186461 and region E: DQ186462–DQ186475.

genes are nevertheless an important source to be investigated to understand the genetic basis of the human-specific traits. On the basis of the observation of segmental duplications in human, Bailey *et al*. (15) estimated that human and chimpanzee genomes may differ by 150–350 transcripts. Brosius (3) predicted about 110 gene differences between humans and chimpanzees assuming the 6 million year divergence time. Marques *et al*. (16) estimated that about one retrogene per million years has emerged on the lineage leading to humans. More recently from the comparison of the human and chimpanzee genomes, the Chimpanzee Sequencing and Analysis Consortium (12) estimated about 200 human-specific retroposed gene copies though the functionality of them remains to be investigated.

In this study, we report the comprehensive characterization of a human-specific gene, *c1orf37-dup*. It was identified in this study to be created through retroposition from the parental intron-containing gene *c1orf37* (the open reading frame 37 on Chromosome 1, CA037_HUMAN, ENSG00000163444). Evolutionary and functional analyses reveal that *c1orf37-dup* has been undergoing rapid functional adaptation driven by strong positive Darwinian selection in humans and encodes a transmembrane protein which may bring new properties to cell surface interactions in human cells.

RESULTS AND DISCUSSION

Identification of the human-specific *c1orf37-dup* gene

In an effort to identify fast evolving human genes, a duplicated copy of the *c1orf37* gene, designated as *c1orf37-dup* (ENSG00000169883, XP_370848.1), was identified on Chromosome 3q25.1 (151,183,864–151,182,148). It contains the hallmarks of retroposed sequences: the lack of introns, a remnant of the poly(A) tail at the 3′ end and 14 bp direct repeats at the both ends (17). It retained the full reading frame (1128 bp) from *c1orf37* without any insertion and deletion (indel) and is predicted to encode a 376-amino acid peptide. Genomic DNA amplification by PCR (see Materials and Methods) with the chimpanzee DNAs and subsequent sequencing disclosed that *c1orf37-dup* is absent from the corresponding genomic location in chimpanzees (Fig. 1). Further investigation in orangutan which shares a common ancestor with the human, chimpanzee and gorilla lineage revealed that orangutan is also devoid of *c1orf37-dup*, reflecting the ancestral state of this genomic locus. These observations demonstrate that *c1orf37-dup* is a gene uniquely created in humans after the divergence of humans and chimpanzees. It is not among the previously reported human-specific retrosequences in Chimpanzee Sequencing and Analysis Consortium (12).

Fast evolution of the *c1orf37-dup* gene

Compared with the parental transcript, *c1orf37-dup* has nine nucleotide changes: two in the 5′-UTR, one in the 3′-UTR and six in the coding region. Our population study shows that they are fixed in human populations. All the six coding region substitutions are non-synonymous and result in amino acid changes.

With the K_a/K_s test, a simple and useful measurement of functional constraint on a protein-coding gene (18), we found *c1orf37-dup* has an excess of non-synonymous substitution rate. The K_a/K_s test measures the ratio of non-synonymous substitution rate (K_a) to synonymous substitution rate (K_s) for a pair of sequences. The ratio is about one for a pseudogene derived from its parental gene because it evolves neutrally. Functional genes have a lower K_a due to the constraint on amino acid changes (i.e. purifying selection), which results in $K_a/K_s < 1$. When a novel function emerges in a gene with strong positive selection involved, $K_a/K_s > 1$ is observed. The K_a/K_s between the *c1orf37-dup* gene and *c1orf37* transcript is 2.08 (0.0073/0.0035, $P < 0.02$, *Z*-test for normal distribution), significantly higher than that of neutrally evolving sequences, suggesting the rapid evolution of *c1orf37-dup* driven by positive selection. In contrast, the parental *c1orf37* gene itself is a much conserved gene. Its homologues are found from *Drosophila* to human. In mammals, the K_a/K_s of the coding region across species is very low, e.g. it is 0 between human and chimpanzee, 0.065 between human and mouse, 0.090 between human and rat and 0.072 between mouse and rat, which are comparable to the average K_a/K_s (0.093) for housekeeping genes between human and mouse (19). Furthermore, *c1orf37* has maintained a single-copy status in a broad range of species, indicating its crucial function and that its copy number is maintained precisely during evolution (7). Consequently, the K_a/K_s between *c1orf37* and *c1orf37-dup* in humans is 27 times the average K_a/K_s in *c1orf37* between human and rodents, reflecting its unusually fast evolution in the short period, since the divergence of human and chimpanzee.

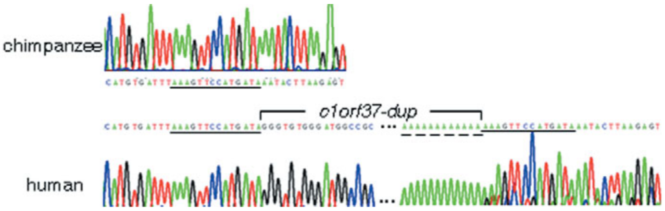

Figure 1. *c1orf37-dup* is present in human but absent in chimpanzee shown by sequencing chromatogram comparison. In the human sequence only the ends of *c1orf37-dup* are shown, and the internal part is indicated by dots. Directional repeat sequences are underlined by solid lines and poly(A) tail is underlined by a dashed line.

Low level of polymorphism and selective sweep affecting the neighboring regions

Rapid adaptive fixation of a new gene is expected to be reflected by the nucleotide polymorphism pattern in populations, as while selection helps to fix the advantageous mutations, it displaces linked nucleotide polymorphisms in the process (5). For this purpose, we studied nucleotide variation features in the *c1orf37-dup* gene and the regions in its neighborhood. When we obtained sequences of the coding region from 61 individuals (122 chromosomes) from different continents, only one polymorphic site, a non-synonymous ATG (Met) to GTG (Val) at the 193rd amino acid, was observed. The nucleotide diversity θ is 1.6×10^{-4} and the frequency of the derived (non-ancestral) GTG codon is high (54%). To obtain more informative sites of nucleotide variation for analyses, we extended sequencing to

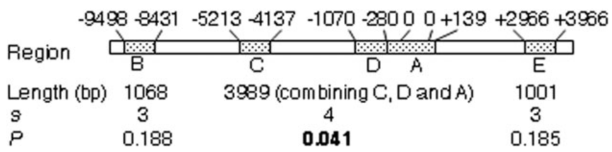

Figure 2. Genomic location of the neighboring regions (dotted boxes) relative to the *c1orf37-dup* gene (between zeros) and comparison between observed and expected polymorphisms in each region are shown. *s* stands for the number of observed polymorphic sites. The *P*-value of significance is shown in bold.

280 bp upstream and 139 bp downstream to *c1orf37-dup* of total 2121 bp in length. The data revealed three more polymorphisms: two singletons locate, respectively, in the 5′ flanking region and the 3′-UTR, and a 1 bp deletion (C) in the 5′-UTR that is linked with the 193GTG. A total of four segregation sites constructing four haplotypes were identified in this fragment despite a large number of samples investigated. A low level of polymorphism ($\theta = 3.5 \times 10^{-4}$) compared with the average θ of the human genome (8.3×10^{-4}) (20) suggests that the evolution of *c1orf37-dup* has been shaped by a recent selective sweep.

If the selective sweep has been sufficiently recent and strong, its effect can be detected not only in the selected gene but also in the surrounding region (5). We labeled the above 2121 bp fragment harboring *c1orf37-dup* as region A and selected four neighboring regions (about 1 kb each) dispersed in 13.5 kb assigned as B, C, D and E which do not contain any annotated functional gene for sequencing (Fig. 2). We compared the observed and the expected numbers of variation in the regions following Wang *et al.* (21) given the expected nucleotide diversity 8.3×10^{-4} from the average θ of the human genome (see Methods and Materials), and the probability values (*P*) are shown in Figure 2. The number of the observed polymorphic sites is significantly lower than the expected ($P = 0.041$) in regions C–D–A, whereas it resumes to match the expected values in regions B and E ($P = 0.188$ and 0.185, respectively). The levels of DNA variation show a trough across this genomic segment, supporting that a strong selective sweep has occurred recently at the *c1orf37-dup* locus in human populations.

Cell localization and functionality of *c1orf37-dup*

The rapid evolution of *c1orf37-dup* instigates more interests on its functional role in cells. Although the extreme conservation of the C1ORF37 protein (peptide ID NP_612400.2) has indicated that it is an important protein in organisms, its function is yet unknown; neither is any homology to other genes or functional domains recognized. A structural analysis with TMpred (see Materials and Methods) showed that the C1ORF37 protein contains two predicted transmembrane regions at the amino acids 150–170 (score 835) and 302–321 (score 2224), suggesting a topology of two termini inside the membrane and a loop on the other side. The transient GFP–C1ORF37 fusion protein expression revealed that GFP signal is largely localized at plasma membranes (Fig. 3A and B). As the transmembrane span (150–170) has a strong preference of the orientation from inside to outside and the span (302–312) amino acid has an orientation preference from outside to inside as indicated in the TMpred analysis, the C1ORF37 protein is possibly a $N_{\rm in}$–$C_{\rm in}$ integral protein with an extracellular loop localized on plasma membranes. In the predicted topology, although remaining open for experimental evidence, the cytoplasmic N-terminal region (150 amino acids) contains two low-complexity strings presumably with important functions (from Ensemble). The extracellular hairpin contains 131 amino acids harboring seven cysteines with the potential to form cystine bonds. On the basis of the information, we speculate that the extracellular hairpin may interact with signals from the environment and the N-terminal region is the reactive region responding to the signals. The GFP–C1ORF37–DUP protein maintains the same localization at plasma membranes (Fig. 3C). Comparing to the C1ORF37 protein, we found that the distribution of the amino acid changes in the C1ORF37–DUP protein is not random over the peptide. Among the six replacements, five replacements N44D, P47T, Q64R, C91A and S106N locate in a span of 66 amino acids in the N-terminal region, and the other replacement R197A is in the extracellular loop close to the 193M/V polymorphic site (Fig. 3D). In a pseudogene, non-synonymous substitutions more likely distribute randomly (22), whereas in a new gene, the changes are expected to evolve in the part of peptide where novel functions may emerge. Therefore, it is implied that novel functions might have evolved in the N-terminal region and the extracellular loop of the C1ORF37–DUP protein.

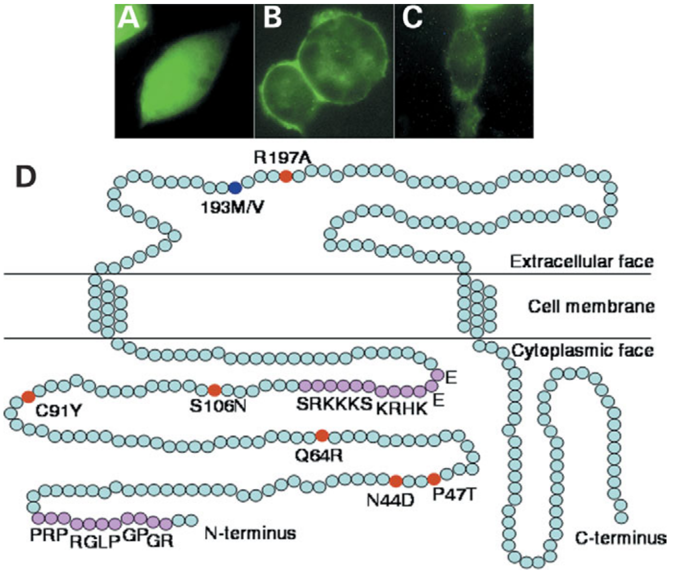

Figure 3. Subcellular location of GFP-tagged proteins in HEK293 cells expressing (**A**) vector pGFP-C1, (**B**) plasmid pGFP-c1orf37 and (**C**) plasmid pGFP-c1orf37-dup. (**D**) The predicted topology of C1ORF37-DUP. Each ball represents an amino acid. Low-complexity regions are shown in pink, replacements in brown and a polymorphic site in dark blue.

Tissue-specific expression

In a retroposition event, the original regulatory elements are generally not co-transferred with the transcript. When mRNA-derived retroposed sequences insert into a promoterless

Table 1. Primer sequences and purposes

Name	Sequence	Purpose
18	TGTCTGCCTTGTAGTATCAAGT	PCR, sequencing
19	AATGAACCTCTACCCTTACATC	PCR, sequencing
U1	GATTACGCCGACTCGGATCT	Sequencing
U2	TGCGTCTGCGACCAGAGTC	Sequencing
U3	TGGTGGCATCCTCAGTACCC	Sequencing
U4	CAAGAGGGAAAACAACTACTATG	Sequencing
U5	AAAAAGTTCCATGATAAATACTTA	Sequencing
L1	GCTTTCTTGACTCGTCCAGAC	Sequencing
L2	CACACAAGCCCGGAGACAGT	Sequencing
L3	GGATGGGAAGCAGTAACTACG	Sequencing
L4	AGATTCTGAAAACCCTGTCACC	Sequencing
Inner-U-D	AGCCGTAGCCAACGCTGT*C*CG*	cDNA PCR
Inner-L-D	TGGAGACAGTTCTCTCATGGA*G*GT*	cDNA PCR
Region B-U	AGCCCTGGTGGCAGCAACTA	PCR, sequencing
Region B-L	TCTCCCCAGTGTAGGATTGAAAGTC	PCR, sequencing
Region C-U	GTTTGCCCCCAGAACTTGGTGT	PCR, sequencing
Region C-L	CCACAGGCTCTGGTACAATAGACAA	PCR, sequencing
Region D-U	ATGTCCATCAGCAGAGGTATG	PCR, sequencing
Region D-L	CACCCAGACACATAAATCCTG	PCR, sequencing
Region E-U	TCACATACTCTACAGCCAGACCAC	PCR, sequencing
Region E-L	GCAGGCAATCCATCAAAACAC	PCR, sequencing
Cloning-U	GGACTCAGATCTATGCCCAAGAGAGGAAAGC	Cloning
Cloning-L	GGTACCGTCGACCGCTCTCAGGGAGAATGGGTACTGA	Cloning

*Mutated nucleotides are shown in italic.

region, they usually degenerate to non-functional pseudogenes (23,24). However, occasionally retroposition leads to active genes when new retrosequences are supported by resident promoter elements or new promoters evolve fortuitously, which often exhibit a different expression pattern from that of the original founder genes (25). Hence, we investigated the *c1orf37-dup*'s transcription capability and its expression pattern in human tissues. Using c1orf37-dup-specific primers (Table 1), its expression was first detected in brain, which bears the most significant changes between human and other primates (Fig. 4). Next, the expression of *c1orf37-dup* was revealed by PCR with human cDNA panels in lung, pancreas, thymus, intestine and blood, but it was undetectable in heart, placenta, liver, muscle, kidney, spleen, prostate, testis, ovary or colon (Fig. 4). These results demonstrate that *c1orf37-dup* is an active gene in transcription and it selectively expresses in several tissues in human. In contrast, the parental *c1orf37* expresses ubiquitously in different tissues based on NCBI's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) and Serial Analysis of Gene Expression map (http://www.ncbi.nlm.nih.gov/SAGE/). These findings indicate that *c1orf37-dup* has acquired transcription regulatory components and become a new gene variant of the parental *c1orf37* gene. Its expression in brain may have important implications relevant to human evolution.

Overall, our study identifies and characterizes a human-specific gene *c1orf37-dup*. The rapid evolution driven by positive selection suggests its unusual functional adaptation in humans. As cell membrane proteins are engaged in the information exchanges between cell–cell and cell–environment, the novel functions of the C1ORF37–DUP protein may endow cells in several human tissues including brain to acquire unique properties in communication or interaction, which do not exist in the ancestral species. Future studies on the functions of the *c1orf37* and *c1orf37-dup* genes will help us to disclose how this new gene creation event is related to human-specific traits formation.

MATERIALS AND METHODS

DNA samples

Sixty-one human genomic DNA samples from different populations in Asia, Africa and Europe were used in this study. The samples were obtained from the following collections: the Kunming Cell Bank of the Chinese Academy of Sciences, Human Genetics Center of Yunnan University, Kunming Blood Center and Shanghai National Genome Center in China. Genomic DNAs from three chimpanzees (*Pan troglodytes*) and one orangutan (*Pongo abelii*) were isolated from the cell lines from the Kunming Cell Bank of the Chinese Academy of Sciences with PUREGENE DNA Isolation Kit (Gentra Systems). Human sample manipulation was approved by Yunnan University and Kunming Institute of Zoology.

Genomic DNA PCR and sequencing

Primers 18 and 19 (Table 1) located in the flanking regions of *c1orf37-dup* were used to amplify the fragment containing this gene with 61 humans (122 chromosomes), three chimpanzees and one orangutan. PCR reactions were assembled quickly and carried out with Takara's ExTaq System at a condition of 95°C 3min, 35 cycles of 94°C 30 s, 54°C 30 s and 72°C 2 min, followed by 72°C 10 min with a hot start. PCR products were purified with Qiagen PCR Purification Kit and subject to sequencing. The internal sequencing primers are listed in Table 1.

Four neighboring genomic regions B, C, D and E were PCR amplified with a sample size of 7, 10, 7 and 7 individuals

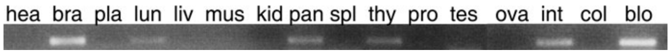

Figure 4. Tissue-specific *c1orf37-dup* expression revealed by cDNA PCRs. Three-letter abbreviations stand for heart, brain, placenta, lung, liver, muscle, kidney, pancreas, spleen, thymus, prostate, testis, ovary, intestine, colon and blood, respectively.

(14, 20, 14 and 14 chromosomes), respectively, and sequenced with their upper and lower primers (Table 1). The PCR condition was 95°C 3 min, 35 cycles of 94°C 30 s, 58°C 30 s and 72°C 1 min, followed by 72°C 7 min with Takara's ExTaq System.

Evolutionary analyses of sequence variation

DNA sequencing reads were assembled, trimmed and aligned with Sequencher 4.0.5 and by manual verification. Synonymous and non-synonymous changes, synonymous and non-synonymous substitution rates and DNA polymorphism (nucleotide diversity θ) were calculated using Dnasp 3.51 (26).

Statistical comparison between observed (s) and expected values of variation followed Wang *et al*. (21). In brief, under the assumption of neutrality, a probability distribution of polymorphism for an expected variation $\theta = 4N\mu$ was calculated following the recursion equation (27).

$$P_n(s) = \sum_{i=0}^{s} P_{n-1}(s-i)\, Q_n(i)$$

where N and μ are effective population size and mutation rate per nucleotide site, respectively and

$$Q_n(i) = \left(\frac{l\theta}{l\theta + n - 1}\right)^i \frac{n-1}{l\theta + n - 1}$$

where l is the length of DNA sequence and n is the number of chromosomes. The average θ of the human genome 8.3×10^{-4} (20) and the chromosome number $n = 15$ were used in the calculation. $n = 15$ was based on 14 chromosomes we sequenced and one chromosome from Entrez sequence.

$$P_2(i) - \left(\frac{\theta}{1+\theta}\right)^i \frac{1}{1+\theta} P_2(i)$$

is the probability of i mutations occurring on the lineages as the most recent common ancestor, which provided the initial condition of the recursion (28).

Plasmid construction, GFP expression and protein structure prediction

The pEGFP-c1orf37 and pEGFP-c1orf37-dup plasmids were constructed to make GFP tagged proteins for cellular localization investigation. The coding regions of *c1orf37* and *c1orf37-dup* were PCR amplified with the cloning primers (Table 1). The PCR products were digested with *Bgl*II and *Sal*I and ligated to pEGFP-C1 vector (Clontech), respectively. Ligations were carried out with T4 ligase (Fermentas) following its protocol. *E. Coli* DHα5 competent cells were transformed and positive clones were amplified in 3ml overnight cultures. Plasmid DNAs were isolated with Qiagen Miniprep Kit.

HEK293 cells were maintained with DMEM supplemented by 10% new born calf serum and 1% L-glutamine at 37°C with 5% CO_2. The pEGFP-c1orf37, pEGFP-c1orf37-dup and vector pEGFP-C1 were transiently expressed in HEK293 cells. Cell spreading and transfection procedures followed the manufacturer's manual of Lipofectamine™ 2000 (Invitrogen) in a 24-well plate. Twenty four hours after the transfection, cells were fixed as suggested by the manual and observed with fluorescent microscopy.

TMpred (http://www.ch.embnet.org/software/TMPRED_form.html) was used to predict protein transmembrane region. Scores over 500 were considered significant.

Expression survey with human tissue cDNAs

The expression of *c1orf37-dup* was investigated by PCR with Human cDNA panels from BD Biosciences, which contained cDNAs from heart, whole brain, placenta, lung, liver, skeletal muscle, kidney, pancreas, spleen, thymus, prostate, testis, ovary, small intestine, colon and blood. Owing to the high degree of sequence similarity between *c1orf37* and *c1orf37-dup*, primers inner-U-D and inner-L-D (Table 1) were designed to amplify exclusively the *c1orf37-dup* cDNA. The *c1orf37* and *c1orf37-dup* have few nucleotide differences in cDNA sequences. On the basis of these differences, the position of the primers was chosen so that the 3′-ending nucleotide of each primer matches to *c1orf37-dup* but not *c1orf37*. In addition, the third nucleotide at the 3′ of each primer was mutated, which resulted in a pair of primers for *c1orf37-dup* with an affordable quantitative loss in PCR product but primers with the least possibility to pick up *c1orf37* cDNA because of two 3′ mismatching nucleotides. PCRs were performed with Expand High Fidelity PCR System (Roche) under a strict condition at 95°C 10 min, 40 cycles of 94°C 30 s and 68°C 1 min and followed by 68°C 10 min. Resulting PCR fragments were cloned with TOPO TA Cloning Kit for Sequencing (Invitrogen) according to its manual and sequenced to confirm their *c1orf37-dup* cDNA identity. Sequencing was carried out with the BigDye3.0 protocol using an ABI3100 autosequencer.

ACKNOWLEDGEMENTS

We thank M. Long for constructive discussions, J.J. Emerson for initial computational analysis of the new gene candidates, W. Xiao, K. Kristiansen, K. Thornton and Wang lab members for helpful comments, X. Fan for sequencing and members in the Kunming Cell Bank of the Chinese Academy of Sciences for help in cell culture. We are grateful to F. Yang and F. Wilson for the critical reading of the manuscript. This project was supported by the Key Project KSCX2-SW-121 from Chinese Academy of Sciences to W.W. and B.S., the Science and Technology Key Project 2005NG08 from Yunnan Province and the Western Chinese Region Development Key Project 2004BA901A07 to C.X. and the grant 30560063 from National Science Foundation of China to H.Y.

Conflict of Interest statement. The authors have no conflicts of interest to declare.

REFERENCES

1. Long, M. (2001) Evolution of novel genes. *Curr. Opin. Genet. Dev.*, **11**, 673–680.
2. Long, M., Betran, E., Thornton, K. and Wang, W. (2003) The origin of new genes, glimpses from the young and old. *Nat. Rev. Genet.*, **4**, 865–875.
3. Brosius, J. (2003) The contribution of RNAs and retroposition to evolutionary novelties. *Genetica*, **118**, 99–116.
4. Walsh, J.B. (1995) How often do duplicated genes evolve new functions? *Genetics*, **139**, 421–428.
5. Li, W.H. (1997) *Molecular Evolution*, Sinaur Associates, Sunderland, Massachusetts.
6. Burki, F. and Kaessmann, H. (2004) Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat. Genet.*, **36**, 1061–1063.
7. Ciccarelli, F.D., von Mering, C., Suyama, M., Harrington, E.D., Izaurralde, E. and Bork, P. (2005) Complex genomic rearrangements lead to novel primate gene function. *Genome Res.*, **15**, 343–351.
8. Gibbons, A. (1998) Which of our genes makes us human? *Science*, **281**, 1432–1434.
9. Nahon, J.L. (2003) Birth of 'human-specific' genes during primate evolution. *Genetica*, **118**, 193–208.
10. Chen, F.C. and Li, W.H. (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.*, **68**, 444–456.
11. Ebersberger, I., Metzler, D., Schwarz, C. and Paabo, S. (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.*, **70**, 1490–1497.
12. Chimpanzee Sequencing Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
13. Enard, W., Khaitovich, P., Klose, J., Zollner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R. *et al*. (2002) Intra- and interspecific variation in primate gene expression patterns. *Science*, **296**, 340–343.
14. Khaitovich, P., Muetzel, B., She, X., Lachmann, M., Hellmann, I., Dietzsch, J., Steigele, S., Do, H.H., Weiss, G., Enard, W. *et al*. (2004) Regional patterns of gene expression in human and chimpanzee brains. *Genome Res.*, **14**, 1462–1473.
15. Bailey, J.A., Yavor, A.M., Viggiano, L., Misceo, D., Horvath, J.E., Archidiacono, N., Schwartz, S., Rocch, M. and Eichler, E.E. (2002) Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.*, **70**, 83–100.
16. Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A. and Kaessmann, H. (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.*, **3**, e357.
17. Weiner, A.M., Deininger, P.L. and Efstratiadis, A. (1986) Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.*, **55**, 631–661.
18. Nekrutenko, A., Makova, K.D. and Li, W.H. (2002) The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.*, **12**, 198–202.
19. Zhang, L. and Li, W.H. (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.*, **21**, 236–239.
20. Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R. and Chakravarti, A. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.*, **22**, 239–247.
21. Wang, W., Thornton, K., Emerson, J.J. and Long, M. (2004) Nucleotide variation and recombination along the fourth chromosome in *Drosophila simulans*. *Genetics*, **166**, 1783–1794.
22. Evans, P.D., Anderson, J.R., Vallender, E.J., Choi, S.S. and Lahn, B.T. (2004) Reconstructing the evolutionary history of microcephalin, a gene controlling human brain size. *Hum. Mol. Genet.*, **13**, 1139–1145.
23. Vanin, E.F. (1985) Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.*, **19**, 253–272.
24. Ophir, R. and Graur, D. (1997) Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene*, **205**, 191–202.
25. Brosius, J. (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene*, **238**, 115–134.
26. Rozas, J. and Rozas, R. (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics*, **15**, 174–175.
27. Hudson, R.R. (1990) Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.*, **7**, 1–42.
28. Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, **7**, 256–276.