carriers. Avoidance learning, which was faster with the larger aggregation, required a strong signal, but the signal had no initial cost when the prey were gregarious. The signal increased detectability only slightly and detectability costs resulting from both the signal and aggregation were counterbalanced by the dilution effect, by decreased per capita encounter probability and possibly by neophobia. Notably, the dilution effect increased the survival of aggregated unpalatable prey even without a strong warning signal. Thus, unpalatability alone could select for grouping under the influence of individual selection, and in groups the evolution of a stronger signal would be favoured by synergistic selection[6,16], which affects individuals of the same phenotype regardless of their ancestral relatedness. Alternative explanations for the initial evolution of warning coloration have been proposed[17], such as random drift, neophobia[14], evolution through individual selection when prey are by some means able to survive an attack[18–20], or the coloration being cryptic from a distance but aposematic when the predator is close[21]. Most of these hypotheses are not mutually exclusive, and different mechanisms may have been important in distinct areas or with different species, thus we do not claim that gregariousness is a prerequisite for the evolution of warning signals. However, given the dilution effect, the small detectability costs of signals in groups and the enhanced learning of strong signals in groups, it seems that gregariousness of unpalatable prey might have enabled the initial appearance of aposematism, and grouping may assist in the survival of established aposematic prey whenever the prey encounter naive predators. □

Methods

Predators and prey

Wild great tits were caught in mist nets around Konnevesi Research Station where the experiments were carried out from January to May 1997 (general methods as in ref. 9). Each bird was trained to open similar paper prey items to the ones that were eventually used in the experiments, but during the training the prey items had no signal. After the experiment we released great tits to the area where they were caught.

Detectability experiment

All of the prey items were palatable as the objective was to find out how group size and signal conspicuousness affect the number of prey attacked owing to detectability differences. Signal 1 (the background signal) was not used because in a separate visibility test with solitary prey items, signal 1 and signal 2 did not differ significantly in their conspicuousness to the great tits (result reported in ref. 9). Before the detectability test, the birds ($n = 11$) were given palatable prey items that displayed all of the signals used so as to avoid neophobic reactions towards any of the signals. Eating or touching the prey item was taken as an indication that the bird had seen the prey, as the birds had no reason to avoid any of the prey types. The trial continued until the bird had attacked 20 prey assemblages, but only the first 15 assemblages were included in the final analysis to avoid the risk that prey depletion during feeding would bias the detectability estimation. The experiment was repeated the next day and the mean values from two trials were used in further calculations, because the trials gave similar results. The data were analysed with a two-way analysis of variance (ANOVA) with main effects (SPSS for Windows version 7.0).

Learning experiment

We tested selection pressures on evolving aposematic prey by using palatable cryptic prey items with signal 1 together with unpalatable items with either signal 2 or signal 4. Signal 3 was not included because it did not differ much from signal 2 (see Fig. 2). Each bird was randomly assigned to one of the six treatment groups (two signal strengths × three group sizes), so that every treatment had 7–9 birds (total $n = 48$ birds).

In each treatment half of the prey items in the aviary were cryptic, palatable and solitary, whereas the other half were aposematic (unpalatable and displaying a signal) and were placed either solitarily, in groups of four or in groups of eight. The number of prey items was always the same: 24 palatable and 24 distasteful items in the aviary. The birds were allowed to taste 15 prey items in each of the five trials. The number of unpalatable prey items eaten in a trial was used as a dependent variable in repeated measures ANOVA (SPSS for Windows version 7.0), with trial as the within-subject factor (corresponding to learning) and group size as a between-subject factor. Separate tests were performed for the two signals. The data for solitary treatment, which serves as the control here, were obtained from another experiment that was performed at the same time[9].

Received 9 February; accepted 13 August 2001.

1. Endler, J. A. Frequency-dependent predation, crypsis and aposematic coloration. *Phil. Trans. R. Soc. Lond. B* **319,** 505–523 (1988).
2. Guilford, T. in *Insect Defenses. Adaptive Mechanisms and Strategies of Prey and Predators* (eds Evans, D. L. & Schmidt, J. O.) 23–61 (State Univ. New York Press, New York, 1990).
3. Mallet, J. & Singer, M. C. Individual selection, kin selection, and the shifting balance in the evolution of warning colours: the evidence from butterflies. *Biol. J. Linn. Soc.* **32,** 337–350 (1987).
4. Schuler, W. & Roper, T. J. Responses to warning coloration in avian predators. *Adv. Study Behav.* **21,** 111–146 (1992).
5. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon, Oxford, 1930).
6. Alatalo, R. V. & Mappes, J. Tracking the evolution of warning signals. *Nature* **382,** 708–710 (1996).
7. Tullberg, B. S., Leimar, O. & Gamberale-Stille, G. Did aggregation favour the initial evolution of warning coloration? A novel world revisited. *Anim. Behav.* **59,** 281–287 (2000).
8. Alatalo, R. V. & Mappes, J. Initial evolution of warning coloration: comments on the novel world method. *Anim. Behav.* **60,** F1–F2 (2000).
9. Lindström, L., Alatalo, R. V., Mappes, J., Riipi, M. & Vertainen, L. Can aposematic signals evolve by gradual change? *Nature* **397,** 249–251 (1999).
10. Treisman, M. Predation and the evolution of gregariousness. I. Models for concealment and evasion. *Anim. Behav.* **23,** 779–800 (1975).
11. Turner, G. F. & Pitcher, T. J. Attack abatement: a model for group protection by combined avoidance and dilution. *Am. Nat.* **128,** 228–240 (1986).
12. Sillén-Tullberg, B. & Leimar, O. The evolution of gregariousness in distasteful insects as a defense against predators. *Am. Nat.* **132,** 723–734 (1988).
13. Hamilton, W. D. Geometry for the selfish herd. *J. Theor. Biol.* **31,** 295–311 (1971).
14. Lindström, L., Alatalo, R. V., Lyytinen, A. & Mappes, J. Predator experience on cryptic prey affects the survival of conspicuous aposematic prey. *Proc. R. Soc. Lond. B* **268,** 357–361.
15. Gagliardo, A. & Guilford, T. Why do warningly-coloured prey live gregariously? *Proc. R. Soc. Lond. B* **251,** 69–74 (1993).
16. Guilford, T. Is kin selection involved in the evolution of warning coloration? *Oikos* **45,** 31–36 (1985).
17. Mallet, J. & Joron, M. Evolution of diversity in warning color and mimicry: polymorphisms, shifting balance, and speciation. *Annu. Rev. Ecol. Syst.* **30,** 201–233 (2000).
18. Järvi, T., Sillén-Tullberg, B. & Wiklund, C. The cost of being aposematic. An experimental study of predation on larvae of *Papilio machaon* by the great tit *Parus major*. *Oikos* **36,** 267–272 (1981).
19. Sillén-Tullberg, B., Wiklund, C. &. Järvi, T. Aposematic coloration in adults and larvae of *Lygaeus equestris* and its bearing on mullerian mimicry: an experimental study on predation on living bugs by the great tit *Parus major*. *Oikos* **39,** 131–136 (1982).
20. Marples, N. M., van Veelen, W. & Brakefield, P. M. The relative importance of colour, taste and smell in the protection of an aposematic insect *Coccinella septempunctata*. *Anim. Behav.* **48,** 967–974 (1994).
21. Weismann, A. *Studies in the Theory of Descent* (Sampson, Low, Marston, Searle and Rivington, London, 1882).

Acknowledgements

We thank H. Nisu, L. Vertainen and the Academic Hobby Crafts Club for their help, and Konnevesi Research Station for the facilities. We also thank A. Chaine, G. Corrigan, A. Kause, R. Leimu, B. Lyon, A. Lyytinen, J. Tuomi and B. Weaver for valuable comments on earlier versions of the manuscript. This study was supported by the Academy of Finland. Authors after M.R. are in alphabetical order.

Correspondence and requests for materials should be addressed to M.R. (e-mail: marianna.riipi@utu.fi) or J.M. (e-mail: mappes@cc.jyu.fi).

Positive selection of a gene family during the emergence of humans and African apes

Matthew E. Johnson*, Luigi Viggiano†, Jeffrey A. Bailey*, Munah Abdul-Rauf‡, Graham Goodwin‡, Mariano Rocchi† & Evan E. Eichler*

* *Department of Genetics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106, USA*
† *DAPEG, Sezione di Genetica, Via Amendola 165/A, 70126 Bari, Italy*
‡ *Section of Molecular Carcinogenesis, Institute of Cancer Research, Haddow Laboratories, 15 Cotswold Road, Sutton, Surrey MS2 5NG, UK*

Gene duplication followed by adaptive evolution is one of the primary forces for the emergence of new gene function[1]. Here we describe the recent proliferation, transposition and selection of a 20-kilobase (kb) duplicated segment throughout 15 Mb of the short arm of human chromosome 16. The dispersal of this segment was accompanied by considerable variation in chromosomal-map location and copy number among hominoid species. In humans, we identified a gene family (*morpheus*) within the duplicated segment. Comparison of putative protein-encoding

exons revealed the most extreme case of positive selection among hominoids. The major episode of enhanced amino-acid replacement occurred after the separation of human and great-ape lineages from the orangutan. Positive selection continued to alter amino-acid composition after the divergence of human and chimpanzee lineages. The rapidity and bias for amino-acid-altering nucleotide changes suggest adaptive evolution of the *morpheus* gene family during the emergence of humans and African apes. Moreover, some genes emerge and evolve very rapidly, generating copies that bear little similarity to their ancestral precursors. Consequently, a small fraction of human genes may not possess discernible orthologues within the genomes of model organisms.

During physical mapping and sequencing of the human genome[2–5], a complex series of duplicated genomic segments were identified that mapped to multiple cytogenetic band positions on chromosome 16 (Fig. 1a). We reassessed the genomic distribution and the extent of duplication of one of these segmental duplications, termed LCR16a (low-copy repeat sequence 'a' from chromosome 16)[3]. Fifteen distinct copies of the duplicated segment were characterized (see Methods and Supplementary Information Fig. 1). These genomic repeats were specific to human chromosome 16, were ~20 kb long, and shared a remarkably high degree of sequence identity (Fig. 1b). Searches for sequence similarity in the expressed sequence divisions of GenBank revealed a previously uncharacterized family of genes within the LCR16a segment (Fig. 1c and see Supplementary Information). Transcripts could be identified for 6 of the 15 genomic copies. We found no significant sequence similarity to this gene family in other organisms either at the nucleotide or protein level (sequence similarity values $E < 10^{-30}$ and $E < 10^{-2}$, respectively), indicating a highly diverged family of human transcripts. Sequence comparison of putative proteins from two full-length human transcripts showed 81% amino-acid sequence identity (see Supplementary Information Fig. 2). In sharp contrast, the corresponding non-coding portions of genomic DNA were 98.1% identical. These data suggested either that the exonic regions were hypermutable or that amino-acid changes had been selected during the evolution of this gene family.

The high degree of genomic sequence similarity among the various human copies (~98%) indicated a recent evolutionary divergence. Analysis by fluorescence *in situ* hybridization (FISH) of primate metaphase chromosomes and interphase nuclei confirmed marked variation in signal intensity, copy number and map location (Fig. 2). Among all Old World monkeys, a single metaphase signal corresponding to one or two copies (by interphase nuclei) was identified distally on chromosome 16. Sequence analysis of a genomic subclone from baboons (data not shown) confirmed an orthologous map position to human sequence 16p13.1—the probable ancestral segment from which all other copies originated. In contrast to the Old World monkeys, the genome of the great apes showed a major proliferation of the LCR16a duplicon. This is particularly evident within the short arm, which is almost completely 'painted' by the LCR16a probe. This effect is most striking in the lineages of humans and African great apes. Using a combination

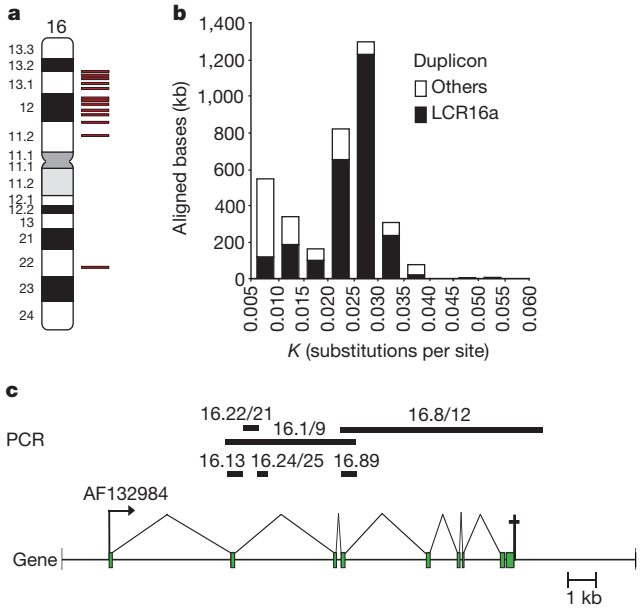

Figure 1 Sequence properties of the LCR16a duplication. **a**, Schematic display of the distribution pattern (red bars) of LCR16a duplications relative to a human chromosome 16 ideogram. The analysis is based on the published human genome project assembly[4,5] and shows the clustering of duplications on the short arm of chromosome 16. **b**, The number of substitutions per site (*K*) among LCR16 duplications as a function of the number of aligned base pairs. Optimal global alignments for all possible pairwise combinations for each duplication were made and the degree of sequence identity for each alignment was computed ($n = 183$ pairwise alignments). LCR16a duplications are compared to all other characterized LCR16 duplications. (See Supplementary Information Fig. 1b for a detailed description of other duplications.) **c**, The gene structure of one member of the gene family (AF132984) is shown (green bars) compared with the 20-kb LCR16a segment from its corresponding genomic locus (AC002045). The analysis indicates eight exons, two strong polyadenylation signals within the 3′ untranslated region, and a putative promoter region overlapping the first exon. PCR products used as probes in this study and their relative location are indicated above the gene structure.

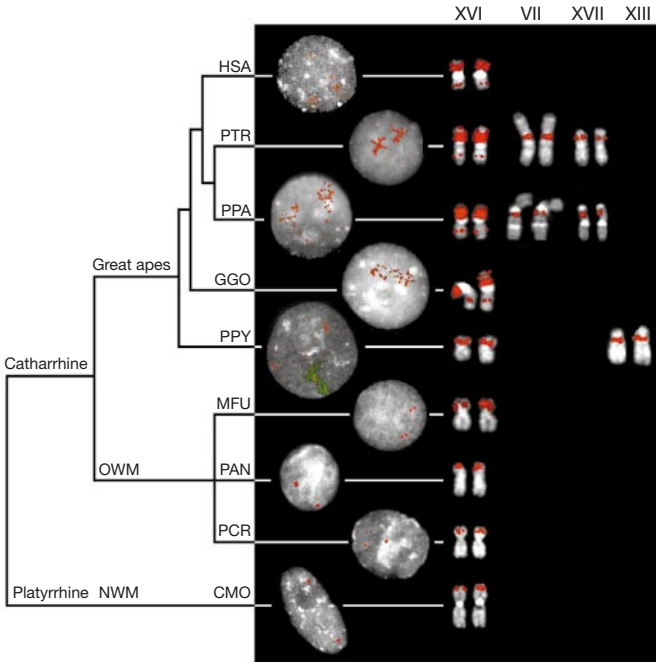

Figure 2 Comparative FISH analysis among primates. Metaphase (right) and interphase nuclei (left) that have been hybridized with probes (16.1/9 and 16.8/12) are shown from a representative panel of New World monkeys (NWM: CMO, *Callicebus mollochus*), Old World monkeys (OWM: MFU, *Macaca fascicularis*; PAN, *Papio anubis*; PCR, *Presbytis cristata*) and hominoid species (HSA, *Homo sapiens*; PTR, *Pan troglodytes*; PPA, *Pan paniscus*; GGO, *Gorilla gorilla*; PPY, *Pongo pygmaeus*). The results are depicted in the context of a generally accepted phylogeny of the species[6]. Roman numerals above metaphase chromosomes accord to standard cytogenetic nomenclature. Note that the multiple copies of the repeat located on XVI among the hominoids seem to paint the short arm of the chromosome. Reciprocal experiments using probes derived from other primate species were used to eliminate the possibility of false negative signal (see Methods). The orangutan (PPY) interphase also shows hybridization of a human chromosome XVI paint (green fluorescence). In this species, copies of the LCR16a duplicon have spread to the pericentromeric region of chromosome XIII.

Table 1 Average pairwise distance ($\bar{K}$) of intron sequence

Species	HSA	PTR	HKL	PHA
HSA	–	0.021	0.034	0.074
PTR	0.002	–	0.038	0.080
HKL	0.004	0.004	–	0.070
PHA	0.008	0.008	0.007	–

HSA, *Homo sapiens* ($n = 14$ sequences); PTR, *Pan troglodytes* ($n = 17$); HKL, *Hylobates* ($n = 4$); and PHA, *Papio hamadryas* ($n = 1$); n is the number of paralogues analysed within each species. $\bar{K}$ is the average genetic distance (Kimura two-parameter model) between groups (above the diagonal); standard errors are given below the diagonal.

of approaches (interphase nuclei, library hybridization and sequence analysis of genomic clones), we estimated the copy number of the duplication in orangutans, gorillas, humans and chimpanzees as 9, 17, 15 and 25–30 copies, respectively. Interestingly, in both orangutans and chimpanzees, copies have been transposed to chromosomes other than 16 (Fig. 2), clearly indicating lineage-specific duplication events.

To more precisely estimate evolutionary timing of the duplications, we resequenced 1,421 base pairs (bp) of non-coding intronic sequence (Fig. 1c) from various human, chimpanzee, gibbon and baboon genomic subclones, and compared the number of nucleotide substitution events both within and between species (Table 1). First, we performed a Tajima's relative rate test using baboon sequence as an outgroup and orthologous pairs of sequence from chimpanzee, human and gibbon. True orthologues were determined by end-sequence analysis of the genomic subclones (see Methods). This allowed duplicated copies that were orthologous by position to be identified within their respective genomes. On the basis of the analysis of four tests ($P = 0.109, 0.239, 0.242$ and 0.093), which indicated that the intronic sequence was evolving neutrally, we accepted the molecular-clock hypothesis. Using 25 Myr as an estimate of the timing of separation between humans and the Old World monkeys[6], we calculated the mean rate of nucleotide substitution for intronic sequence as $(1.5 \pm 0.14) \times 10^{-9}$ substitutions per site per year. This estimate is remarkably consistent with previously published neutral rates between human and Old World monkeys ($1.2–1.8 \times 10^{-9}$) (ref. 7). On the basis of this rate of nucleotide substitution, we predict that the duplication events identified within the gibbons and orangutans occurred independently from those of humans and great apes. In the case of humans and chimpanzees, our analysis indicates that duplications occurred both before and after the separation of these two lineages (Fig. 3a). The observed quantitative and qualitative differences in the localization of some of the copies (Fig. 2) support this conclusion.

Alignments of the human paralogous segments revealed that regions corresponding to coding exons were conspicuously hypervariable (10% nucleotide divergence when compared with intronic sequences that exhibited ~2% divergence). The increased frequency of substitution suggested rapid genic evolution had occurred along with the genomic dispersal of the LCR16a duplication. Increased substitutions among exons are a hallmark of genes undergoing adaptive evolution[8–11]. A common test of positive selection is to

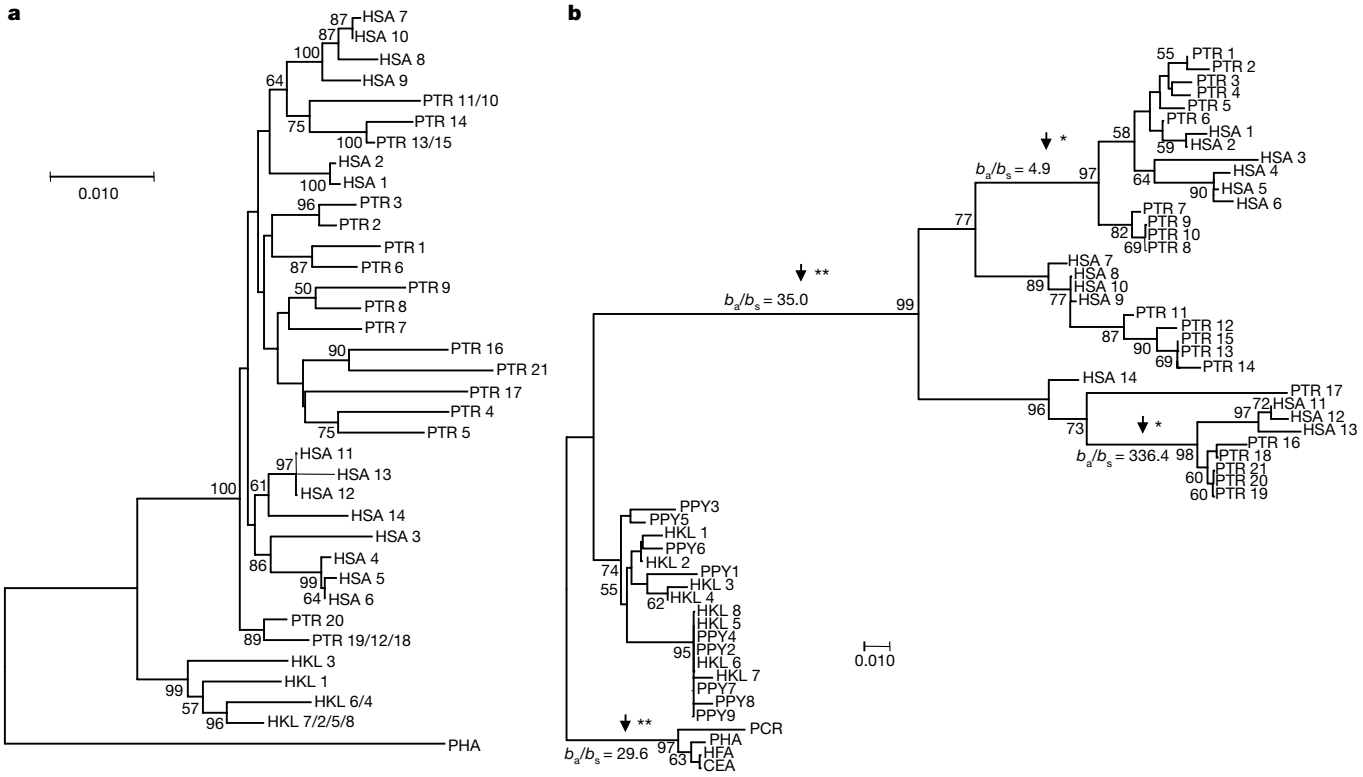

Figure 3 Phylogeny of coding and non-coding portions of the LCR16a duplication. Neighbour-joining phylogenetic trees for 1,421 bp of intronic sequence (**a**; introns 2, 3 and 4; Fig. 1c) and 186 bp of exon 2 (**b**) are compared. Extreme positive selection for exon 2 is indicated on the branch separating humans and African apes from the orangutan lineage (a 35-fold excess of amino-acid changes when compared with the neutral model). Note the significantly shorter branch lengths for flanking non-coding intronic sequences, which are consistent with nucleotide sites evolving at a neutral rate. More than 95% of the informative sites for the phylogenetic tree of exon 2 are the result of amino-acid-altering nucleotide changes. Branches showing significant positive selection are indicated by arrows with accompanying b_a/b_s quotients (estimated amino-acid replacement and synonymous changes per branch per site[12]). Significance was calculated based on the difference (asterisk, $P < 0.05$; double asterisk, $P < 0.01$). Sequence for various duplicate copies are identified by species acronym (PHA, *Papio hamadryas*; HKL, *Hylobates klossi*; and see Fig. 2) and a number corresponding to clone and/or accession within GenBank (Supplementary Information Table 3). Scale bar, Jukes–Cantor corrected distance. The midpoint of all trees was set to one-half the distance between gibbon and baboon sequence taxa. Only bootstrap values >50% are shown ($n = 1,000$ replicates). A similar topology showing positive selection for exon 4 sequence was obtained (Supplementary Information Fig. 3).

compare the number of non-synonymous substitutions per site (K_a) to the number of synonymous substitutions per site (K_s)[7]. K_a/K_s quotients significantly greater than 1.0 are taken as evidence for positive selection. We assessed the K_a/K_s quotient for two of the most rapidly diverging exons (exons 2 and 4; Fig. 3b and Supplementary Information Fig. 3). Genomic subclones were obtained for the various copies of LCR16a from chimpanzees, gorillas, orangutans, gibbons and several Old World monkeys, and the exonic regions were comparatively sequenced (see Methods). Average K_a and K_s for all between-group and within-group comparisons was calculated independently for each exon (MEGA2, Modified Nei–Gojobori method; Tables 2 and 3). A statistical test of the difference of average K_a and K_s values both between species and for multiple copies within species was used as a measure of significance.

In the case of exon 2, highly significant K_a/K_s quotients ($P < 0.0005$) were observed among all comparisons involving either humans or chimpanzees. The most extreme positive selection was observed between humans and Old World monkeys ($K_a/K_s = 13.0$, $P < 10^{-5}$) and between chimpanzees and Old World monkeys ($K_a/K_s = 11.8$, $P < 10^{-5}$). This level of amino-acid replacement translates into ~43% amino-acid divergence between these species and a rate of amino-acid replacement of $\sim 1.0 \times 10^{-8}$ changes per site per year for this exon. This is far in excess (20-fold) of most typical estimates[7] of protein divergence between Old World and great ape species. Highly significant differences were also observed when comparing paralogues between chimpanzee and human sequences ($K_a/K_s = 5.0$, $P < 0.0001$) with an average amino-acid divergence of 23% among the paralogous exons. To identify more precisely when the major episode of positive selection occurred, we estimated the number of synonymous and non-synonymous nucleotide substitutions per site for each branch of a phylogenetic tree using the method first proposed by Zhang *et al.*[12]. A major burst of positive selection seems to have occurred after the separation of the human and chimpanzee lineages from the orangutan (<12 Myr ago, $K_a/K_s = 35.0$), with subsequent protein-diversification events occurring during the emergence of chimpanzee and human species (see Fig. 3). A comparison with gorillas (Table 2) confirms that the major effect occurred in a common ancestor to humans and African apes. In stark contrast, the paralogues within orangutan and gibbon species have not experienced bursts of rapid positive selection (Table 2).

Similar to exon 2, analysis of exon 4 sequences showed a significant episode of positive selection after the separation of the chimpanzee/human and orangutan lineages ($K_a/K_s = 4.67$, $P < 0.05$; Supplementary Information Fig. 3). Although there is a marked increase in the number of putative amino-acid replacements (~30% between chimpanzees/humans and Old World monkeys), much more modest K_a/K_s quotients (1.81–2.45, Table 3) are observed in comparisons between and within species. This reduction is primarily due to the greater number of synonymous events that have occurred concurrently with non-synonymous changes. These events have occurred precisely within the putative coding regions of the exons and do not extend into flanking intronic sequences. Trivial explanations for the enhanced rates of non-synonymous and synonymous changes were examined, including unusual CpG (cytosine–guanine) content, codon bias and the presence of hypermutable repeat sequences within the exonic regions. No evidence in support of these alternatives could be found. It should be noted, however, that alternative splicing has been observed for exon 4 among human complementary DNAs (for example AF229069, D86974 and Supplementary Information Fig. 2). It is possible that some paralogues in different species have also experienced alternative splicing, resulting in new protein products with open reading frames that no longer conform to that predicted by our human reference cDNA (AF132984). The result of such an apparent frameshift would be to increase both synonymous and non-synonymous rates if both splice variants were represented in each species. Furthermore, no distinction in this study has been made between functional and non-functional paralogues, because this would require detailed expression and protein analyses. We felt that this treatment was conservative as pseudogene comparisons and alternative splicing would tend to neutralize both adaptive and purifying selection constraints. Consequently, such events would cause K_a/K_s quotients to approximate unity and reduce our power to detect positive selection.

Although the precise function of this gene family is unknown, previous examples of positive selection have included either genes involved in xenobiotic recognition of macromolecules (immunoglobulin genes, venom toxins, lysozymes)[10,12–14] or genes associated with male reproduction[8,9,11,15]. In many of these cases, positive selection has occurred in concert with duplication events. Delineation of the function of this gene family will require detailed experimental analysis. In humans, multiple transcripts ($n = 284$) with open reading frames have been recovered, demonstrating clear transcriptional and splicing potency. Analysis by polymerase chain reaction with reverse transcription (RT-PCR) confirms a broad distribution of this gene family in most human tissues (Supplementary Information Fig. 4). Finally, immunolocalization studies performed with constructs fused with green fluorescent protein (GFP) reveal a clear localization to the nuclear membrane for at least one translated member of this gene family. Colocalization of these products with antibodies raised against membrane-bound nucleo-

Table 2 Positive selection of exon 2

Exon 2	$\bar{K}_a$ (s.e.)	$\bar{K}_s$ (s.e.)	$\bar{K}_a/\bar{K}_s$	$\bar{K}_a/\bar{K}_s$ (s.e.)	*Z*-value	*P*
HSA–PTR	0.189 (0.035)	0.042 (0.021)	4.5	0.147 (0.043)	3.42	<0.0001
HSA–GGO	0.176 (0.031)	0.066 (0.026)	2.67	0.110 (0.036)	3.06	<0.01
HSA–PPY	0.350 (0.065)	0.097 (0.048)	3.61	0.254 (0.080)	3.18	<0.0005
HSA–HKL	0.345 (0.062)	0.098 (0.045)	3.52	0.247 (0.077)	3.21	<0.0005
HSA–OW	0.429 (0.080)	0.033 (0.016)	13.00	0.396 (0.081)	4.89	<0.00001
PTR–GGO	0.182 (0.035)	0.069 (0.028)	2.64	0.114 (0.041)	2.78	<0.01
PTR–PPY	0.341 (0.064)	0.101 (0.048)	3.38	0.240 (0.077)	3.11	<0.0005
PTR–HKL	0.334 (0.062)	0.102 (0.046)	3.27	0.232 (0.075)	3.09	<0.01
PTR–OW	0.423 (0.078)	0.036 (0.016)	11.75	0.386 (0.078)	4.95	<0.00001
GGO–PPY	0.361 (0.067)	0.112 (0.048)	3.22	0.248 (0.083)	2.99	<0.01
GGO–HKL	0.350 (0.066)	0.113 (0.047)	3.10	0.244 (0.081)	2.77	<0.01
GGO–OW	0.420 (0.078)	0.054 (0.024)	7.78	0.366 (0.077)	4.75	<0.0001
PPY–HKL	0.025 (0.011)	0.038 (0.024)	0.66	–0.012 (0.020)	–0.55	NS
PPY–OW	0.113 (0.030)	0.071 (0.046)	1.59	0.042 (0.050)	0.84	NS
HKL–OW	0.105 (0.029)	0.072 (0.044)	1.46	0.033 (0.051)	0.65	NS
HSA–HSA	0.190 (0.032)	0.046 (0.030)	4.75	0.150 (0.041)	3.66	<0.001
PTR–PTR	0.181 (0.039)	0.046 (0.022)	3.93	0.135 (0.046)	2.93	<0.01
GGO–GGO	0.512 (0.031)	0.067 (0.030)	2.27	0.085 (0.037)	2.30	<0.01
PPY–PPY	0.033 (0.013)	0.037 (0.027)	0.89	–0.004 (0.023)	–0.17	NS
HKL–HKL	0.021 (0.012)	0.044 (0.028)	0.48	–0.026 (0.024)	–1.08	0.95
OW–OW	0.022 (0.012)	0 (0)	NA	0.022 (0.011)	2.00	<0.05

HSA, *Homo sapiens* ($n = 15$ paralogous sequences); PTR, *Pan troglodytes* ($n = 21$); GGO, *Gorilla gorilla* ($n = 21$); HKL, *Hylobates klossi* ($n = 8$); PPY, *Pongo pygmaeus* ($n = 9$); OW, Old World monkeys (*Cercopithecus aethiops*, *Papio anubis*, *Papio hamadryas* and *Macaca fascicularis*; each species represented by a single sequence). NS, no significant positive selection detected.

Table 3 Positive selection of exon 4

Exon 4	$\bar{K}_a$ (s.e.)	$\bar{K}_s$ (s.e.)	$\bar{K}_a/\bar{K}_s$	$\bar{K}_a/\bar{K}_s$ (s.e.)	*Z*-value	*P*
HSA–PTR	0.099 (0.018)	0.056 (0.021)	1.77	0.043 (0.026)	1.65	<0.05
HSA–PPY	0.229 (0.043)	0.149 (0.053)	1.54	0.080 (0.066)	1.21	NS
HSA–HKL	0.233 (0.046)	0.143 (0.049)	1.63	0.090 (0.063)	1.36	NS
HSA–OW	0.275 (0.048)	0.177 (0.064)	1.55	0.098 (0.078)	1.26	NS
PTR–PPY	0.260 (0.048)	0.150 (0.054)	1.73	0.110 (0.068)	1.62	NS
PTR–HKL	0.261 (0.051)	0.143 (0.050)	1.83	0.118 (0.066)	1.79	<0.05
PTR–OW	0.292 (0.051)	0.180 (0.066)	1.62	0.112 (0.080)	1.40	NS
PPY–HKL	0.057 (0.020)	0.037 (0.021)	1.54	0.021 (0.024)	0.88	NS
PPY–OW	0.103 (0.025)	0.048 (0.023)	2.15	0.056 (0.033)	1.70	<0.05
HKL–OW	0.107 (0.026)	0.049 (0.025)	2.18	0.057 (0.032)	1.78	<0.05
HSA–HSA	0.078 (0.018)	0.066 (0.027)	1.18	0.012 (0.035)	0.34	NS
PTR–PTR	0.089 (0.015)	0.045 (0.016)	1.98	0.044 (0.019)	2.32	<0.01
PPY–PPY	0.060 (0.019)	0.039 (0.025)	1.54	0.021 (0.028)	0.75	NS
HKL–HKL	0.066 (0.021)	0.040 (0.025)	1.65	0.026 (0.028)	0.93	NS
OW–OW	0.093 (0.020)	0.055 (0.028)	1.69	0.039 (0.032)	1.22	NS

HSA, *Homo sapiens* ($n = 15$ paralogous sequences); PTR, *Pan troglodytes* ($n = 19$); HKL, *Hylobates klossi* ($n = 7$); PPY, *Pongo pygmaeus* ($n = 14$); OW, Old World monkeys (*Cercopithecus aethiops*, *Papio anubis*, *Papio hamadryas*, *Presbytis Cristata* and *Allenopithecus*, each species represented by a single copy sequence). NS, no significant positive selection detected.

porin (p62)[16] further indicates that this particular human copy may associate with the nuclear pore complex (Supplementary Information Fig. 5).

Our analysis has revealed an extraordinary degree of evolutionary plasticity, at the level of both the genome and the gene. We provide evidence for the evolution of a hominoid gene family by recent duplication and positive selection. Can additional examples within the human proteome be expected? Preliminary analysis of the human genome suggests that as much as 5–7% of all human sequences may have been duplicated within the last 30 Myr of evolution. The abundance of segmental duplications may be an important reservoir for the emergence of other hominoid genes that do not possess definitive orthologues in the genomes of model organisms. □

Methods

Human genome analysis

Searches for sequence similarity in GenBank (release 121.0) identified 41 human accessions that contained a complete copy of the LCR16a repeat. Because the degree of sequence similarity among these copies approached levels of allelic variation (98–99%), comparison of unique sequences, flanking the duplications, and partial sequencing of chromosome 16 cosmids (LA16NC02) were used to distinguish various paralogues from allelic overlap. The human cosmid library is derived from a single chromosome 16 haplotype, thereby allowing sequence variants to effectively classify duplicated copies[17]. A suite of genomic software tools were used to analyse and characterize the duplications, including PARASIGHT[5] to delineate the junction sequences and the extent of overlap for each duplicated segment, ALIGN to perform optimal global pairwise alignments between copies, and sim4 to optimally compare cDNA with genomic DNA[17]. Only pairwise sequence alignments greater than 1 kb with at least 90% identity were considered in this analysis (see Supplementary Information). Unique sequence differences within the predicted exons from genomic sequences compared with expressed sequence tag sequences were used to identify transcriptionally competent loci. Software for analysis of protein structure (http://bmerc-www.bu.edu/psa and http://maple.bioc.columbia.edu/predictprotein) predicted the presence of single transmembrane domain flanked by α-helical secondary structure for the AF132984 open reading frame (347 amino acids).

Fluorescence *in situ* hybridization

Chromosome metaphase and interphase nuclei were prepared from lymphoblastoid cell lines representative of five hominoid species (*Homo sapiens*, *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla* and *Pongo pygmaeus)*, three Old World monkey (*Pan anubis*, *Presbytis cristata* and *Cercopithecus aethiops*) and one New World monkey (*Callicebus mollochus*). *In situ* hybridizations were performed under standard conditions[18] with two genomic probes (16.1/9 and 16.8/12; Fig. 1c) subcloned from one of the paralogous copies (AC002039). To eliminate the effect of cross-hybridization of common repeat sequences, probes were blocked by using repetitive DNA (C_{ot}) before hybridization. At least 20 independent metaphase and interphase nuclei were examined in the determination of copy number and chromosomal band location. The combined probes spanned 11.5 kb of genomic sequence and included 7 exons of the AF132984 cDNA. Reciprocal experiments using probes derived from baboons and gibbons were used to confirm the specificity of hybridization. When necessary, hybridizations were performed in conjunction with human whole-chromosome painting probes to confirm chromosomal assignment (orangutan, Fig. 2).

Library hybridization and sequencing

Large-insert genomic libraries from human (LA16NC02), chimpanzee (RPCI-43, *P. troglodytes*), gibbon (DKZ-140, *Hylobates klossi*) and the olive baboon (RPCI-41, *Papio hamadryas*) were hybridized with PCR-amplified products (16.1/9 and 16.8/12; see Supplementary Information Table 3 for conditions and oligonucleotide sequences). All hybridizations were performed as previously described[17]. A total of 156 genomic clones (70 human, 75 chimpanzee, 10 gibbon and 1 baboon) were comparatively sequenced. (In all, 1,753 bp were examined, partitioned into 1,421 bp of intronic sequence and 332 bp of sequence from exons 2 and 4.) All PCR products (forward and reverse reactions) were directly sequenced using a modified dye-terminator sequencing protocol[17]. Non-human sequences were deemed to be paralogous if more than two sequence differences were observed within 150 bp of coding sequence. All paralogues were encoded by species name and numbered according to clone and/or accession identifier (Supplementary Information Table 4). End sequences generated from the cloning site (T7 and T3 or T3 and SP6) were used to further position specific paralogous copies with respect to the human genome reference. In all cases, the duplicated sequence was flanked either directly or within ~70 kb by non-duplicated unique sequence. As a result, end-sequence analysis allows a subset of bacterial artificial chromosome (BAC) clones from different species to be unambiguously placed on the basis of alignment to unique sequence on either side spanning the duplication. For exons 2 and 4, additional sequences were generated by TA-subcloning of PCR-amplified product from orangutans (*P. pygmaeus)* and direct PCR sequencing of products from various Old World monkeys (*Macaca fascicularis*, *Presbytis cristata* and *Cercopithecus aethiops*).

Sequence analysis

Estimates of genetic distance (pairwise deletion) were calculated using the Jukes and Cantor one-parameter model (when transition/transversion quotients $i/v \approx 1.0$) or Kimura's two-parameter model (when $i/v \approx 2.0$)[19]. A Tajima's relative rate test was performed[20] using orthologous sequence pairs (HSA13 versus PTR3, PTR8 versus HKL1, HKL1 versus HSA13, and HSA3 versus PTR17; see Supplementary Information Table 4) from human, chimpanzee and gibbon intronic sequences with the baboon sequence (PHA) as the outgroup. Four such tests were used to accept the molecular-clock hypothesis for the non-coding sequences under study in this analysis. Estimates of duplication timing were based on 1,421 bp of non-coding sequence and were calculated using the formula $r = K/2T$, (where r is the rate of nucleotide substitution; K is the number of substitutions for site; and T is the time of separation), with the baboon sequence as a reference orthologue and an estimated time of separation from the hominoid lineage of 25 Myr[6]. For exonic sequence, the average number of synonymous (K_s) and non-synonymous (K_a) substitutions per site were estimated using the modified Nei–Gojobori method[12,21]. To test for positive darwinian selection, we calculated the difference ($D = K_a - K_s$) within and between groups for all pairwise comparisons of paralogues. Groups are here defined as species. Owing to the large number of pairwise analyses performed, significance levels should be corrected for multiple comparisons. Because the comparisons are not independent of one another, the usual Bonferroni method cannot be used. Instead, the average difference for all comparisons and its associated standard error were computed. Initially, all possible comparisons were made among the sequenced exons and the difference between amino-acid and synonymous substitutions was calculated for each pairwise comparison. The differences were averaged between and within groups (within groups included multiple duplicate copies within each species). The variance for the average difference was estimated using the bootstrap method (n = 1,000 replicates) and a one-tailed Z-test ($Z = D/\sigma$) to determine the level of significance[22]. Positive selection was defined as a significant positive difference. Evolutionary trees of multiple aligned sequences (ClustalW) were generated using neighbour-joining distance estimates (MEGA2). Only bootstrap values >50% are indicated in the tree topology. Internal branch estimates of the number of synonymous (b_s) and non-synonymous (b_a) substitutions per site were determined by the method of Zhang *et al.*[12].

Received 13 March; accepted 29 June 2001.

1. Ohno, S. *Evolution by Gene Duplication* (Springer, Berlin, 1970).
2. Stallings, R., Whitmore, S., Doggett, N. & Callen, D. Refined physical mapping of chromosome 16-specific low-abundance repetitive DNA sequences. *Cytogenet. Cell Genet.* **63,** 97–101 (1993).
3. Loftus, B. *et al.* Genome duplications and other features in 12 Mbp of DNA sequence from human chromosome 16p and 16q. *Genomics* **60,** 295–308 (1999).
4. The International Human Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–920 (2001).
5. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11,** 1005–1017 (2001).
6. Goodman, M. The genomic record of humankind's evolutionary roots. *Am. J. Hum. Genet.* **64,** 31–39 (1999).
7. Li, W. *Molecular Evolution* (Sinauer, Sunderland, 1997).
8. Wyckoff, G. J., Wang, W. & Wu, C. I. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403,** 304–309 (2000).
9. Nurminsky, D. I., Nurminskaya, M. V., De Aguiar, D. & Hartl, D. L. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. *Nature* **396,** 572–575 (1998).
10. Duda, T. F. & Palumbi, S. R. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *Proc. Natl Acad. Sci. USA* **96,** 6820–6823 (1999).
11. Vacquier, V. D., Swanson, W. J. & Lee, Y. H. Positive Darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence? *J. Mol. Evol.* **44,** S15–S22 (1997).
12. Zhang, J., Rosenberg, H. F. & Nei, M. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl Acad. Sci. USA* **95,** 3708–3713 (1998).
13. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335,** 167–170 (1988).
14. Messier, W. & Stewart, C. B. Episodic adaptive evolution of primate lysozymes. *Nature* **385,** 151–154 (1997).
15. Ting, C. T., Tsaur, S. C., Wu, M. L. & Wu, C. I. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* **282,** 1501–1504 (1998).
16. Davis, L. I. & Blobel, G. Identification and characterization of a nuclear pore complex protein. *Cell* **45,** 699–709 (1986).
17. Horvath, J., Schwartz, S. & Eichler, E. The mosaic structure of a 2p11 pericentromeric segment: A strategy for characterizing complex regions of the human genome. *Genome Res.* **10,** 839–852 (2000).
18. Lichter, P. *et al.* High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science* **247,** 64–69 (1990).
19. Jukes, T. H. & Cantor, C. R. in *Mammalian Protein Metabolism* (ed. Munro, H. N.) 21–123 (Academic, New York, 1969).
20. Tajima, F. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135,** 599–607 (1993).
21. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3,** 418–426 (1986).
22. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, New York, 2000).

Supplementary information is available on *Nature's* World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature.*

Acknowledgements

We thank W. E. Kutz and D. Zivkovic for technical assistance and sequencing analyses. This work was supported by grants from the National Institutes of Health and the US Department of Energy to E.E.E., and grants from Progretti di Interesse Nationale (PRIN), Centro Eccelenza (CE), Ministero per la Ricerca Scientifica e Tecnologica (MURST) and Telethon to M.R. We are grateful to C. I. Wu, A. Chakravarti, D. Cutler, D. Locke, G. Matera and H. Willard for comments on this manuscript.

Correspondence and requests for materials should be addressed to E.E.E. (e-mail: eee@po.cwru.edu). All sequences have been deposited in GenBank under accession numbers AF364182–AF364299.

A forkhead-domain gene is mutated in a severe speech and language disorder

Cecilia S. L. Lai*†, Simon E. Fisher*†, Jane A. Hurst‡, Faraneh Vargha-Khadem§ & Anthony P. Monaco*

** Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK*
‡ Department of Clinical Genetics, Oxford Radcliffe Hospital, Oxford OX3 7LJ, UK
§ Developmental Cognitive Neuroscience Unit, Institute of Child Health, Mecklenburgh Square, London WC1N 2AP, UK
† These authors contributed equally to this work

Individuals affected with developmental disorders of speech and language have substantial difficulty acquiring expressive and/or receptive language in the absence of any profound sensory or neurological impairment and despite adequate intelligence and opportunity[1]. Although studies of twins consistently indicate that a significant genetic component is involved[1–3], most families segregating speech and language deficits show complex patterns of inheritance, and a gene that predisposes individuals to such disorders has not been identified. We have studied a unique three-generation pedigree, KE, in which a severe speech and language disorder is transmitted as an autosomal-dominant monogenic trait[4]. Our previous work mapped the locus responsible, SPCH1, to a 5.6-cM interval of region 7q31 on chromosome 7 (ref. 5). We also identified an unrelated individual, CS, in whom speech and language impairment is associated with a chromosomal translocation involving the SPCH1 interval[6]. Here we show that the gene *FOXP2*, which encodes a putative transcription factor containing a polyglutamine tract and a forkhead DNA-binding domain, is directly disrupted by the translocation breakpoint in CS. In addition, we identify a point mutation in affected members of the KE family that alters an invariant amino-acid residue in the forkhead domain. Our findings suggest that *FOXP2* is involved in the developmental process that culminates in speech and language.

Investigations of the KE family (Fig. 1) have been central to discussions regarding the innate aspects of language ability[4,5,7–9]. Affected members have a severe impairment in the selection and sequencing of fine orofacial movements, which are necessary for articulation (referred to as a developmental verbal dyspraxia; MIM 602081)[4,8,9]. The disorder is also characterized by deficits in several facets of language processing (such as the ability to break up words into their constituent phonemes) and grammatical skills (including production and comprehension of word inflections and syntactical structure)[7,8].

Although the mean non-verbal IQ of affected members is lower than that of unaffected members[8], there are affected individuals in the family who have non-verbal ability close to the population average, despite having severe speech and language difficulties; therefore, non-verbal deficits cannot be considered as characteristic of the disorder. Functional and structural brain-imaging studies of affected members of the KE family have suggested that the basal ganglia may be a site of bilateral pathology associated with the trait[9]. Although there has been some debate over which feature of the phenotype constitutes the core deficit in this disorder, all the different studies agree that the gene disrupted in the KE family is likely to be important in neural mechanisms mediating the development of speech and language.

After our initial localization of SPCH1 to 7q31 (ref. 5), we used a bioinformatic approach to construct a transcript map of the crucial interval containing nearly 8 megabases of completed genomic sequence[6]. In addition, we reported molecular cytogenetic studies of an unrelated patient CS, who has a speech and language disorder that is strikingly similar to that of the KE family, associated with a *de novo* balanced reciprocal translocation t(5;7)(q22;q31.2)[6]. As observed for affected members of the KE family, CS presents with a severe orofacial dyspraxia despite normal early feeding and gross motor development. For both KE and CS phenotypes, there is substantial impairment of expressive and receptive language abilities. In both cases, general intelligence is relatively spared: although there is some lowering of IQ, deficits are more profound in the verbal domain.

Fluorescence *in-situ* hybridization (FISH) with a series of bacterial artificial chromosome (BAC) clones enabled us to map the 7q31.2 breakpoint of CS to a single clone, named NH0563O05, and did not reveal any additional associated genomic rearrangements in the vicinity of the translocation[6]. We discovered that the NH0563O05 clone contains several exons from CAGH44, a brain-expressed transcript encoding a large stretch of consecutive polyglutamines[6] (Fig. 2). A previous study of CAGH44 had determined only the first 869 base pairs (bp) of coding sequence from a partial transcript of the gene, in which no in-frame stop codon had been reached[10]. Investigation of this 5′ part of the open reading frame (ORF) in the KE family did not detect any sequence variant co-segregating with the speech and language disorder[6].

To isolate the complete coding region of this candidate gene, we obtained the genomic sequence of NH0563O05 and adjacent BAC clones. Computer-based investigation of these data, using database search tools and gene prediction programs, enabled us to assemble the sequence of a hypothetical 2.5-kilobase (kb) transcript comprising 17 exons and containing a complete ORF of about 2.1 kb (Fig. 2). We verified the predicted transcript sequence experimentally (see Methods), confirming the exon–intron structure of the gene and identifying alternative splicing of two additional exons at the 5′ end of the gene in all tissues examined (Fig. 2b). The carboxy-terminal portion of the predicted protein sequence encoded by this gene

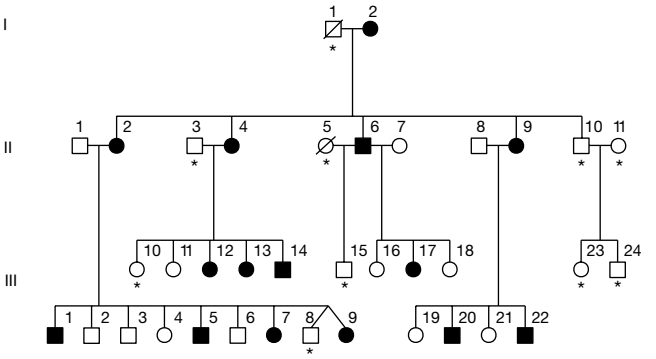

Figure 1 Pedigree of the KE family. Affected individuals are indicated by filled symbols. Asterisks indicate those individuals who were unavailable for genetic analyses. Squares are males, circles are females, and a line through a symbol indicates that the person is deceased.