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

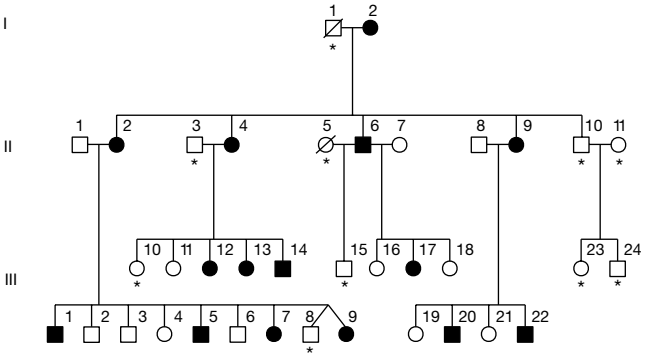


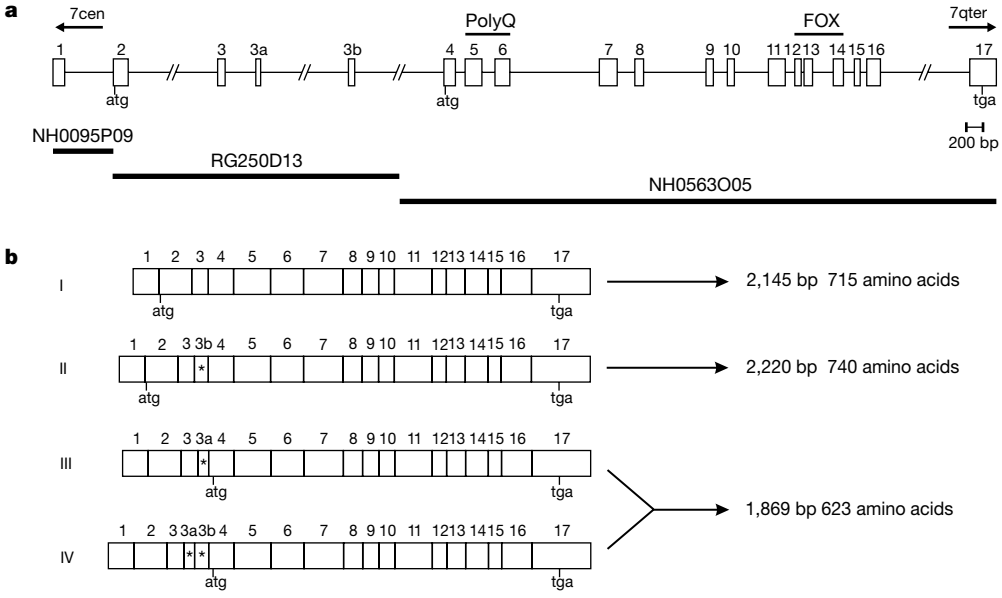
Figure 1 Pedigree of the KE family. Affected individuals are indicated by filled symbols. Asterisks indicate those individuals who were unavailable for genetic analyses. Squares are males, circles are females, and a line through a symbol indicates that the person is deceased.

contains a segment of 84 amino acids (encoded by exons 12–14) that shows high similarity to the characteristic DNA-binding domain of the forkhead/winged-helix (FOX) family of transcription factors[11–14] (Fig. 2c). The complete gene has been therefore designated FOXP2, in accordance with the standard nomenclature proposed for this rapidly growing gene family[14].

Northern blot analysis (see Supplementary Information) of several human adult tissues showed that there is broad expression of a roughly 6.5-kb transcript. This transcript was also observed in fetal tissues, with strong expression in brain. Similarly, an investigation of the murine homologue of FOXP2 has demonstrated expression in a range of adult and fetal mouse tissues[15]. Using *in situ* hybridization, it was also found that murine FOXP2 is expressed in defined regions of the central nervous system during mouse embryogenesis, including the neopallial cortex and the developing cerebral hemispheres[15].

We used additional FISH experiments and Southern blot analysis of DNA from CS to investigate further the relationship between the translocation and the FOXP2 locus. We thereby localized the translocation breakpoint to a 200-bp region in the intron between exons 3b and 4 (Fig. 3). These results indicate that disruption of FOXP2 is implicated in the aetiology of the speech and language disorder of this patient.

We screened the newly defined coding regions (exons 1, 3b and 8–17) of FOXP2 for mutations in the KE family. A G-to-A nucleotide transition was detected in exon 14 of affected individuals, and



c

```
FOXP2  MMQESATETISNSSMNQNGMSTLSSQLDAGS-RDGRSSGDTSS-EVSTVELLHLQQQQ--ALQAARQLLLQQQ----TSGLKSPKSSDKQRPLQVPVSVA 92
       MMQES TET SN S  QNG    +  L+ G  R+GRS+G+T + ++   +L H QQQQ  ALQ ARQLLLQQQ     SGLKSPK +DKQ  LQVPVSVA
FOXP1  MMQESGTETKSNGSAIQNGSGGSNHLLECGGLREGRSNGETPAVDIGAADLAHAQQQQQQALQVARQLLLQQQQQQQQVSGLKSPKRNDKQPALQVPVSVA 100

FOXP2  MMTPQVITPQQMQQILQQQVLSPQQLQALLQQQQAVMLQQQQLQEFYKKQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQH 192
       MMTPQVITPQQMQQILQQQVLSPQQLQ LLQQQQA+MLQQQQLQEFYKKQQEQL LQLLQQQ                                   H
FOXP1  MMTPQVITPQQMQQILQQQVLSPQQLQVLLQQQQALMLQQQQLQEFYKKQQEQLQLQLLQQQ----------------------------------H 163

FOXP2  PGKQAKEQQQQQQQQQQLAAQQLVFQQQLLQMQQLQQQQHLLSLQRQGLISIPPGQAALPVQSLPQAGLSPAEIQQLWKEVTGVHSMEDN-GIKHGGLDL 291
        GKQ KEQQQ        +A QQL FQQQLLQMQQLQQQ HLLSLQRQGL++I PGQ ALP+Q L Q G+ P E+QQLWKEVT  H+ E+  G  H  LDL
FOXP1  AGKQPKEQQQ-------VATQQLAFQQQLLQMQQLQQQ-HLLSLQRQGLLTIQPGQPALPLQPLAQ-GMIPTELQQLWKEVTSAHTAEETTGNNHSSLDL 254

FOXP2  TTNNSSSTTSSNTSKASPPITHHSIVNGQSSVLSARRDSSSHEETGASHTLYGHGVCKWPGCESICEDFGQFLKHLNNEHALDDRSTAQCRVQMQVVQQL 391
       TT   SS+  S TS    P   H+  NGQ SV + +R+S SHEE   SH LYGHGVCKWPGCE++CEDF  FLKHLN+EHALDDRSTAQCRVQMQVVQQL
FOXP1  TTTCVSSSAPSKTSLIMNP---HASTNGQLSVHTPKRESLSHEEHPHSHPLYGHGVCKWPGCEAVCEDFQSFLKHLNSEHALDDRSTAQCRVQMQVVQQL 351

FOXP2  EIQLSKERERLQAMMTHLHMRPSEPKPSPKPLNLVSSVTMSKNMLETSPQSLPQTPTTPTAPVTPITQGPSVITPASVPNVGAIRRRHSDKYNIPMSS-E 490
       E+QL+K++ERLQAMMTHLH++ +EPK +P+PLNLVSSVT+SK+  E SPQSLP TPTTPTAP+TP+TQGPSVIT  S+  VG IRRR+SDKYN+P+SS +
FOXP1  ELQLAKDKERLQAMMTHLHVKSTEPKAAPQPLNLVSSVTLSKSASEASPQSLPHTPTTPTAPLTPVTQGPSVITTTSMHTVGPIRRRYSDKYNVPISSAD 451

FOXP2  IAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFAYFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEYQKRRSQKITG 590
       IA N EFYKNA+VRPPFTYA+LIRQAI+ES ++QLTLNEIY+WFTR FAYFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVE+QKRR QKI+G
FOXP1  IAQNQEFYKNAEVRPPFTYASLIRQAILESPEKQLTLNEIYNWFTRMFAYFRRNAATWKNAVRHNLSLHKCFVRVENVKGAVWTVDEVEFQKRRPQKISG 551

FOXP2  SPTLVKNIPTSLGYGAALNASLQAALAESSLPLLSNPGLINNASSGLLQAVHEDLNGSLDHIDSN-GNSSPGCSPQPHIHSIHVKEEPVIAEDEDCPMSL 689
       +P+L+KN+ +S  Y   LNA+LQA++AE+S+PL +   + N     L  A+ E+LNG+++H +SN  +SSPG SP   +H +HVKEEP+  E+ + P+SL
FOXP1  NPSLIKNMQSSHAYCTPLNAALQASMAENSIPLYTTASMGNPTLGNLASAIREELNGAMEHTNSNESDSSPGRSPMQAVHPVHVKEEPLDPEEAEGPLSL 651

FOXP2  VTTANHSPELEDDREIEEEPLSEDLE 715
       VTTANHSP+ + DR+ E+EP++ED+E
FOXP1  VTTANHSPDFDHDRDYEDEPVNEDME 677
```

Figure 2 Identification of the human FOXP2 gene. **a**, Representation of human FOXP2 gene structure. Boxes represent exons, with positions of initiation and termination codons indicated. The scale shown applies only to exons; the entire region spans more than 267 kb of genomic DNA. Exons encoding polyglutamine tracts (PolyQ) and the forkhead domain (FOX) are indicated. The gene includes regions corresponding to expressed sequence tag ye50f03.r1 (exon 1), the partial CAGH44 transcript (exons 2–7) and a partial cDNA clone YX52E07 (exons 11–15). BAC genomic sequence entries are aligned beneath the gene structure. **b**, Alternative splicing of exons 3a and 3b (indicated by asterisks) leads to four different transcripts. 'I' was originally identified by genomic predictions. 'II' contains exon 3b, which inserts 75 bp in-frame into the coding region. 'III' and 'IV' include the 58-bp exon 3a, which shifts the frame such that the ORF begins in exon 4, rather than exon 2. **c**, Amino-acid sequence encoded by human FOXP2 (transcript 'I'), aligned with human FOXP1 (accession AAG47632). The 40-residue and 10-residue stretches of polyglutamine in FOXP2 are reduced to only three glutamines each in FOXP1. FOXP2 transcript 'II' inserts 25 amino acids between residues 86 and 87. Transcripts 'III' and 'IV' give a shorter product beginning with the methionine at residue 90.

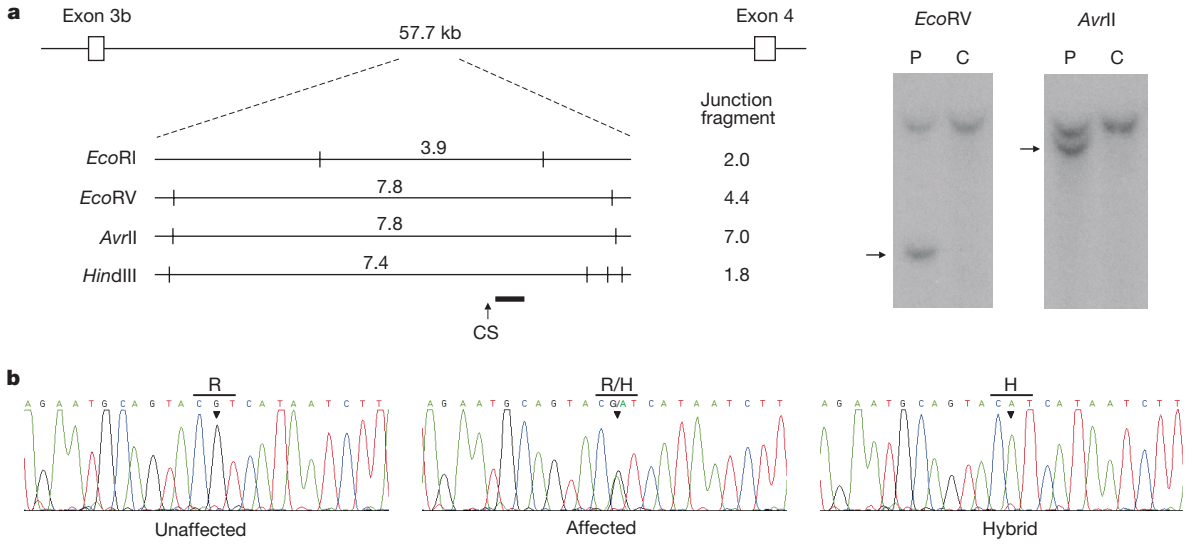


Figure 3 Disruption of FOXP2 in patients with severe speech and language disorder. **a**, Localization of the CS translocation breakpoint. A 499-bp probe (indicated by a thick black line) from the intron between exon 3b and exon 4 detected abnormal restriction fragments on Southern blots with four different enzymes, *Eco*RI, *Eco*RV, *Avr*II and *Hin*dIII. A scaled restriction map of the normal locus is shown, with estimated sizes (in kilobases) of detected junction fragments displayed at the side. The *Hin*dIII results indicate that the CS breakpoint maps to a region of ~200 bp on the centromeric side of the probe. Examples of Southern blot hybridizations with digested DNA from the patient (P) and a control (C) are also shown for two of the enzymes, with the junction fragments indicated by arrows. **b**, Direct sequencing of exons from FOXP2 detected a G-to-A transition causing an R553H substitution in the forkhead domain in family KE. All affected individuals from the KE pedigree were heterozygous for this mutation, whereas all unaffected individuals were homozygous for the wild type (see Fig. 1). Somatic cell hybrids containing only the chromosome 7 associated with the speech and language disorder[6] were hemizygous for the mutation.

shown to co-segregate perfectly with the speech and language disorder in the KE pedigree (Fig. 3). Using a restriction-enzyme-based assay, we showed that the mutation was absent in 364 independent chromosomes from normal Caucasian controls (data not shown), indicating that it does not represent a naturally occurring polymorphism. The mutation is predicted to result in an arginine-to-histidine substitution (R553H) in the forkhead DNA-binding domain of FOXP2 (Fig. 4). Forkhead (or winged-helix) domains adopt a characteristic structure, comprising three amphipathic α-helices followed by two large loops (called 'wings'), in which the third α-helix is presented to the major groove of the target DNA[12,16]. The R553H change occurs in this third helix, which is the most highly conserved part of the forkhead domain[12], adjacent to a histidine residue that makes a direct base contact with the target DNA[16].

The R553 amino acid is invariant in all the currently known members of the large family of forkhead proteins, in species ranging from yeast to human (see http://www.biology.pomona.edu/fox.html). Furthermore, it has been proposed as an invariant feature of all homeodomain recognition helices[12]. Therefore, we suggest that this arginine residue is crucially important for the function of the forkhead domain, and that the histidine substitution observed in affected members of the KE family disrupts the DNA-binding and/or transactivation properties of FOXP2. The alternative hypothesis—that the R553H change is in linkage disequilibrium with a pathogenic mutation in a neighbouring gene and that the disorder in the translocation patient actually results from positional inactivation of this other gene—is highly unlikely.

Many members of the forkhead family are known to be key regulators of embryogenesis[13]. Mutations in FOX genes have been implicated in specific human disorders, including congenital

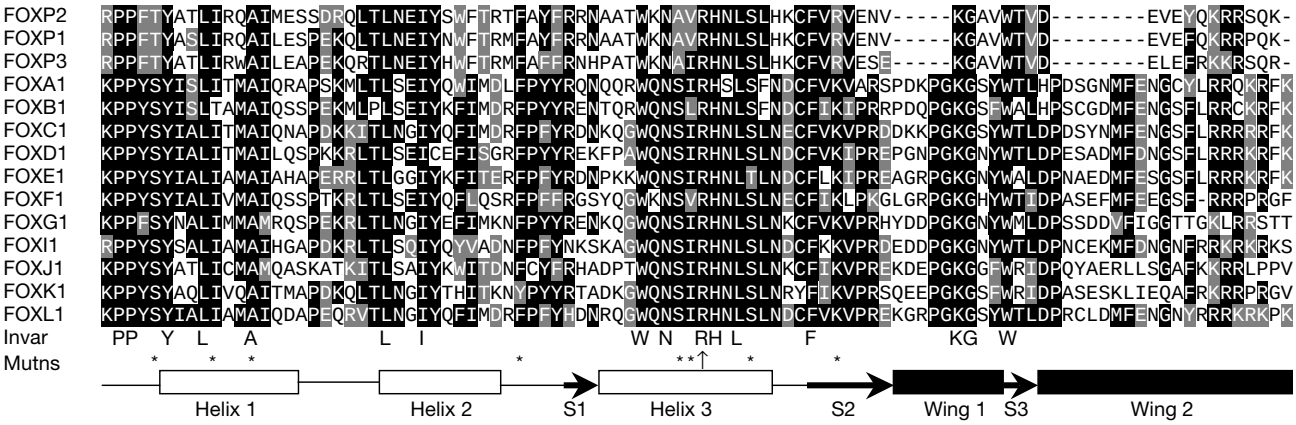


Figure 4 Forkhead domains of the three known FOXP proteins aligned with representative proteins from several branches of the FOX family. All sequences are from *Homo sapiens*. Residues that are invariant in this selection of forkhead proteins are given beneath the alignment. Asterisks show sites of the substitution mutations in FOXC1, FOXE1 and FOXF2 that have been previously implicated in human disease states[17–19,23,24]. The upwards arrow indicates the site of the R553H substitution identified in FOXP2 in affected members of the KE pedigree. The proposed structure of the forkhead domain as established by X-ray crystallography[16] is shown, containing three α-helices, three β-strands (S1–3) and two wings.

glaucoma (FOXC1)[17,18], thyroid agenesis (FOXE1)[19], lymphedema–distichiasis (LD) syndrome (FOXC2)[20], blepharophimosis/ptosis/epicanthus inversus (BPES) syndrome (FOXL2)[21], and anterior-segment dysgenesis associated with cataracts (FOXE3)[22]. The mouse phenotype *scurfy* and a similar syndrome found in humans both result from disruption of FOXP3 (refs 23–25), a gene that is closely related to FOXP2.

A significant number of the mutations identified in FOX genes are missense changes, and all of these result in substitution at residues in the forkhead domain[17–19,23,24], as observed here for FOXP2 (Fig. 4). Frameshift and nonsense mutations yielding truncated protein products that lack a forkhead domain have also been identified[17,18,20,23]. In addition, there have been reports of balanced translocations causing positional effect inactivation of FOXC1, FOXC2 and FOXL2 in glaucoma[17], lymphedema–distichiasis[20] and BPES[21], respectively. Data from those studies[17–20,23,24], as well as from mouse models[25–27] and *in vitro* functional assays[19], indicate that inactivation or loss of the forkhead domain is a general mechanism by which mutation of FOX genes can lead to human disease states. Investigations of forkhead-domain mutations associated with autosomal dominant traits suggests that the resulting disorders are a consequence of haplo-insufficiency during embryological development[17,18,20,27]. The finding that duplications involving FOXC1 can cause anterior-chamber defects of the eye[28,29] provides further evidence that the correct gene dosage of forkhead transcription factors is important in embryogenesis.

In addition to the forkhead domain, the FOXP2 protein also contains a stretch of 40 consecutive glutamines followed by a second stretch of only 10 glutamines. Abnormal expansion of variable polyglutamine tracts has been implicated in several hereditary neurodegenerative disorders[30]. The polyglutamine region of FOXP2 is encoded by a mixture of CAG and CAA codons, making it highly stable in normal individuals[10]. Although polyglutamine tracts have been found in many transcription-related proteins[30] this is the first report of such a domain in a FOX family member. The amino-acid sequence of FOXP2 shows remarkable similarity throughout its length to FOXP1, another member of the P branch of the forkhead family that has been identified in humans (68% identity; 80% similarity). However, an intriguing difference between these two human paralogues is that the polyglutamine tracts of FOXP2 are reduced markedly in FOXP1 (Fig. 2c); thus, comparison of the properties of the two proteins might shed light on the role of polyglutamine repeats in non-pathological processes.

In conclusion, we have shown that the FOXP2 gene is directly disrupted by a translocation in a patient with a speech and language disorder, and that a mutation affecting a crucial residue of the forkhead domain of this putative transcription factor co-segregates with affection status in the KE family. We propose that, in both cases, FOXP2 haplo-insufficiency in the brain at a key stage of embryogenesis leads to abnormal development of neural structures that are important for speech and language. This is the first gene, to our knowledge, to have been implicated in such pathways, and it promises to offer insights into the molecular processes mediating this uniquely human trait. □

Methods

Bioinformatic analyses

We obtained BAC genomic sequence data from the Washington University Genome Sequencing Center database (http://genome.wustl.edu/gsc). Genomic sequence data were analysed with database search tools and gene prediction software, as implemented in the NIX package (http://www.hgmp.mrc.ac.uk/NIX). Amino-acid sequences of FOXP2 and FOXP1 in Fig. 2c were aligned using BLAST2 (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html). Forkhead-domain sequences from human FOX proteins in Fig. 4 were aligned using ClustalW, accessed through the Baylor College of Medicine Search Launcher (http://searchlauncher.bcm.tmc.edu:9331/multi-align/multi-align.html).

FOXP2 mRNA sequence and genomic structure

We used a reverse-transcriptase polymerase chain reaction (RT–PCR)-based approach to confirm the FOXP2 mRNA sequence that had been predicted by bioinformatics. Primers were designed from putative exonic sequence and used to amplify by PCR first-strand complementary DNA from a range of adult tissues, which was obtained from Clontech. Products were sequenced as described[6] and compared with the predicted sequence.

Expression analyses of FOXP2

Adult and fetal northern blots were obtained from Clontech and hybridized according to the manufacturers' instructions, using a cDNA probe isolated from exons 8–11 of FOXP2.

Translocation mapping

We performed FISH on metaphase spreads of cells from CS, using a series of roughly 10-kb genomic probes obtained from the NH0563O05 BAC clone, as described[6]. In parallel, we ran Southern blot analyses of several restriction fragments spanning the FOXP2 locus, comparing digested DNA from CS with that from unaffected controls, according to standard procedures.

Mutation search

On the basis of genomic sequence information, we designed primers to flank each FOXP2 exon. These were used for PCR amplification of DNA from affected and unaffected individuals of the KE family, and from hybrid cell lines containing the affected chromosome 7 (ref. 6). We sequenced products as described[6]. The G-to-A transition detected in exon 14 of affected individuals destroys a restriction site for the enzyme *Mae*II (A|CGT). An assay using this restriction enzyme was developed to test for the exon 14 change in 182 unrelated normal controls.

GenBank accession numbers

BAC genomic sequence data, AC073626, AC003992 and AC020606; human FOXP2 mRNA sequence, AF337817.

Received 13 February; accepted 27 July 2001.

1. Bishop, D. V. M., North, T. & Donlan, C. Genetic basis for specific language impairment: evidence from a twin study. *Dev. Med. Child Neurol.* **37,** 56–71 (1995).
2. Tomblin, J. B. & Buckwalter, P. R. Heritability of poor language achievement among twins. *J. Speech Lang. Hear. Res.* **41,** 188–199 (1998).
3. Dale, P. S. *et al.* Genetic influence on language delay in two-year-old children. *Nature Neurosci.* **1,** 324–328 (1998).
4. Hurst, J. A., Baraitser, M., Auger, E., Graham, F. & Norell, S. An extended family with a dominantly inherited speech disorder. *Dev. Med. Child Neurol.* **32,** 347–355 (1990).
5. Fisher, S. E., Vargha-Khadem, F., Watkins, K. E., Monaco, A. P. & Pembrey, M. E. Localization of a gene implicated in a severe speech and language disorder. *Nature Genet.* **18,** 168–170 (1998).
6. Lai, C. S. L. *et al.* The SPCH1 region on human 7q31: genomic characterization of the critical interval and localization of translocations associated with speech and language disorder. *Am. J. Hum. Genet.* **67,** 357–368 (2000).
7. Gopnik, M. & Crago, M. B. Familial aggregation of a developmental language disorder. *Cognition* **39,** 1–50 (1991).
8. Vargha-Khadem, F., Watkins, K., Alcock, K., Fletcher, P. & Passingham, R. Praxic and nonverbal cognitive deficits in a large family with a genetically transmitted speech and language disorder. *Proc. Natl Acad. Sci. USA* **92,** 930–933 (1995).
9. Vargha-Khadem, F. *et al.* Neural basis of an inherited speech and language disorder. *Proc. Natl Acad. Sci. USA* **95,** 12695–12700 (1998).
10. Margolis, R. L. *et al.* cDNAs with long CAG trinucleotide repeats from human brain. *Hum. Genet.* **100,** 114–122 (1997).
11. Lai, E., Clark, K. L., Burley, S. K. & Darnell, J. E. Jr Hepatocyte nuclear factor 3/fork head or "winged helix" proteins: a family of transcription factors of diverse biologic function. *Proc. Natl Acad. Sci. USA* **90,** 10421–10423 (1993).
12. Li, C. & Tucker, P. W. DNA-binding properties and secondary structural model of the hepatocyte nuclear factor 3/fork head domain. *Proc. Natl Acad. Sci. USA* **90,** 11583–11587 (1993).
13. Kaufmann, E. & Knöchel, W. Five years on the wings of fork head. *Mech. Dev.* **57,** 3–20 (1996).
14. Kaestner, K. H., Knöchel, W. & Martinez, D. E. Unified nomenclature for the winged helix/forkhead transcription factors. *Genes Dev.* **14,** 142–146 (2000).
15. Shu, W., Yang, H., Zhang, L., Lu, M. M. & Morrisey, E. E. Characterization of a new subfamily of winged-helix/forkhead (fox) genes that are expressed in the lung and act as transcriptional repressors. *J. Biol. Chem.* **276,** 27488–27497 (2001).
16. Clark, K. L., Halay, E. D., Lai, E. & Burley, S. K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364,** 412–420 (1993).
17. Nishimura, D. Y. *et al.* The forkhead transcription factor gene FKHL7 is responsible for glaucoma phenotypes which map to 6p25. *Nature Genet.* **19,** 140–147 (1998).
18. Mears, A. J. *et al.* Mutations of the forkhead/winged-helix gene, FKHL7, in patients with Axenfeld–Rieger anomaly. *Am. J. Hum. Genet.* **63,** 1316–1328 (1998).
19. Clifton-Bligh, R. J. *et al.* Mutation of the gene encoding human TTF-2 associated with thyroid agenesis, cleft palate and choanal atresia. *Nature Genet.* **19,** 399–401 (1998).
20. Fang, J. *et al.* Mutations in FOXC2 (MFH-1), a forkhead family transcription factor, are responsible for the hereditary lymphedema–distichiasis syndrome. *Am. J. Hum. Genet.* **67,** 1382–1388 (2000).
21. Crisponi, L. *et al.* The putative forkhead transcription factor FOXL2 is mutated in blepharophimosis/ptosis/epicanthus inversus syndrome. *Nature Genet.* **27,** 159–166 (2001).
22. Semina, E. V., Brownell, I., Mintz-Hittner, H. A., Murray, J. C. & Jamrich, M. Mutations in the human forkhead transcription factor FOXE3 associated with anterior segment ocular dysgenesis and cataracts. *Hum. Mol. Genet.* **10,** 231–236 (2001).
23. Wildin, R. S. *et al.* X-linked neonatal diabetes mellitus, enteropathy and endocrinopathy syndrome is the human equivalent of mouse scurfy. *Nature Genet.* **27,** 18–20 (2001).

24. Bennett, C. L. *et al.* The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3. *Nature Genet.* **27,** 20–21 (2001).
25. Brunkow, M. E. *et al.* Disruption of a new forkhead/winged-helix protein, scurfin, results in the fatal lymphoproliferative disorder of the scurfy mouse. *Nature Genet.* **27,** 68–73 (2001).
26. De Felice, M. *et al.* A mouse model for hereditary thyroid dysgenesis and cleft palate. *Nature Genet.* **19,** 395–398 (1998).
27. Smith, R. S. *et al.* Haploinsufficiency of the transcription factors FOXC1 and FOXC2 results in aberrant ocular development. *Hum. Mol. Genet.* **9,** 1021–1032 (2000).
28. Lehmann, O. J. *et al.* Chromosomal duplication involving the forkhead transcription factor gene FOXC1 causes iris hypoplasia and glaucoma. *Am. J. Hum. Genet.* **67,** 1129–1135 (2000).
29. Nishimura, D. Y. *et al.* A spectrum of FOXC1 mutations suggests gene dosage as a mechanism for developmental defects of the anterior chamber of the eye. *Am. J. Hum. Genet.* **68,** 364–372 (2001).
30. Cummings, C. J. & Zoghbi, H. Y. Fourteen and counting: unraveling trinucleotide repeat diseases. *Hum. Mol. Genet.* **9,** 909–916 (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 are deeply indebted to the KE family whose continued cooperation has made this research possible. We also thank CS and family for agreeing to participate in this study. We thank D. C. Jamison and E. D. Green for facilitating completion of the 7q31 genomic sequence; M. Fox, S. Jeremiah and S. Povey for the chromosome 7 hybrids; E. R. Levy for assistance with cytogenetic analyses; D. I. Stuart, E. Y. Jones and R. M. Esnouf for advice on structural analyses of forkhead domains; L. Rampoldi for assistance with northern blots; and E. Dunne for help with sequence analyses of other 7q31 candidate genes. Chromosome 7 sequence data were generated by the Washington University Genome Sequencing Center. This study was funded by the Wellcome Trust. A.P.M. is a Wellcome Trust Principal Research Fellow.

Correspondence and requests for materials should be addressed to A.P.M. (e-mail: anthony@well.ox.ac.uk).

Genome sequence of *Yersinia pestis*, the causative agent of plague

J. Parkhill*, B. W. Wren†, N. R. Thomson*, R. W. Titball‡, M. T. G. Holden*, M. B. Prentice§, M. Sebaihia*, K. D. James*, C. Churcher*, K. L. Mungall*, S. Baker*, D. Basham*, S. D. Bentley*, K. Brooks*, A. M. Cerdeño-Tárraga*, T. Chillingworth*, A. Cronin*, R. M. Davies*, P. Davis*, G. Dougan‖, T. Feltwell*, N. Hamlin*, S. Holroyd*, K. Jagels*, A. V. Karlyshev†, S. Leather*, S. Moule*, P. C. F. Oyston‡, M. Quail*, K. Rutherford*, M. Simmonds*, J. Skelton*, K. Stevens*, S. Whitehead* & B. G. Barrell*

* *The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK*
† *Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK*
‡ *Chemical and Biological Sciences, Dstl, Porton Down, Salisbury, Wiltshire SP4 0JQ, UK*
§ *Department of Medical Microbiology, St Bartholomew's and the Royal London School of Medicine and Dentistry, London EC1A 7BE, UK*
‖ *Centre for Molecular Microbiology and Infection, Department of Biological Sciences, Imperial College of Science, Technology and Medicine, London SW7 2AZ, UK*

The Gram-negative bacterium *Yersinia pestis* is the causative agent of the systemic invasive infectious disease classically referred to as plague[1], and has been responsible for three human pandemics: the Justinian plague (sixth to eighth centuries), the Black Death (fourteenth to nineteenth centuries) and modern plague (nineteenth century to the present day). The recent identification of strains resistant to multiple drugs[2] and the potential use of *Y. pestis* as an agent of biological warfare mean that plague still poses a threat to human health. Here we report the complete genome sequence of *Y. pestis* strain CO92, consisting of a 4.65-megabase (Mb) chromosome and three plasmids of 96.2 kilobases (kb), 70.3 kb and 9.6 kb. The genome is unusually rich in insertion sequences and displays anomalies in GC base-composition bias, indicating frequent intragenomic recombination. Many genes seem to have been acquired from other bacteria and viruses (including adhesins, secretion systems and insecticidal toxins). The genome contains around 150 pseudogenes, many of which are remnants of a redundant enteropathogenic lifestyle. The evidence of ongoing genome fluidity, expansion and decay suggests *Y. pestis* is a pathogen that has undergone large-scale genetic flux and provides a unique insight into the ways in which new and highly virulent pathogens evolve.

Yersinia pestis is primarily a rodent pathogen, usually transmitted subcutaneously to humans by the bite of an infected flea, but also transmitted by air, especially during pandemics of disease. Notably, *Y. pestis* is very closely related to the gastrointestinal pathogen *Yersinia pseudotuberculosis*, and it has been proposed that *Y. pestis* is a clone that evolved from *Y. pseudotuberculosis* (probably serotype O:1b (ref. 3)) 1,500–20,000 years ago[4]. Thus *Y. pestis* seems to have rapidly adapted from being a mammalian enteropathogen widely found in the environment, to a blood-borne pathogen of mammals that is also able to parasitize insects and has limited capability for survival outside these hosts. Horizontally acquired DNA may be significant in having enabled *Y. pestis* to adapt to new hosts; conversely, the identification of gene remnants produced through genome decay may be associated with a redundant enteric lifestyle. Given the historical importance of plague and the need to understand the evolution and pathogenesis of such a potentially devastating pathogen, we undertook the genome sequencing of *Y. pestis* CO92 (biovar Orientalis), a strain recently isolated from a fatal human case of primary pneumonic plague contracted from an infected cat[5].

The general features of the genome are shown in Fig. 1 and Table 1. The most striking large-scale features in the genome are anomalies in GC bias. All bacterial genomes sequenced to date have a small but detectable bias towards G on the leading strand of the bidirectional replication fork[6]. Anomalies in this plot can be caused by the very recent acquisition of DNA (such as prophages) or by the inversion or translocation of blocks of DNA. The three anomalies visible in the *Y. pestis* plot (see Supplementary Information; see also http://www.sanger.ac.uk/Projects/Y_pestis/) are each bounded by insertion sequence elements, suggesting that they could be the result of recent recombination between these perfect repeats. To investigate this, we designed polymerase chain reaction (PCR) primers to test for the presence and absence of the predicted translocation, and for the orientation of the two inversions (see Supplementary Information). PCR confirmed the position of the translocation, but, intriguingly, the results for the two inversions showed that both orientations were present in the same DNA preparation, with the inverse orientation predominating. This suggests genomic rearrangement during growth of the organism. The results were similar in DNA from three different subcultures of CO92 and investigation of other strains indicated that similar rearrangements may have occurred (see Supplementary Information). These results demonstrate that the *Y. pestis* genome is fluid, and capable of frequent intragenomic recombination *in vitro*; the rapid emergence of new ribotypes of *Y. pestis* biovar Orientalis in the environment following pandemic spread[7] shows that chromosomal rearrangements are common *in vivo*. The effects of these rearrangements on the biology and pathogenicity of the organism are unknown.

Gene acquisition has been important in the evolution of *Y. pestis*. In addition to the 70-kb virulence plasmid (pYV/pCD1) found in all pathogenic *Yersinia*, *Y. pestis* has acquired two unique plasmids that encode a variety of virulence determinants. A 9.5-kb plasmid (pPst/pPCP1) encodes the plasminogen activator Pla (ref. 8), a putative invasin that is essential for virulence by the subcutaneous route. A 100–110-kb plasmid (pFra/pMT1) encodes murine toxin and the F1 capsular protein, which have been shown to have a role in the transmission of plague. No conjugation apparatus is