Hum Genet (2002) 111 : 136–144
DOI 10.1007/s00439-002-0768-5

ORIGINAL INVESTIGATION

**Heather A. Bruce · Russell L. Margolis**

# *FOXP2*: novel exons, splice variants, and CAG repeat length stability

Received: 11 January 2002 / Accepted: 23 May 2002 / Published online: 16 July 2002


**Abstract** FOXP2 is a transcription factor containing a polyglutamine tract, a zinc-finger motif, and a forkhead DNA-binding domain. The *FOXP2* gene is located on 7q31. A missense mutation in the forkhead domain (exon 14) and a balanced reciprocal translocation t(5;7)(q22;q31.2) with a breakpoint between exons 3b and 4 have recently been associated with a speech and language disorder (SPCH1). The role of FOXP2 in this neurodevelopmental disorder suggests that mutations in *FOXP2* could cause other neuropsychiatric disorders. To begin investigation of this possibility, we examined the genomic structure and CAG/CAA repeat region of *FOXP2*. We detected little polymorphism and no expansions in the *FOXP2* CAG/CAA repeat in 142 individuals with progressive movement disorders. We found evidence of alternate splice variants and six previously undetected exons: three 5′ untranslated exons (s1, s2, s3), two additional untranslated exons (2a and 2b) between exons 2 and 3, a translated exon (4a) between exons 4 and 5, and a longer version of exon 10 (10+) that contains an alternate stop codon and produces a truncated protein (FOXP2-S). Our results suggest that *FOXP2* spans at least 603 kb of genomic DNA, more than twice the previously defined region, and provide evidence of a promoter region flanking exon s1. This demonstration of additional *FOXP2* exons and splice variants should facilitate understanding of FOXP2 function and the search for additional *FOXP2* mutations.

Electronic database information: accession numbers and URLs for the data in this article are as follows:
Nucleotide sequence data reported are available in the DDBJ/EBML/GenBank databases under accession numbers AF454830 and AF467252–AF467259
National Center for Biotechnology Information server, www.ncbi.nlm.nih.gov/blast/blast.cgi

H.A. Bruce · R.L. Margolis
Laboratory of Genetic Neurobiology, Division of Neurobiology, Department of Psychiatry,
Johns Hopkins University School of Medicine,
Baltimore, MD 21287, USA

R.L. Margolis (✉)
Program in Cellular and Molecular Medicine,
Johns Hopkins University School of Medicine,
Baltimore, MD 21287, USA

*Present address:*
R. L. Margolis
Meyer 2-181, 600 N. Wolfe Street, Baltimore, MD 21287, USA,
e-mail: rmargoli@jhmi.edu,
Tel.: +1-410-6144262, Fax: +1-410-9558233

## Introduction

FOXP2 belongs to a family of transcription factors (Kaestner et al. 2000) containing the DNA-binding forkhead/winged helix domain. FOXP2 also contains a zinc-finger motif (Shu et al. 2001) and one of the longest polyglutamine stretches in the human genome. The initial report of *FOXP2* described 19 exons, two of which (3a and 3b) are variably spliced into transcripts, covering approximately 300 kb on chromosome 7q31 (Lai et al. 2001).

We first detected a 5′ fragment of *FOXP2* in a cDNA library screen designed to identify genes containing CAG repeats encoding polyglutamine tracts, as part of our effort to test the hypothesis that such genes may be involved in neuropsychiatric disorders (Margolis et al. 1997, 1999). The polyglutamine tract, near the N terminal of the FOXP2 protein, consists of 40 consecutive glutamines encoded by $(CAG)_4CAA(CAG)_4(CAA)_2(CAG)_2(CAA)_2(CAG)_3(CAA)_5(CAG)_2(CAA)_2(CAG)_5CAA(CAG)_5CAA$ CAG followed eight residues later by ten consecutive glutamines encoded by $(CAG)_7CAACAGCAA$. A glutamine-rich sequence flanks both sides of the repeat domain. The possibility that this repeat may undergo expansion mutation is supported by the recent report that expansion of a repeat of similar length and structure in the TATA-binding protein (TBP) gene causes the neurodegenerative disorder spinocerebellar ataxia type 17 (Nakamura et al. 2001; Koide et al. 1999).

FOXP2 was recently associated with a speech and language disorder in the large KE family and an unrelated individual (CS; Lai et al. 2001). The phenotype is complex, and several different aspects have been described (Vargha-

Khadem et al. 1995, 1998; Hurst et al. 1990; Gopnik and Crago 1991). The central deficit appears to be in the coordination of orofacial movements; the lower face and mouth of affected individuals are relatively immobile (Vargha-Khadem et al. 1998). There are also deficits in language skills (Vargha-Khadem et al. 1995; Gopnik and Crago 1991). Structural studies have shown abnormalities in several brain areas, including a bilateral reduction in the volume of the caudate nucleus thought to be the underlying pathological cause of the orofacial dyspraxia (Vargha-Khadem et al. 1998). The known *FOXP2* mutations that are associated with this disorder are a missense mutation in exon 14 in the forkhead domain in the KE family and a t(5;7)(q22;q31.2) translocation with a break point between exons 3b and 4 in individual CS (Lai et al. 2001). This association with a complex disorder affecting the basal ganglia, among other brain regions, and with clinical manifestations that include prominent abnormalities of speech and language confirms the importance of FOXP2 as a modulator of brain development and function and as a candidate for related neuropsychiatric disorders.

To facilitate investigation of the role of FOXP2 in such disorders, we sought to characterize further the genomic structure of *FOXP2* and to identify alternative splice variants. In addition, we examined the stability of the *FOXP2* glutamine-encoding region. Our results demonstrate the presence of previously undetected 5′ untranslated exons and internal exons, a splice variant that encodes a truncated version of the FOXP2 protein lacking the forkhead domain (FOXP2-S), and relative stability in the length of the glutamine-repeat encoding region.

## Materials and methods

We first identified CAGH44 and P22 (an overlapping clone) in a fetal brain cDNA library screen by using a $(CAG)_{15}$ probe (Margolis et al. 1997). CAGH44 was initially mapped to chromosome 6. Subsequently, the availability of additional genomic sequence indicated that the 3′ end of CAGH44 was chimeric and, eliminating that region from consideration, CAGH44 mapped to chromosome 7q31 consistent with the published data (Lai et al. 2001). Two probes from the overlap between P22 and CAGH44 (5′-ATCTCTGCTGCCAGCATCTAATTGGCTGCTTAGAGTGCT-CATTCC-3′ and 5′-AGGTGGAATGGAGATGAGTCCCTGAC-GCTGAAGGCTGAGCAGATG-3′) were used to screen human frontal cortex and fetal brain libraries (Stratagene), yielding clones FC2A1, FCA1, and HF2B2. To search for additional 3′ exons, we performed RACE (rapid Amplification of cDNA ends, LTI) with human brain poly-A RNA (Clontech) and two sets of nested primers (set 1: 5′-GGCCAGGCAGCACTTCC-3′ and 5′-TCCA-ATCGCTGCCTCAAG-3′; set 2: 5′-CCAAAGCCTCACCACC-3′ and 5′-GGGCCTCTCACACTCTCT-3′), designed from the 3′ end of sequence obtained from the above clones. Primer set 1 yielded clone 3Race700, primer set 2 yielded clone BA4. To confirm the 3′ end of the open reading frame derived from this group of clones, we performed reverse transcription/polymerase chain reaction (RT-PCR) on a striatal cDNA library (Stratagene) with a vector-specific primer and nested primer set 2, yielding clone STR 7. We then performed RT-PCR to amplify the entire open reading frame of *FOXP2-S* from exon 2 through exon 10+, by using RNA from human amygdala and priming cDNA synthesis from sequence in the 10+ UTR (primer EX3′ 5′-TGTTACATTGGTAGAGC-3′). PCR was performed with EX5′ (5′-ATTAAGTCATGATGCAG-3′) and Q40 EX3′ as the outer primers, and 5′-GATCCTCGAGTT-TACTGTTTATAAAGCAATATGCACT-3′ and 5′-GATCGGA-TCCCCATGCAGGAATCTGCGACAGAGACAA-3′ as inner primers. This experiment yielded two independent clones, Amyg 2a.2 and Amyg 8a.4.

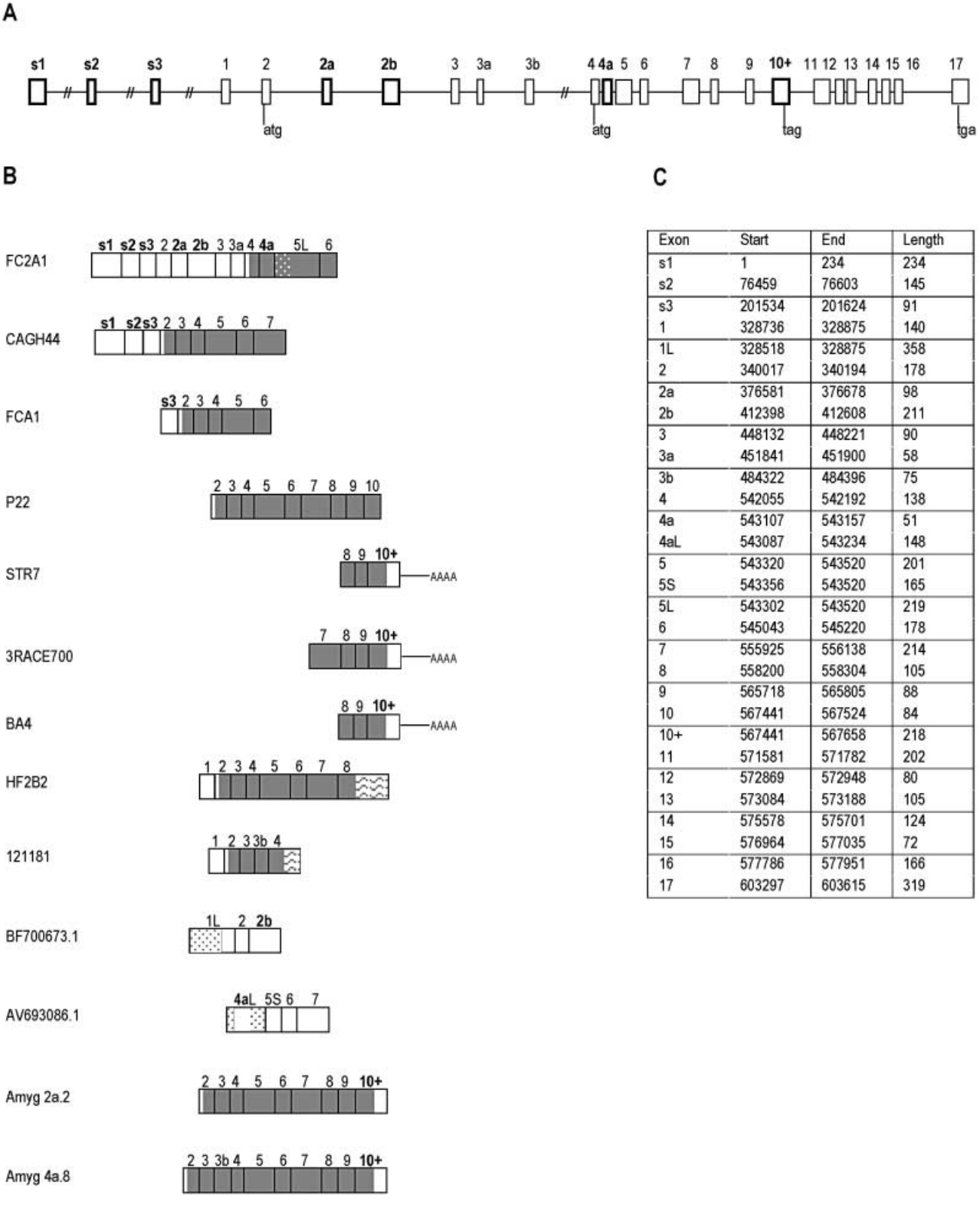
We performed basic local alignment search tool (BLASTN) searches of GenBank with our clones by using the National Center for Biotechnology Information server (www.ncbi.nlm.nih.gov/blast/blast.cgi), which yielded IMAGE clone 121181 and expressed sequence tags (ESTs) AW490098.1, BF700673.1, AV693086.1, BB656124, and BB660527.

To confirm the expression of the novel exons and splice variants we performed a series of RT-PCRs with the primers listed in Table 1. We synthesized cDNA with oligo dT priming (Superscript First Strand Synthesis System, Invitrogen) from poly-A RNA derived from human adult brain, fetal brain, caudate nucleus, and lung (Clontech). We performed PCR with Buffer J and *Taq* polymerase (Invitrogen) over 36 amplification cycles. PCR products were visualized on 1.2% (for products larger than 1000 bp) or 3% (for products smaller than 500 bp) agarose gels stained with ethidium bromide. To rectify variations in cDNA quality, digitalized images from the same PCR were spliced together.

To test repeat length stability and screen for possible expansion mutations, we amplified across the region encoding $Q_{40}X_8Q_{10}$ (by

**Table 1** RT-PCR primers

| Primer | Sequence | Tm (°C) | Target(s) |
|---|---|---|---|
| s1.5′-1 | AGCAGCTGCCCGGACTCG | 67 | s1–17, s1–10+, s1-s3 |
| 1.5′-1 | ACTGTTTTCTGTGCTGGCTTTTT | 61 | 1–17 |
| 1.5′-2 | TGAAGTTGAAACCGGGAAGTTTGC | 67 | 1–10+ |
| s3.5′-1 | GGACTCCGTTTCAAGGTTAATTCG | 64 | s3–2a, s3–2b, s3–2 |
| 4.5′-1 | GTGTCAGTGGCCATGATGACTCC | 66 | 4–4a |
| s1.5′-2 | TGTGTTGTTTGGGGGCTTCTG | 57 | s1–2 |
| 17.3′-1 | TGTCCCTTCACGCTGAGGTTTCA | 68 | s1–17, 1–17 |
| 10+.3′-1 | TGCCGTATTTTTCATCACACTCATTG | 65 | s1–10+ |
| 10+.3′-2 | TGTTCATTTTCTTAGTCTGTTACATTGG | 60 | 1–10+ |
| 2a.3′-1 | ACTGCTTGAGCCCAGGAGTTCCA | 68 | s3–2a |
| 2b.3′-1 | AACCTGGGTGTAAGAGACAACATC | 60 | s3–2b |
| 4a.3′-1 | GGCAGCTCTGAAATTTTCCAATCC | 66 | 4–4a |
| s3.3′-1 | GGAAAATTCAATCCCAGCATCAC | 63 | s1-s3 |
| 2.3′-1 | GGTGTCACCACTTGATCTTCCAT | 62 | s3–2 |
| 2.3′-2 | TCTGTCGCAGATTCCTGCATC | 56 | s1–2 |




C

| Exon | Start | End | Length |
|---|---|---|---|
| s1 | 1 | 234 | 234 |
| s2 | 76459 | 76603 | 145 |
| s3 | 201534 | 201624 | 91 |
| 1 | 328736 | 328875 | 140 |
| 1L | 328518 | 328875 | 358 |
| 2 | 340017 | 340194 | 178 |
| 2a | 376581 | 376678 | 98 |
| 2b | 412398 | 412608 | 211 |
| 3 | 448132 | 448221 | 90 |
| 3a | 451841 | 451900 | 58 |
| 3b | 484322 | 484396 | 75 |
| 4 | 542055 | 542192 | 138 |
| 4a | 543107 | 543157 | 51 |
| 4aL | 543087 | 543234 | 148 |
| 5 | 543320 | 543520 | 201 |
| 5S | 543356 | 543520 | 165 |
| 5L | 543302 | 543520 | 219 |
| 6 | 545043 | 545220 | 178 |
| 7 | 555925 | 556138 | 214 |
| 8 | 558200 | 558304 | 105 |
| 9 | 565718 | 565805 | 88 |
| 10 | 567441 | 567524 | 84 |
| 10+ | 567441 | 567658 | 218 |
| 11 | 571581 | 571782 | 202 |
| 12 | 572869 | 572948 | 80 |
| 13 | 573084 | 573188 | 105 |
| 14 | 575578 | 575701 | 124 |
| 15 | 576964 | 577035 | 72 |
| 16 | 577786 | 577951 | 166 |
| 17 | 603297 | 603615 | 319 |

**Fig. 1A–C** Structure of *FOXP2*. **A** Genomic structure of *FOXP2*. Novel exons are in *bold*. Exons longer than 200 bp are drawn twice as wide as smaller exons. Exon 10+ includes exon 10. *Hash marks* Introns longer than 50,000 bp. **B** *FOXP2* human cDNAs and ESTs. *Dark gray* Areas of open reading frame, *white* untranslated regions, *dotted pattern* use of alternate donor/acceptor splice sites that extend the length of the given exon (these exons are designated by *L* following the respective exon *number*), *S after exon number* use of an alternate donor/acceptor splice site that shortens the corresponding exon, *wavy line pattern* probable retained intronic sequence, *line with four As* poly-A tail. All pictured exons use canonical donor/acceptor splice sites except the variants of exons 4a and 5 in AV693086.1. Some cDNAs and ESTs do not contain the entire exon that is pictured at their 3′ or 5′ end. Sources: *FC2A1, FCA1* adult frontal cortex, *CAGH44, P22, HF2B2* fetal brain, *3RACE700, BA4* adult whole brain, *STR7* adult striatum, *Amyg 2A.2, Amyg 4A.4* amygdala, *121181* fetal liver and spleen, *AV693086.1* hepatocellular carcinoma, *BF700673.1* fetal primitive neuroectoderm. **C** Precise positions and lengths of *FOXP2* exons. The relative chromosomal location of each exon is listed; the first basepair of s1 is considered as basepair 1

using primers 5′-GCAAGAGCAGTTACATCT-3′ and 5′-GGAAGACAAGCTGCTGGG-3′) in DNA extracted from 142 subjects referred to the Neurogenetics Testing Laboratory at Johns Hopkins University for testing for repeat expansion diseases. These individuals had tested negative for the known repeat expansions and had progressive movement disorders of unknown etiology involving the cerebellum or the basal ganglia. Allele length was assessed by using 6% polyacrylamide gel electrophoresis of radiolabeled PCR products or automated detection of fluorescently labeled primers (Perkin-Elmer).

## Results

Previously unidentified 5′ exons were found in three independent human cDNAs from two different sources (Fig. 1B) and mouse ESTs AW490098.1, BB656124, and BB660527. The three exons, termed s1, s2, and s3, may be alternately spliced with exon 1, as no clone contains both exon 1 and an "s" exon. The "s" exons are untranslated and do not affect the open reading frame of *FOXP2*, which starts in exon 2. Mouse EST AW490098.1 aligns to exons s1, s2, s3, 2, and 3 at 83% identity. The region that aligns to exons 2 and 3 matches the first 199 bp of the open reading frame of mouse Foxp2 at 99%, indicating that the 5′ untranslated region (UTR) exons s1, s2, and s3 are part of the genomic structure of both mouse and human *FOXP2*. In mouse ESTs BB656124 and BB660527, exon s1 extends an additional 137–141 bp 5′ to the most 5′ FOXP2 human cDNA sequence that we have identified. This upstream mouse s1 sequence is 90% identical to human genomic sequence, suggesting that human exon s1 may extend further upstream than the arbitrary start site denoted in Fig. 1A.

We used Promoter Inspector (http://genomatix.gsf.de/promoterinspector; experimentally derived specificity of 85%, sensitivity of 50%; Scherf et al. 2000) to analyze the 5-kb sections of genomic sequence 5′ to the 5′ end of exons s1, s2, s3, and 1. Defining the first basepair of the most 5′ transcript of exon s1 as basepair 1, promoter regions were predicted in proximity to s1 (at –235 to +220 and –1531 to –1208; Fig. 2A), and no promoters were predicted in proximity to exons s2, s3, and 1. Analysis of 1 kb genomic sequence 5′ to the 3′ end of exons s1, s2, s3, and 1 by using the TSSG human pol II recognition algorithm (http://searchlauncher.bcm.tmc.edu/seq-search/gene-search.html; Prestridge 1995) also indicated a promoter region adjacent to exon s1, with a transcription start site at position –149 (Fig. 2A). No transcription start sites were predicted for exons s2, s3, or 1. Like the promoter regions of the previously studied fox genes *foxf1* (Chang and Ho 2001) and *foxj1* (Brody et al. 1997), the exon s1 promoter region of *FOXP2* contains no TATA box and is GC-rich (Fig. 2B). The GC content climbs to a peak of 80% at approximately –1400 bp and again at the start of exon s1; between these peaks, the GC content averages 60% (http://bioweb.pasteur.fr/seqanal/interfaces/cpgplot.html). The algorithm GrailEXP v3.3 (Hyatt et al. 1997; http://compbio.ornl.gov/grailexp/) predicts CpG islands at –1543 bp to –1245 bp and –291 to +305 bp. No CpG island was predicted in the 5′ flanking region of exon 1. The GC content of the 1.5 kb sequence upstream of exon 1 (and also of exons s2 and s3) is approximately 40% (Fig. 2C).

We used MatInspector (http://genomatix.gsf.de/mat_fam; Quandt et al. 1995) to search the 5′ flanking region of exon s1 (–1500 bp to +100 bp) for the presence of transcription-factor-binding sites (Fig. 2B). Five sites, more than double the frequency expected by chance alone, were predicted for SP1, a factor that can initiate transcription from TATAless promoters (Pugh and Tjian 1990). The Wilm's tumor (WT) suppressor site, expected to occur 0.97 times per 1 kb by chance, occurred eight times; *FOXD1*, another member of the forkhead family, is regulated by the WT suppressor (Ernstsson et al. 1996). The E2F-binding site with an e-value of 0.06 occurs in six places, three of which are within 300 bases of the start of s1. Four AP2 sites, three GC box sites, two nuclear factor 1 sites, three Egr-1/NGF1-A sites, four Ras-responsive-element-binding protein 1 sites, five neural restrictive silencer element sites, two Elk-1 sites, and one CREB site were also predicted. The nearest CCAAT box to exon s1 is at –1070 bp and may not be relevant as functional CCAAT boxes typically reside 30–150 bp upstream of the transcription start site.

We also used Matinspector to examine the flanking region of exon 1 (Fig. 2D). Unlike the flanking region of exon s1, there are no SP1, NRSE, WT, RREBP1, GC Box, Egr-1/NGF1-A, AP2, NF1, Elk-1, or CCAAT boxes, and there is just one E2F site. Instead, the exon 1 flanking region is rich in octamer-binding protein 1 sites (10), SEF1 sites (4), GATA3 sites (5), and other Fox-protein-binding sites (5). Three MyT1 sites, three GFI1 sites, and two CREB sites are also present. The nearest TATA box is at –1004 bp.

Three novel internal exons, viz., 2a, 2b, and 4a, were found in clone FC2A1 (Fig. 1). It is unlikely that exons 2a and 2b are translated, as the open reading frame that starts in exon 2 and that includes these exons ends with stop codons within them. An alternative start codon in exon 4, also used in transcripts containing exon 3a (Lai et al. 2001), may serve as the primary start site if these exons are part of the transcript. Exon 4a contains an open reading frame, as does the 18 additional base pairs in exon 5L present in this clone because of the use of an alternate donor splice site (Fig. 3). Exon 2b, accompanied by exons 1L and 2, is also found in human EST BF700673.1 (Fig. 1). Exon 4a plus exons 5–7 are in EST AV693068.1 (Fig. 1). In this EST, because of the use of alternate donor/acceptor splice sites, exon 4aL includes additional sequence extending 20 bp 3′ and 77 bp 5′ to exon 4a in FC2A1, and exon 5S lacks the first 36 bp of the 201 bp of exon 5. An extension of exon 4 that continues 288 bp beyond the published exon 4 sequence with an in-frame stop codon occurring after 33 bp and a poly-A signal present after 193 bp was found in clone 121181 (Fig. 1). There is no poly-A tail and no additional evidence supporting this exon, suggesting that it may be a retained intronic sequence. An extension of exon 8 is present in clone HF2B2 (Fig. 1); it

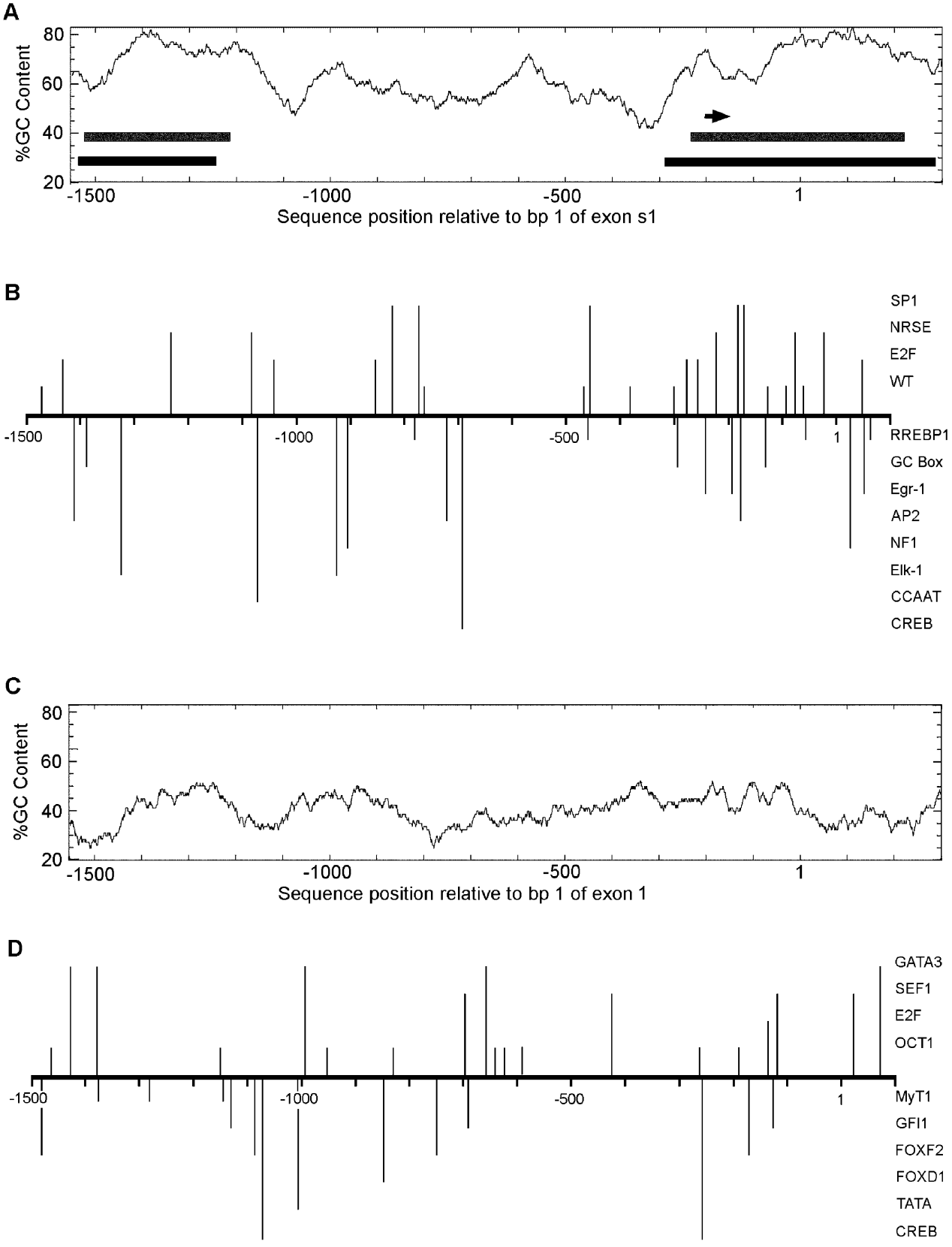



**Fig. 2A–D** Analysis of the 5′ regions flanking exons s1 and 1. **A, C** GC content in a 100-nucleotide sliding window of exon s1 and exon 1, respectively. The first basepair of each exon is numbered *1*. *Black rectangles* CpG islands, *gray rectangles* promoter regions predicted by Promoter Inspector, *black arrow* transcription start site for the promoter predicted by the TSSG algorithm. **B, D** Schematic representation of MatInspector-predicted transcription-factor-binding sites for exons s1 and 1, respectively

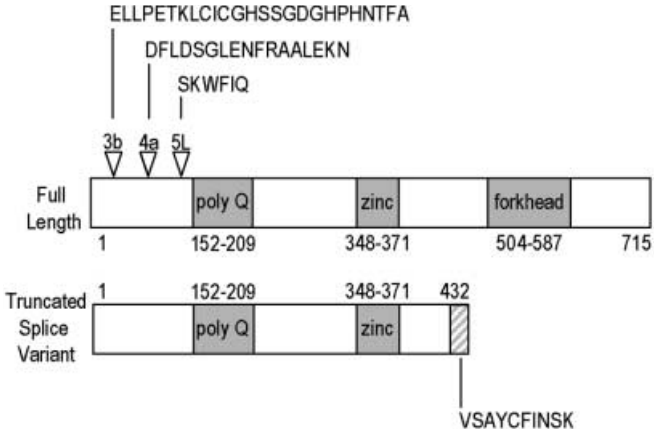



**Fig. 3** FOXP2 and FOXP2-S proteins. Both full length and truncated FOXP2 (FOXP2-S) are depicted. The glutamine repeat region, zinc finger, and forkhead domain are in *gray* with their respective amino acids *numbered*. *Above* the full length FOXP2 are the additional amino acids encoded when the three variably spliced exons with open reading frames (3b, 4a, and 5L) are present. FOXP2-S is 432 amino acids long. The last 10 amino acids are encoded by exon 10+ and are not present in the full length FOXP2

continues an additional 1041 bp and contains a stop codon immediately after the published end of exon 8, and poly-A signals after 67, 483, and 512 bp. There is no poly-A tail and no further evidence supporting this exon, so it is also likely to be a retained intronic sequence.

Exon 10+ was detected in five independent clones from three different sources of cDNA (Fig. 1B) and encodes ten additional amino acids not present in the published exon 10 (Fig. 3). A polyadenylation signal (AATAAA) is present 78 bp 3′ to the stop codon, and the poly-A tail begins after an additional 27 bp. The extra 134 bp that are present in exon 10+, but not present in exon 10, are conserved at 91% in the mouse genome. The cDNAs Amyg 2a.2 and Amyg 4a.8 (Fig. 1B) contain the entire open reading frame of FOXP2-S from exon 2–10+ and support the existence of this alternative splice variant.

We performed a series of RT-PCR experiments to confirm the expression of the novel *FOXP2* exons and splice variants detected by cDNA library screens and to investigate the possibility of tissue-specific splicing. In addition to whole brain, we examined fetal brain to search for developmental differences in the expression patterns of *FOXP2* splice variants. Since caudate is the most profoundly affected region in individuals with SPCH1 (Vargha-Kahadem et al. 1998) and *FOXP2* is important for lung development (Shu et al. 2001), we also examined these two tissues. RT-PCR demonstrated the expression of *FOXP2* transcripts that begin with either exon s1 or exon 1 and end at exon 10+ or exon 17 (Fig. 4A–D). Both *FOXP2* and *FOXP2-S* are expressed in adult brain, fetal brain, caudate, and lung. Both *FOXP2* and *FOXP2-S* exist in alternate forms that begin with either exon s1 or exon 1. There is some evidence of splice variants in all transcripts, as expected.

Given the multiple possible alternative splice variants of similarly sized small exons, we designed RT-PCR experiments to check specifically for the presence of exons s1, s2, s3, 2a, 2b, and 4a. Exons 4a, 2a, and 2b are present in at least some transcripts in all tissue types (Fig. 4E–G). Figure 4E and F show exons 2a and 2b amplified from transcripts containing exon s3; exons 2a and 2b are also expressed in transcripts containing exon 1 (data not shown). In Fig. 4F, the smaller product is consistent with the inclusion of exon 2 between exons s3 and 2b (predicted product length of 266 bp), whereas the large product is consistent with the inclusion of exons 2 and 2a between exons s3 and 2b (predicted product length of 364 bp). Transcripts containing exons s1, s2, and s3 are present in all tissue types (Fig. 4H), and trace amounts of a product 145 bp shorter is consistent with the existence of a transcript in which exon s2 is spliced out (data not shown).

We designed several RT-PCR experiments to investigate the possibility that the “s” exons and exon 1 are mu-

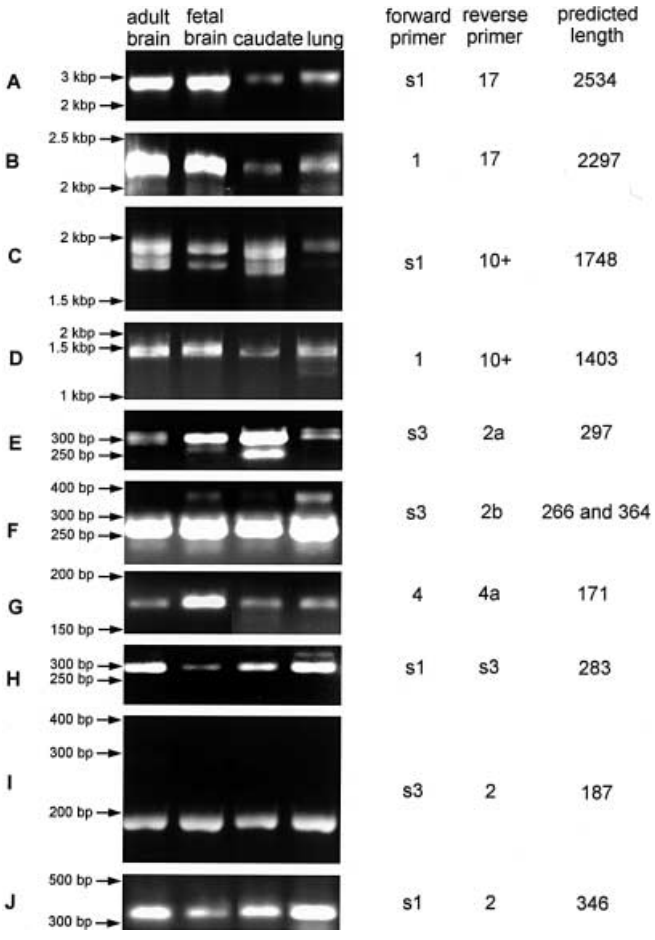



**Fig. 4A–J** RT-PCR of *FOXP2* novel exons and splice variants. PCR primers, derived from specific exons listed adjacent to the PCR results, were used to determine the presence of exons and splice variants in human whole brain, fetal brain, caudate nucleus, and lung (see Table 1 for primer sequences). In **A–D**, the listed length is the predicted RT-PCR product of each reaction assuming the absence of the variably spliced exons 2a, 2b, 3a, 3b, and 4a. The detection of products of other lengths is consistent with the inclusion of these exons in various combinations. The two predicted lengths in **F** represent products with and without exon 2a (see Fig. 1C for individual exon lengths)

tually exclusive and may represent the use of distinct promoter regions. We found no evidence of transcripts with exons between S3 and 2 (Fig. 4I). RT-PCR from exons s1, s2, or s3 to exon 1 never produced a product (data not shown). RT-PCR from exon s1 to exon 2 yielded a main product consistent with the presence of exons s1, s2, s3, and exon 2 (Fig. 4J), and a barely discernable product, 145 bp smaller than the main product, consistent with the absence of exon s2 (data not shown). Given that exon 2 appears to be invariant, these results in aggregate further suggest that *FOXP2* transcripts contain either the s1–s3 exon group or exon 1, but not both.

We also investigated the possibility that an expansion in the glutamine-repeat-encoding region of *FOXP2* might be associated with a neurodegenerative disorder. Repeat length analysis in 142 individuals with progressive movement disorders of unknown etiology revealed that the polyglutamine-encoding region of *FOXP2* is only minimally polymorphic. Two individuals had a single additional triplet in one allele, whereas others were homozygotes for a single band of the expected size. No expansions were detected.

## Discussion

Our results demonstrate the presence of additional 5′ exons, internal exons, and alternate splice variants of *FOXP2*, with no distinct differences in expression of splice variants in the four tissue types examined. The implication, supported by our sequence analysis of the 5′ region of *FOXP2*, is that at least one *FOXP2* promoter lies upstream of exon s1, more than 300 kb 5′ to the previously described exon 1. The presumed promoter region lacks a TATA box but is CpG-rich, resembling the *FOXF2* (FREAC-2, FKHL6) promoter (Blixt et al. 1998). No repeat expansions were detected in the CAG/CAA polyglutamine encoding region of *FOXP2*.

Establishing the promoter(s) for *FOXP2* is of considerable importance in understanding the spatial and temporal regulation of *FOXP2* expression. Several lines of evidence indicate that the expression of FOX genes is tightly regulated and specific levels are necessary for normal growth and development. FOX genes are known to play key regulatory roles in embryonic development, cell differentiation, and oncogenesis (Kaufmann and Knochel et al. 1996). Mutations in FOX genes cause several human developmental disorders: blepharophimosis, ptosis, and epicanthus inversus syndrome (BPES, *FOXL2*; Crisponi et al. 2001); immune dysregulation, polyendocrinopathy, enteropathy, and X-linked syndrome (IPEX, *FOXP3*; Wildin et al. 2001; Bennett et al. 2001a, 2001b); anterior segment ocular dysgenesis (ASOD and cataract, *FOXE3*; Semina et al. 2001); lymphedema and distichiasis (LD, *FOXC2*; Fang et al. 2000); anterior chamber defects associated with glaucoma (*FOXC1*; Nishimura et al. 1998; Mears et al. 1998); and thyroid agenesis (*FOXE1*; Clifton-Bligh et al. 1998). A wide spectrum of mutations have been implicated (De Baere et al. 2001; Erickson et al. 2001; Bell et al. 2001; Bennett et al. 2001a, 2001b), many affecting the forkhead domain and several thought to cause disease through a gene-dosage effect or through other mechanisms that affect expression levels. Haploinsufficiency of several FOX genes is associated with autosomal dominant diseases: BPES (Crisponi et al. 2001), LD (Fang et al. 2000), and anterior chamber defects with glaucoma (Nishimura et al. 1998; Mears et al. 1998). Either haploinsufficiency (Smith et al. 2000) or duplication of *FOXC1* (Lehmann et al. 2000; Nishimura et al. 2001) causes abnormal ocular development. In addition, translocation events that occur far upstream from *FOXC2*, *FOXC1*, and *FOXL2* are among the probable causes of LD (Fang et al. 2000), anterior chamber defects with glaucoma (Nishimura et al. 1998; Mears et al. 1998), and BPES (Crisponi et al. 2001), respectively, through position effects that interfere with control of expression. Lastly, a polyadenylation signal mutation in *FOXP3* is believed to cause IPEX through increased degradation of *FOXP3* mRNA (Bennett et al. 2001a, 2001b).

We have identified a truncated splice variant of *FOXP2*, viz., *FOXP2-S*, that does not contain the forkhead domain. A similar splice variant of mouse *Foxp1* (which has a high degree of homology to mouse *Foxp2* and human *FOXP2*) has also been described in which exons encoding a portion of the forkhead domain are excluded, resulting in the loss of forkhead function (Shu et al. 2001). Both this splice variant and *FOXP2-S* include the C2H2 zinc-finger domain and the polyglutamine domain. The zinc-finger domains in mouse and human FOXP2 proteins are identical, and in mouse, this domain has been shown to function as an independent transcriptional repressor (Shu et al. 2001). Overexpression of FOXP2-S in mammalian cells leads to cytoplasmic aggregation and toxicity (J.K. Cooper and C. Nucifora, unpublished data). The cytoplasmic localization of the over-expressed FOXP2-5, although potentially influenced by aggregation, is consistent with the recent report that the highly similar forkhead domain of FOXP3 is required for nuclear localization and DNA binding (Schubert et al. 2001). Thus, it is unlikely that FOXP2-S functions as a transcription factor.

The absence of an expansion of the *FOXP2* CAG/CAA repeat region in 142 cases of idiopathic neurodegenerative movement disorders demonstrates that expansion of this region is not a common cause of this type of disorder, although very rare expansions associated with neurodegenerative movement disorders or an association between expansion and other types of diseases cannot be excluded. The very minimal length polymorphism of the repeat region (heterozygosity of about 1%) is consistent with the absence of a repeat expansion, since most repeats that are known to undergo expansion are highly polymorphic in the normal population, typically with heterozygosity scores over 70% and a wide range of repeat length (Margolis and Ross 1998). For instance, the repeat in TBP that expands to cause SCA17 is highly polymorphic in the normal population, with heterozygosity ranging from 71% to 85% among different ethnic groups (Rubinsztein

et al. 1996) and length variation at the protein level of between 25 and 42 consecutive glutamines. Nonetheless, the *FOXP2* glutamine-encoding domain remains a candidate site for neurodegenerative diseases linked to the 7q31 region, such as a recently described autosomal dominant disorder characterized by sensory and motor neuropathy with ataxia (Brkanac et al. 2002).

Our detection of additional *FOXP2* exons may prove of value in establishing cell and animal models of FOXP2 function and also for detecting mutations in individuals with speech and language disorders similar to those seen in the KE family. The recent report that no *FOXP2* mutations have been detected in individuals with autism or specific language impairment (Newbury et al. 2002) emphasizes the importance of establishing the complete genomic structure of *FOXP2*. For instance, in both IPEX (associated with *FOXP3*) and ASOD (associated with *FOXC1*), mutations outside of presumed coding regions are postulated to explain the failure to identify mutations in affected pedigrees with linkage to the relevant FOX gene chromosomal locus (Wildin et al. 2001; Nishimura et al. 2001). Such mutations may occur in unidentified coding regions or noncoding regions where mutations may affect transcriptional regulation or RNA splicing. Our preliminary Northern analysis (with a cDNA probe spanning exons 2–4, data not shown), consistent with that reported by Lai et al. (2001), has demonstrated a 6-kb message that is approximately twice the sum of the lengths of all known *FOXP2* exons. It therefore seems likely that additional *FOXP2* exons remain to be identified.

**Acknowledgements** We thank John Kleiderlein, John Hwang, Daniel Gorelick-Feldman, and Kirsten Bottoms for technical assistance, and gratefully acknowledge the guidance and support of Christopher A. Ross, Susan E. Holmes, Doda Rudnicki, Logan Bruce, and Eric Johnson. This work was supported by a Johns Hopkins University Provost's Research Award to H.A.B., and NIH grants NS38054, MH02175, and NS16375.

## References

Bell R, Brice G, Child AH, Murday VA, Mansour S, Sandy CJ, Collin JRO, Brady F, Callen DF, Burnand K, Mortimer P, Jeffery S (2001) Analysis of lymphoedema-distichiasis families for *FOXC2* mutations reveals small insertions and deletions throughout the gene. Hum Genet 108:546–551

Bennett CL, Brunkow ME, Ramsdell F, O'Briant KC, Zhu Q, Fuleihan RL, Shigeoka AO, Ochs HD, Chance PF (2001) A rare polyadenylation signal mutation of the *FOXP3* gene (AAUAAA→AAUGAA) leads to the IPEX syndrome. Immunogenetics 53:435–439

Bennett CL, Christie J, Ramsdell F, Brunkow ME, Ferguson PJ, Whitesell L, Kelly TE, Saulsbury FT, Chance PF, Ochs HD (2001) The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of *FOXP3*. Nat Genet 27:20–21

Blixt A, Mahlapuu M, Bjursell C, Darnfors C, Johannesson T, Enerback S, Carlsson P (1998) The two-exon gene of the human forkhead transcription factor FREAC-2 (FKHL6) is located at 6p25.3. Genomics 53:387–390

Brkanac Z, Fernandez M, Matsushita M, Lipe H, Wolff J, Bird TJ, RaskindWH (2002) Autosomal dominant sensory/motor neuropathy with ataxia (SMNA): linkage to chromosome 7q22-q32. Am J Med Genet 114:450–457

Brody SL, Hackett BP, White RA (1997) Structural characterization of the mouse *Hfh4* gene, a developmentally regulated forkhead family member. Genomics 45:509–518

Chang VW, Ho Y (2001) Structural characterization of the mouse *Foxf1a* gene. Gene 267:201–211

Clifton-Bligh RJ, Wentworth JM, Heinz P, Crisp MS, John R, Lazarus JH, Ludgate M, Chatterjee VK (1998) Mutation of the gene encoding human TTF-2 associated with thyroid agenesis, cleft palate and choanal atresia. Nat genet 19:399–401

Crisponi L, Deiana M, Loi A, Chiappe F, Uda M, Amati P, Bisceglis L, Zelante L, Nagaraja R, Porcu S, Ristaldi MS, Marzella R, Rocchi M, Nicolino M, Lienhardt-Roussie A, Nivelon A, Verloes A, Schlessinger D, Gasparini P, Bonneau D, Cao A, Pilia G (2001) The putative forkhead transcription factor *FOXL2* is mutated in blepharophimosis/ptosis/epicanthus inversus syndrome. Nat Genet 27:159–166

De Baere E, Dixon MJ, Small KW, Jabs EW, Leroy BP, Devriendt K, Gillerot Y, Mortier G, Meire F, Van Maldergem L, Courtens W, Hjalgrim H, Huang S, Liebaers I, Van Regemorter N, Touraine P, Praphanphoj V, Verloes A, Udar N, Yellore V, Chalukya M, Yelchits S, De Paepe A, Kuttenn F, Fellous M, Veitia R, Messiaen L (2001) Spectrum of *FOXL2* gene mutations in blepharophimsis-ptosis-epicanthus inversus (BPES) families demonstrates a genotype phenotype correlation. Hum Mol Gen 10:1591–1600

Erickson RP, Dagenais SL, Caulder MS, Downs CA, Herman G, Jones MC, Kerstjens-Frederikse WS, Lidral AC, McDonald M, Nelson CC, Witte M, Glover TW (2001) Clinical heterogeneity in lymphoedema-distichiasis with *FOXC2* truncating mutations. J Med Genet 38:761–766

Ernstsson S, Pierrou S, Hulander M, Cederberg A, Hellqvist M, Carlsson P, Enerback S (1996) Characterization of the human forkhead gene FREAC-4 evidence for regulation by Wilm's tumor suppressor gene (WT-1) and p53. J Biol Chem 271: 21094–21099

Fang J, Dagenais SL, Erickson RP, Arlt MF, Glynn MW, Gorski JL, Seaver LH, Glover TW (2000) 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

Gopnik M, Crago MB (1991) Familial aggregation of a developmental language disorder. Cognition 39:1–50

Hurst JA, Baraitser M, Auger E, Graham F, Norell S (1990) An extended family with a dominantly inherited speech disorder. Dev Med Child Neurol 32:347–355

Hyatt D, Shah M, Olman V, Mural R, Xu Y, Uberbacher EC (1997) Automated gene identification in large-scale genomic sequences. J Comp Biol 4:325–338

Kaestner KH, Knochel W, Martinez DE (2000) Unified nomenclature for the winged helix/forkhead transcription factors. Gene Dev 14:142–146

Kaufmann E, Knochel W (1996) Five years on the wings of forkhead. Mech Dev 57:3–20

Koide R, Kobayashi S, Shimohata T, Ikeuchi T, Maruyama M, Saito M, Yamada M, Takahashi H, Tsuji S (1999) A neurological disease caused by an expanded CAG trinucleotide repeat in the TATA-binding protein gene: a new polyglutamine disease? Hum Mol Gen 8:2047–2053

Lai CSL, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. Nat Genet 413:519–523

Lehmann OJ, Ebenezer ND, Jordan T, Fox M, Ocaka L, Payne A Leroy BP, Clark BJ, Hitchings RA, Povey S, Khaw PT, Bhattacharya SS (2000) Chromosomal duplication involving the forkhead transcription factor gene *FOXC1* causes iris hypoplasia and glaucoma. Am J Hum Genet 67:1129–1135

Margolis RL, Ross CA (1998) Detection of unstable trinucleotide repeat loci: genome and cDNA screening. In: Wells RD, Warren ST (eds) Genetic instabilities and hereditary neurological diseases. Academic Press, San Diego, pp 431–438

Margolis RL, Abraham MA, Gatchell SB, Li SH, Kidwai AS, Breschel TS, Stine OC, Callahan C, McInnis MG, Ross CA (1997) cDNAs with long CAG trinucleotide repeats from human brain. Hum Genet 100:114–122

Margolis RL, McInnis MG, Rosenblatt A, Ross CA (1999) Trinucleotide repeat expansion and neuropsychiatric disease. Arch Gen Psychiatry 56:1019–1031

Mears AJ, Jordan T, Mirzayans F, Dubois S, Kume T, Parlee M, Ritch R, Koop B, Kuo W, Collins C, Marshall J, Gould DB, Pearce W, Carlsson P, Enerback S, Morissette J, Bhattacharya S, Hogan B, Raymond V, Walter MA (1998) Mutations of the forkhead/winged-helix gene, *FKHL7*, in patients with Axenfeld-Rieger anomaly. Am J Hum Genet 63:1316–1328

Nakamura K, Jeong S, Uchihara T, Anno M, Nagashima K, Nagashima T, Ikeda S, Tsuji S, Kanazawa I (2001) SCA 17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. Hum Mol Gen 10:1441–1448

Newbury DF, Bonroa E, Lamb JA, Fisher SE, Lai CS, Baird G, Jannoun L, Slonims V, Stott CM, Merricks MJ, Bolton PF, Bailey AJ, Monaco AP (2002) FOXP2 is not a major susceptibility gene for autism or specific language impairment. Am J Hum Genet 70:1318–1327

Nishimura D, Swiderski R, Alward WLM, Searby CC, Patil SR, Bennet SR, Kanis AB, Gastier JM, Stone EM, Sheffield VC (1998) The forkhead transcription factor gene *FKHL7* is responsible for glaucoma phenotypes which map to 6p25. Nat Genet 19:140–147

Nishimura D, Searby CC, Alward WL, Walton D, Craig JE, Mackey DA, Kawase K, Kanis AB, Patil SR, Stone EM, Sheffield VC (2001) 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

Prestridge DS (1995) Predicting Pol II promoter sequences using transcription factor binding sites. J Mol Biol 249:923–932

Pugh BF, Tjian R (1990) Mechanism of transcriptional activation by SP1: evidence for coactivators. Cell 61:1187–1197

Quandt K, Frech K, Karas H, Wingender E, Werner T (1995) MatInd and MatInspector – new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res 23:4878–4884

Rubinsztein DC, Leggo J, Crow TJ, DeLisi LE, Walsh C, Jain S, Paykel ES (1996) Analysis of polyglutamine-coding repeats in the TATA-binding protein in different human populations and in patients with schizophrenia and bipolar affective disorder. Am J Med Genet 67:495–498

Scherf M, Klingenhoff A, Werner T (2000) Highly specific localization of promoter regions in large genomic sequences by Promoter Inspector: a novel context analysis approach. J Mol Biol 297:599–606

Schubert LA, Jeffery E, Zhang Y, Ramsdell F, Ziegler SF (2001) Scurfin (FOXP3) acts as a repressor of transcription and regulates T cell activation. J Biol Chem 276:37672–37679

Semina EV, Brownell I, Mintz-Hittner HA, Murray JC, Jamrich M (2001) Mutations in the human forkhead transcription factor *FOXE3* associated with anterior segment ocular dysgenesis and cataracts. Hum Mol Genet 10:231–236

Shu W, Yang H, Zhang L, Lu MM, Morrisey EE (2001) 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

Smith RS, Zabaleta A, Kume T, Savinova OV, Kidson SH, Martin JE, Nishimura DY, Alward WLM, Hogan BLM, John SWM (2000) Haploinsufficiency of the transcription factors *FOXC1* and *FOXC2* results in aberrant ocular development. Hum Mol Gen 9:1021–1032

Vargha-Khadem F, Watkins KE, Alcock KJ, Fletcher P, Passingham RE (1995) 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

Vargha-Khadem F, Watkins KE, Price CJ, Ashburner J, Alcock KJ, Connelly A, Frackowiak RSJ, Friston KJ, Pembrey ME, Mishkin M, Gadian DG, Passingham RE (1998) Neural basis of an inherited speech and language disorder. Proc Natl Acad Sci USA 95:12695–12700

Wildin RS, Ramsdell F, Peake J, Faravelli F, Casanova J, Buist N, Levy-Lahad E, Mazzella M, Goulet O, Perroni L, Bricarelli FD, Byrne G, McEuen M, Proll S, Appleby M, Brunkow M (2001) X-linked neonatal diabetes mellitus enteropathy and endocrinopathy syndrome is the human equivalent of mouse scurfy. Nat Genet 27:18–20