The *SPCH1* Region on Human 7q31: Genomic Characterization of the Critical Interval and Localization of Translocations Associated with Speech and Language Disorder

Cecilia S. L. Lai,[1,*] Simon E. Fisher,[1,*] Jane A. Hurst,[2] Elaine R. Levy,[1] Shirley Hodgson,[3] Margaret Fox,[4] Stephen Jeremiah,[4] Susan Povey,[4] D. Curtis Jamison,[5] Eric D. Green,[5] Faraneh Vargha-Khadem,[6] and Anthony P. Monaco[1]

[1]Wellcome Trust Centre for Human Genetics, Oxford University, [2]Department of Clinical Genetics, Oxford Radcliffe Hospital, Oxford; [3]Genetics Centre, Guy's Hospital, [4]MRC Human Biochemical Genetics Unit, University College London, and [6]Cognitive Neuroscience Unit, Institute of Child Health, Mecklenburgh Square, London; and [5]National Human Genome Research Institute, National Institutes of Health, Bethesda

The KE family is a large three-generation pedigree in which half the members are affected with a severe speech and language disorder that is transmitted as an autosomal dominant monogenic trait. In previously published work, we localized the gene responsible (*SPCH1*) to a 5.6-cM region of 7q31 between D7S2459 and D7S643. In the present study, we have employed bioinformatic analyses to assemble a detailed BAC-/PAC-based sequence map of this interval, containing 152 sequence tagged sites (STSs), 20 known genes, and >7.75 Mb of completed genomic sequence. We screened the affected chromosome 7 from the KE family with 120 of these STSs (average spacing <100 kb), but we did not detect any evidence of a microdeletion. Novel polymorphic markers were generated from the sequence and were used to further localize critical recombination breakpoints in the KE family. This allowed refinement of the *SPCH1* interval to a region between new markers 013A and 330B, containing ~6.1 Mb of completed sequence. In addition, we have studied two unrelated patients with a similar speech and language disorder, who have de novo translocations involving 7q31. Fluorescence in situ hybridization analyses with BACs/PACs from the sequence map localized the t(5;7)(q22;q31.2) breakpoint in the first patient (CS) to a single clone within the newly refined *SPCH1* interval. This clone contains the *CAGH44* gene, which encodes a brain-expressed protein containing a large polyglutamine stretch. However, we found that the t(2;7)(p23;q31.3) breakpoint in the second patient (BRD) resides within a BAC clone mapping >3.7 Mb distal to this, outside the current *SPCH1* critical interval. Finally, we investigated the *CAGH44* gene in affected individuals of the KE family, but we found no mutations in the currently known coding sequence. These studies represent further steps toward the isolation of the first gene to be implicated in the development of speech and language.

Introduction

Between 2% and 5% of children who are otherwise normal have significant difficulties in acquiring expressive and/or receptive language, despite adequate intelligence and opportunity (Bishop et al. 1995). Strong evidence for genetic influences in developmental disorders of speech and language was found in a twin study that showed significant heritability for expressive subtypes of language impairment, both with and without articulation disorder (Bishop et al. 1995). The vast majority of

families segregating such disorders do not follow a simple Mendelian inheritance pattern, so that the results of conventional parametric linkage analysis should usually be viewed with caution.

However, in 1990, Hurst et al. described a rare case of a large extended family (known as KE) in which the speech and language disorder is clearly inherited with an autosomal dominant, monogenic mode of transmission. Some reports have suggested that the core deficit of the disorder in this family is an inability to use grammatical suffixation rules, such as those for tense, number, and gender (Gopnik 1990; Gopnik and Crago 1991). Other studies have shown that the phenotype is not as selective and is characterized by difficulties with many aspects of grammar and expressive language (Hurst et al. 1990; Vargha-Khadem and Passingham 1990; Vargha-Khadem et al. 1995, 1998). Furthermore, the phenotype involves grossly defective articulation (verbal dyspraxia), such that the speech of affected in-
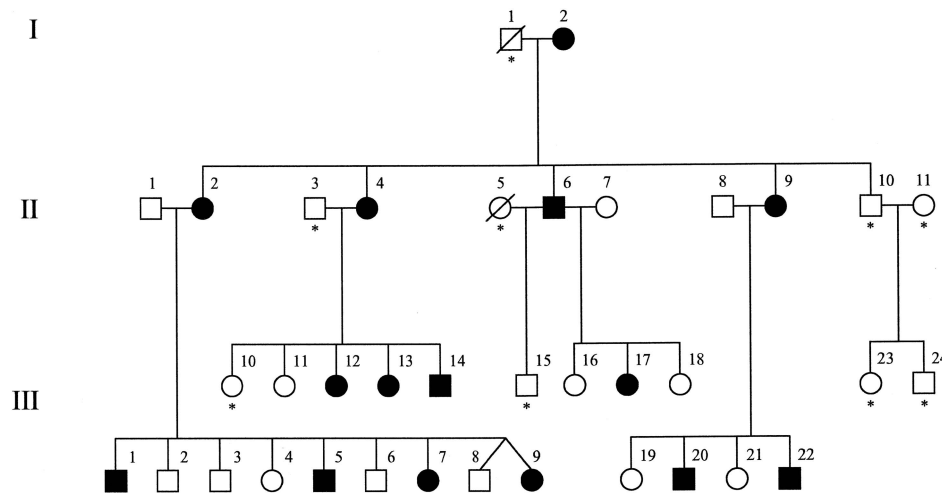
Figure 1   Pedigree of family KE, affected by speech and language disorder. Blackened symbols indicate affected individuals. Asterisks indicate those individuals who were unavailable for linkage analysis.

dividuals is largely incomprehensible to the naive listener (Hurst et al. 1990; Vargha-Khadem et al. 1995). In addition, there is evidence of moderate nonverbal cognitive impairment in some affected family members (Vargha-Khadem et al. 1995). Recent brain-imaging studies of affected individuals from this family revealed functional and structural abnormalities in both cortical and subcortical motor-related areas of the frontal lobe, particularly the basal ganglia (Vargha-Khadem et al. 1998). The findings of Vargha-Khadem et al. (1995, 1998) suggest that the central aspect of the disorder may be the disruption of selection and sequencing of fine orofacial movements, leading to deficits in the development of language skills.

We previously identified a region on chromosome 7 which cosegregates with the speech and language disorder in the KE family (maximum LOD score 6.62 at recombination fraction [θ] 0), confirming autosomal dominant inheritance with full penetrance (Fisher et al. 1998). Fine mapping using all the available Généthon microsatellites from the region allowed us to localize the gene responsible (termed SPCH1) to a ~5.6-cM interval of 7q31, flanked by markers D7S2459 and D7S643 (Fisher et al. 1998). The 7q31 band is well characterized at the physical level, with extensive coverage in YAC, BAC, and PAC clones (Bouffard et al. 1997). In addition, numerous BAC and PAC clones from this region are being actively sequenced, as part of a broader effort to sequence human chromosome 7, at the Washington University Genome Sequencing Center (WU-GSC). Several known genes and a large number of anonymous expressed-sequence tags (ESTs) have been mapped to 7q31 by YAC-based or radiation-hybrid mapping (Schuler et al. 1996; Bouffard et al. 1997).

In the present study, bioinformatic analyses have been employed to assemble a detailed BAC-/PAC-based sequence map of 7q31, which includes 152 sequence tagged sites (STSs), 20 known genes, and 50 anonymous transcripts, thus providing a framework for the positional cloning of the *SPCH1* gene. We have used 120 of these STSs to screen the affected chromosome 7 from the KE family in a search for microdeletions. In addition, we have generated novel polymorphic markers from the sequence and have used these to extract additional linkage information from the KE family, in order to refine the *SPCH1* interval. Finally, we report the localization of translocation breakpoints associated with speech and language disorder in two individuals, relative to the newly refined *SPCH1* critical interval. We have demonstrated that one of these translocations maps to the same BAC as the *CAGH44* gene, which encodes a polyglutamine-repeat protein present in brain, and, therefore, we have investigated this gene further in the KE family.

Subjects and Methods

Family KE

In 1987, the KE family (see fig. 1) was referred for genetic counseling by the director of a school for children with speech and language problems, at which many family members had been pupils. The condition was characterized as a developmental verbal dyspraxia (Hurst et al. 1990).

Patients Bearing Chromosome Translocations

CS is a 5.5-year-old boy with language impairment and verbal dyspraxia. He has a de novo balanced reciprocal translocation t(5;7)(q22;q31.2) which was iden-

tified prior to his birth via amniocentesis. Examination at birth showed no abnormalities, but he was referred back to the genetics team at age 2 years because of concerns of delayed speech and mild motor delay. Subsequent assessment at age 3 years 6 mo, performed by means of the Bailey scale, gave an overall mental development in the mildly delayed range. Although nonverbal skills were in the normal range, there was impairment in both understanding and expression of speech. A diagnosis was made of oral dyspraxia. By 4 years of age, CS was able to put two words together, and his understanding had progressed. Fine and gross motor development had also improved, although there was still evidence of mild impairment. There is no history of speech and language disorder in the family of CS, and none of his siblings (one full sibling and three half-siblings) have any language problems. His mother reports that he has never been able to laugh spontaneously or to sneeze.

BRD is an 8-year-old boy with a history of receptive and expressive language problems, accompanied by behavioral difficulties and low-range intellectual abilities, despite normal physical/motor development. He continues to show difficulties following verbal instructions in school, and has word-finding and sequencing problems accompanied by poor articulation. An MRI scan at age 6 years 11 months detected a small dysembryoplastic neuroepithelial tumor in his right temporal lobe. A more detailed description of the clinical phenotype of BRD is given in Warburton et al. (2000). Cytogenetic analysis previously revealed a de novo balanced reciprocal translocation, t(2;7)(p23;q31.3) (Warburton et al. 2000).

Construction of Human-Hamster Somatic Cell Hybrids

The Chinese hamster mutant cell line a23, deficient in thymidine kinase, was cultured in 5% CO_2 as a monolayer in Dulbecco's modification of Eagle medium (DMEM, Life Technologies) supplemented with 10% fetal calf serum (FCS). White cells were separated from peripheral blood from three affected members of the KE family (II-2, II-9, and III-20) using Histopaque (Sigma). Each sample was combined in equal proportions with trypsinized a23 cells and was placed in serum-free medium. After the cells were spun down, the medium was aspirated. The cell mixture was resuspended in 50% polyethylene glycol (PEG, molecular weight 1450, Sigma), was washed, and was set up in culture with DMEM and 10% FCS at low cell density. After 24 h, the culture medium was replaced with the same medium containing hypoxanthine, aminopterin, and thymidine (HAT, Life Technologies). Once hybrid colonies reached 2–4 mm in diameter (14 d after fusion), they were picked into 24-well plates. DNA prepared from samples of 1×10^5–5×10^5 cells was tested by PCR for chromosome 7–specific markers.
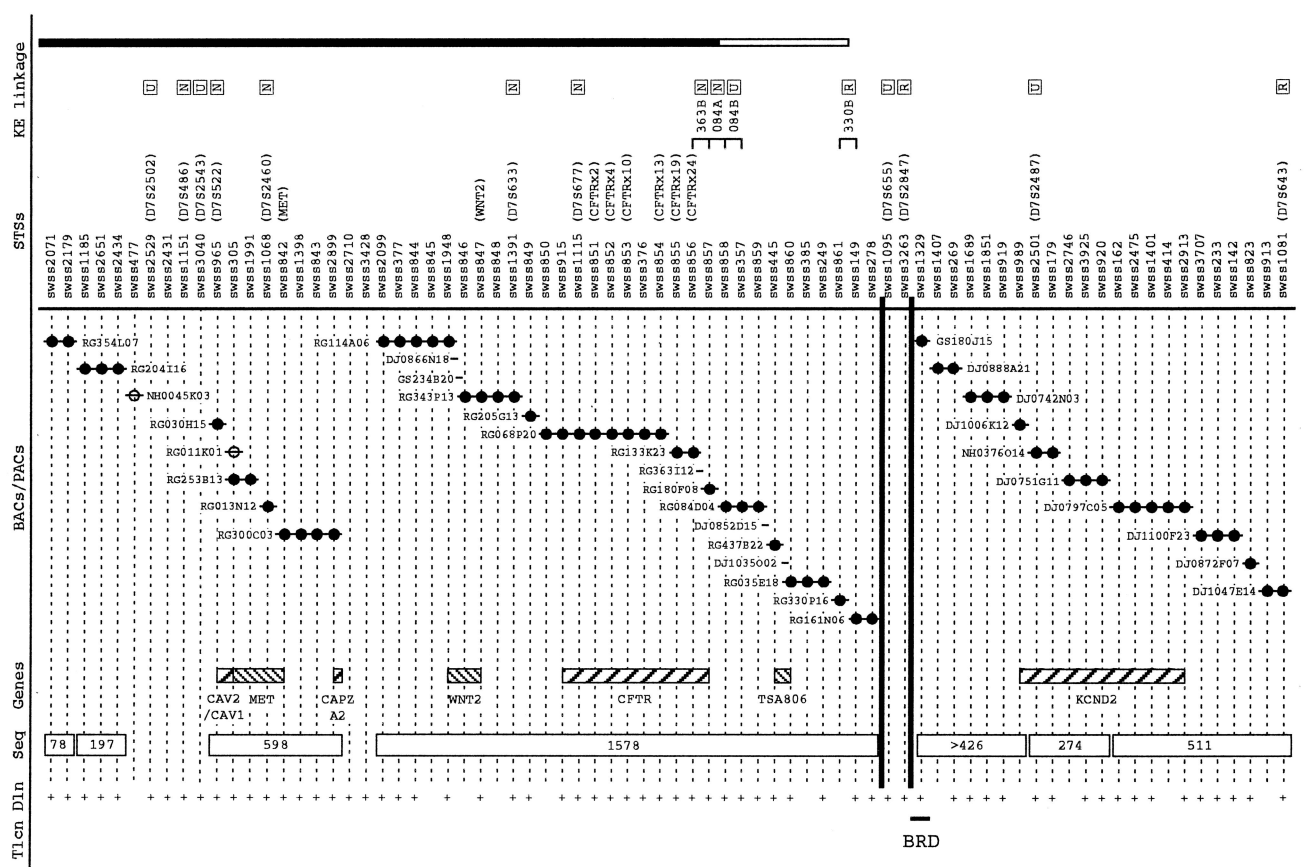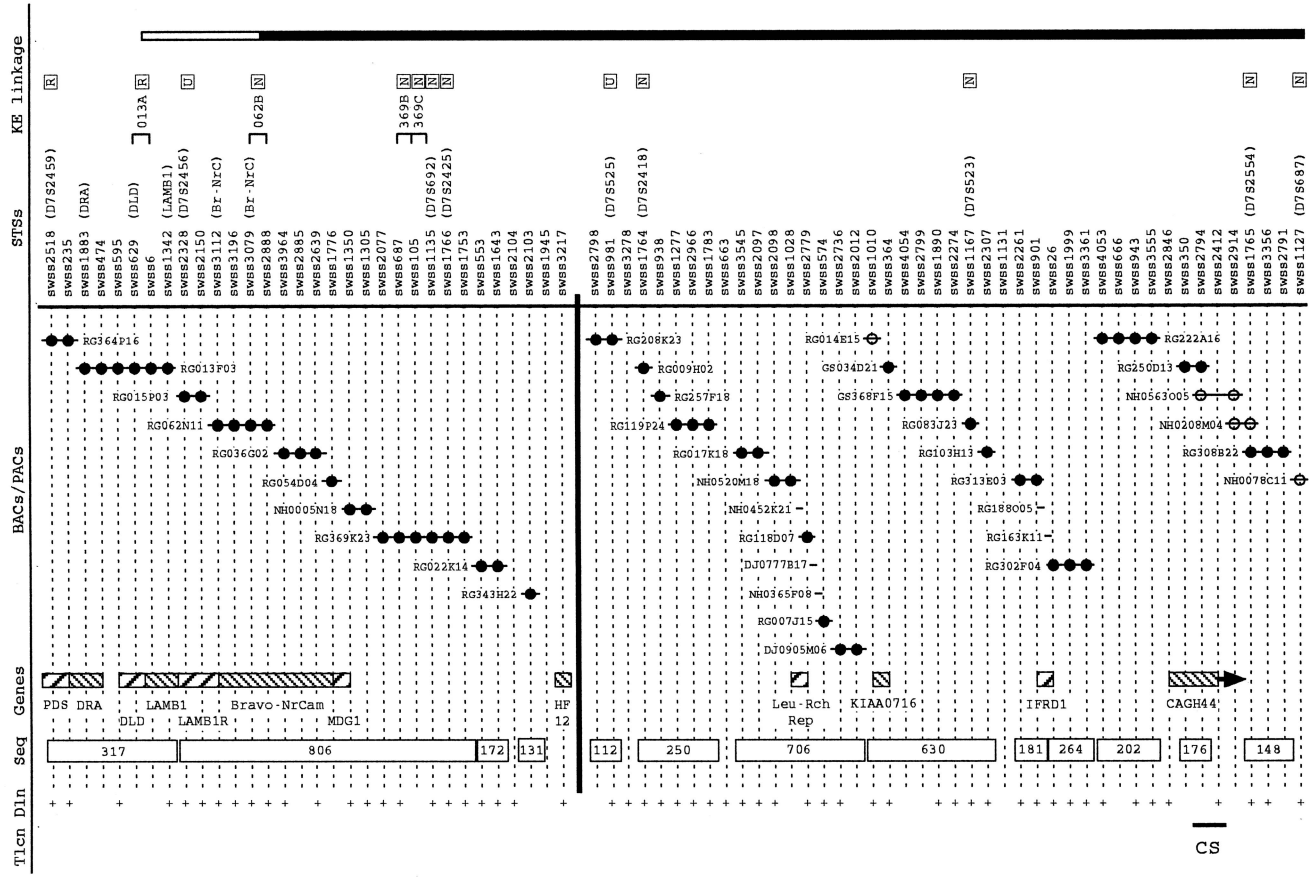
Genotyping and Linkage Analysis

Primers flanking novel polymorphic repeats were designed using the PRIMER program, accessed through the United Kingdom Human Genome Mapping Project resource center. Fluorescence-based semiautomated genotyping was performed as described (Fisher et al. 1998). Linkage analyses were run under the assumption that the disorder in the KE pedigree is due to a single autosomal dominant locus with full penetrance, as described (Fisher et al. 1998).

Fluorescence In Situ Hybridization (FISH)

PHA-stimulated T lymphocytes or lymphoblastoid cells lines were harvested by conventional techniques, and fixed suspensions were dropped onto slides. Slides were denatured at 70°C in 70% formamide/2 × SSC for 2 min 30 s, were incubated in cold 2 × SSC, and were serially dehydrated in 70%, 90%, and 100% (twice) ethanol at room temperature. Probe DNA was labeled by nick translation with biotin (Gibco BRL BioNick Labeling System) or Digoxigenin (DIG; Roche) following manufacturers' protocols. FISH of BACs and PACs was performed as described (Millwood et al. 1997). Biotinylated probes were visualized with two layers of FITC-conjugated streptavidin (green; Vector Labs) and biotinylated goat anti-streptavidin (Vector Labs). DIG-labeled probes were visualized with mouse anti-DIG antibodies (Roche), followed by Cy5-conjugated rabbit anti-mouse and goat anti-rabbit antibodies (pseudocolored blue). Chromosomes were counterstained with Vectorshield containing propidium iodide (Vector Labs). The slides were viewed on a Nikon Optiphot, and images were captured with a Bio-Rad MRC 1024 laser-scanning confocal microscope and Lasersharp software.

Determination of Genomic Organization of CAGH44

Exon/intron boundaries for exons 1–2 of *CAGH44* were identified by comparison of the reported mRNA sequence (accession U80741) to completed BAC sequence. Boundaries for exons 3–6 were determined using either long-range PCR or vectorette (Munroe et al. 1994). For the former, we amplified products from human genomic or BAC DNA, using the Expand Long Template System (Boehringer Mannheim), with primers designed from *CAGH44* mRNA. Products were sequenced using BigDye Terminator Cycle Sequencing kits (PE Applied Biosystems). For the vectorette method, libraries were made by complete digestion of BAC DNA using frequent-cutting restriction enzymes followed by ligation to annealed vectorette bubble anchors. Fragments containing exon/intron boundary sequences were amplified from vectorette libraries by PCR with the *Not*I-A primer in combination with specific primers derived from available *CAGH44* sequences.

Results

Construction of a BAC-/PAC-Based Sequence Map Spanning the SPCH1 *Interval of 7q31*

Preliminary analysis of chromosome 7 physical-mapping data indicated that 152 STSs map to the original *SPCH1* interval, between D7S2459 and D7S643, specifically within four YAC contigs, designated N, O, P, and Q (Bouffard et al. 1997). Although many BAC and PAC clones from this region have been sequenced and deposited as separate records in the public databases, much of the sequence information is not fully cross-referenced and integrated with other related sequences or with other mapping data. Therefore, we compared the sequence for all 152 STSs mapping between D7S2459 and D7S643 to the "non-redundant," "high-throughput genome sequence," and WU-GSC human sequencing project-specific databases using the BLAST algorithm (Altschul et al. 1997). At the time of the most recent analyses (March 2000), 132 of these STSs were found in fully sequenced BACs/PACs, whereas an additional 4 were detected in clones for which partial sequence was available. Thus, only 16 STSs were not found in a sequenced clone.

BLAST analysis of the insert ends of sequenced clones facilitated the establishment of overlaps among adjacent clones and orientation of various sequence blocks. These electronic analyses allowed us to assemble a detailed BAC-/PAC-based sequence map of 7q31 consisting of several blocks of contiguous sequence, the largest of which exceeds 1.5 Mb in length (fig. 2). Within these sequenced regions, we have been able to accurately determine marker order and intermarker distances. In a number of cases, the marker order differs from that previously established in physical-mapping studies. We estimate that the D7S2459–D7S643 interval currently contains >7.75 Mb of completed sequence. No sequence has been obtained, as yet, from the region covered by YAC contig P (sWSS1095-sWSS3263), and this region is estimated from physical mapping studies to span ~1.25 Mb (Bouffard et al. 1997). Therefore, these analyses indicate that the D7S2459–D7S643 interval is likely to be ≥9 Mb in size. Of note, all pairs of adjacent STSs within assembled sequence blocks are separated by <220 kb.

Transcript Map of 7q31

GeneMap '99 includes 147 ESTs that have been mapped to the interval between anchor markers D7S2459 and D7S480 using radiation-hybrid panels G3 and GB4. Analysis of these ESTs suggested that they cluster into <96 transcripts, including 18 genes that have been fully sequenced and/or encode proteins of known function. Electronic PCR and BLAST searches revealed that the majority (>85%) of the ESTs map to sequenced BACs/PACs. A total of 50 of the anonymous transcripts and 13 of the known genes from GeneMap '99 could be localized precisely within the D7S2459–D7S643 sequence contigs. (Eleven anonymous transcripts and three known genes are contained in sequenced clones that map proximal or distal to the D7S2459–D7S643 interval; the remaining 19 transcripts were not found in BAC/PAC sequences.) An additional six genes not present in the GeneMap '99 interval (*PDS, LAMB1R, MDG1, CAGH44, TSA806,* and *KCND2*) were identified and were localized within the D7S2459–D7S643 contigs via annotations in BAC/PAC sequence records. Finally, one of the STSs (sWSS3217) residing within an as-yet-un-sequenced region was found, by homology screening, to correspond to the HF.12 gene. Figure 2 shows the relative positions of all the known genes detected, with additional summary information provided in table 1.

Search for a 7q31 Microdeletion in Family KE

Previously, cytogenetic analyses of affected members of the KE family failed to detect any chromosomal ab-

---

Figure 2   Sequence map of the D7S2459–D7S643 interval. Previously ordered using YACs, 152 STSs provide a framework for the depicted BAC-/PAC-based sequence map. The most likely marker order was established by analysis of available genomic sequence data, as described in the text. For framework markers previously developed from known polymorphisms or genes, the corresponding D7 number or gene name is given in brackets after the STS name. There are three gaps in the YAC contig map of this interval, indicated by thick vertical lines. Sequence contigs, assembled using available BAC-/PAC-derived sequence data (represented by black or white circles), are aligned beneath the ordered STSs. White circles indicate that only "working draft" sequence is currently available for that clone. The prefixes RG, GS, and NH correspond to BACs derived from the Research Genetics, Genome Systems, and RPCI-11 libraries, respectively; the prefix DJ corresponds to PACs derived from the RPCI PAC library. Note that each depicted BAC/PAC is associated with a GenBank sequence record and that, in some cases, the clone contains more DNA than is represented by the sequence record (because of trimming of the sequence to minimize overlaps with adjacent clones). Many of these sequences can be assembled into large contiguous blocks, as indicated by white rectangles beneath the contigs (with sizes shown in kb). The positions of 20 known genes are indicated by shaded rectangles. *CAGH44* has only been partially characterized, so it is currently unclear how far it extends relative to the BAC/PAC contigs (indicated by an arrow). Plus signs (+) below the contigs indicate STSs used for analysis of hybrid cell lines containing the affected chromosome 7 from family KE. Positions of novel polymorphic markers generated from sequence data are given above the ordered STSs. Results of linkage analysis of family KE with all polymorphic markers are summarized at the top of the map: R = recombinant; N = nonrecombinant; U = uninformative. The interval indicated between 013A and 330B is the new critical region for *SPCH1*. Within this, all informative markers from 062B to 084A have been shown to cosegregate precisely with the disorder. The positions of the CS and BRD translocation breakpoints, localised by FISH analysis, are indicated at the bottom of the map.

normality. However, these studies did not rule out a microdeletion within the SPCH1 interval. We therefore chose to screen affected individuals for the absence of STSs within the critical region. This analysis required the prior separation of the chromosome 7 harboring the mutation from the normal homologue. Somatic hybrid cell lines containing human chromosome 7 on a hamster background were derived from three affected individuals of the KE family (see Methods). DNA from these cell lines was genotyped with the Généthon markers D7S692 and D7S522, previously shown to be polymorphic in the affected individuals. In each case, one allele precisely cosegregates with SPCH1 (Fisher et al. 1998). This allowed identification of hybrid cell lines containing only an affected chromosome 7, those containing only a normal chromosome 7, and those containing both copies of chromosome 7. Hybrid cell lines harboring only the affected chromosome 7 were tested for the presence of 120 STSs mapping within the D7S2459–D7S643 interval (see fig. 2). The average distance between these STSs is <100 kb. In all cases, the expected PCR product was generated from the hybrid cell lines but not from hamster DNA controls, indicating that all 120 STSs are present on the affected chromosome 7 in the KE family.

Generation of Novel Polymorphic Markers in 7q31 and Fine Mapping of the SPCH1 *Locus*

Our previous linkage study of the KE family with all the Généthon markers from 7q31 identified critical recombinations in affected female III-12 and unaffected male III-3 that defined D7S2459 and D7S643 as proximal and distal limits for SPCH1, respectively (Fisher et al. 1998) (see fig. 3). Within this ~5.6-cM region, we demonstrated that the ~3.6-cM interval between D7S692/D7S2425 and CFTR cosegregates perfectly with the disorder (see figs. 2 and 3). We therefore used data from the BAC/PAC sequence contigs to develop novel polymorphic markers from the D7S2459–D7S692 and CFTR–D7S643 intervals, in order to refine further the positions of the III-12 proximal and III-3 distal recombination breakpoints, respectively. Note that the only Généthon markers from these intervals (D7S2456, D7S655, and D7S2487) were uninformative with respect to the disorder in the KE family.

Finished sequence data from these intervals was searched for stretches of unbroken tandem dinucleotide repeats with copy numbers of ≥16, which would be predicted to be polymorphic (Weber 1990). PCR primers flanking these repeats were used for semiautomated genotyping of the KE family. The results were analyzed with parametric linkage programs and haplotypes were determined as in our previous linkage study. Positions of new markers relative to the STS map could be accurately established from the sequence data (see fig. 2). Genotyping with four novel markers at the proximal

end (table 2) revealed that the recombination breakpoint in individual III-12 maps between 013A and 062B-369B-369C. This refines the SPCH1 interval by a few hundred kb at the proximal end, and excludes three known genes as candidates: PDS, DRA, and DLD (fig. 2). At the distal end of the SPCH1 interval, physical mapping data indicated that a polymorphic tetranucleotide repeat, D7S2847, lies between CFTR and D7S643. Genotyping of the KE family with D7S2847 demonstrated that the recombination in individual III-3 maps proximal to this marker. Investigation of four new markers generated from the sequence data between CFTR and D7S2847 (table 2) revealed that the III-3 recombination breakpoint is localized between 363B-084A and 330B. (084B was uninformative with respect to the disorder.) This result excludes a region of >2.65 Mb, containing the KCND2 gene, from the distal end of the SPCH1 interval. In addition, it limits the SPCH1 locus to within YAC contigs N and O, with only one uncloned gap in the region. Haplotypes of the relevant markers for critical individuals from the KE family are shown in figure 3.

Two Translocation Breakpoints in 7q31 Associated with Speech and Language Disorder

We have investigated de novo balanced translocations involving 7q31 in two unrelated patients with speech and language disorder, CS 46,XY t(5;7)(q22;q31.2) and BRD 46,XY t(2;7)(p23;q31.3) (see Subjects and Methods for clinical descriptions). The 7q31 breakpoints were localized using two-color FISH to metaphase spreads, with BACs/PACs selected from proximal and distal ends of the 7q31 contigs. The breakpoints of both patients were found to map in the D7S2459–D7S643 interval (fig. 4), suggesting that the translocations might be relevant to the etiology of speech and language disorder.

Metaphase FISH with a series of further clones identified from the contigs mapped the CS breakpoint between RG250D13, which consistently hybridized to the derivative 7 [der(7)] and RG308B22, which consistently hybridized to the derivative 5 [der(5)]. Two BACs have been identified on the basis of fingerprint data to span the gap between these clones (fig. 2), and these are in the process of being sequenced. Whereas clone NH0208M04 mapped to der (5), NH0563O05 gave signals on both der(7) and der(5), suggesting that this BAC crosses the 7q31 breakpoint (fig. 4).

A recent study using FISH with YAC clones reported that the BRD translocation mapped between CFTR and D7S643 (Warburton et al. 2000). We have been able to confirm this and to localize the breakpoint at higher resolution using BAC/PAC clones. These analyses indicate that the BRD breakpoint maps within GS180J15, which gives signals on both der(7) and der(2) (fig. 4). Therefore, the 7q31 breakpoints of CS and BRD are separated by ≥3.75 Mb of DNA, as estimated from

Table 1

Known Genes Residing between D7S2459 and D7S643

Gene[a]	Protein	Function	Accession Number[b]
PDS	Pendrin	Anion transporter; mutated in Pendred syndrome	AF030880
DRA	Down-regulated in adenoma	Anion transporter; mutated in congenital chloride diarrhoea	L02785
DLD	Dihydrolipoamide dehydrogenase	Mitochondrial matrix protein; mutated in infantile lactic acidosis	NM_000108
LAMB1	Laminin beta-1 chain precursor	Component of extracellular matrix	M61916
LAMB1R	Novel laminin beta-1 related protein	Highly similar to *LAMB1*	AF172277
Bravo-NrCAM	Neuronal cell adhesion molecule	Implicated in axonal pathfinding in developing nervous system	NM_005010
MDG1	Microvascular endothelial differentiation gene 1	Expression of rat homologue increased during tube-forming process of microvascular endothelial cells	AB026908
HF12	Zinc finger protein	Involved in cell differentiation/proliferation	X07290
Leu-Rch Rep	Leucine-rich repeat protein	Possible role in neural development via protein-protein interactions	AC004142
KIAA0716	Large protein expressed in brain	Unknown	AB018259
IFRD1	Interferon-related developmental regulator	Growth factor-sensitive positive regulator of cell differentiation	Y10313
CAGH44	Polyglutamine repeat protein from brain	Long polyglutamine tract containing 40 consecutive glutamines	U80741
CAV2	Caveolin 2	Component of caveolae membranes; upregulated in response to neuronal cell injury	AF035752
CAV1	Caveolin 1	Component of caveolae membranes; interacts directly with G-protein alpha subunits	Z18951
MET	Hepatocyte growth factor receptor	Tyrosine protein kinase; mutated in papillary renal carcinoma	J02958
CAPZA2	F-actin capping protein	Binds growing ends of actin filaments; blocks exchange of subunits	U03269
WNT2	Wingless-type MMTV integration site 2	May be signaling protein affecting development of discrete regions of tissues	X07876
CFTR	Cystic fibrosis transmembrane conductance regulator	ATP-binding chloride transporter; mutated in cystic fibrosis	NM_000492
TSA806	Testis-specific ankyrin motif protein	Possible role in cell cycle control or cell fate determination	D78334
KCND2	Brain-specific potassium channel KV4.2	Involved in dendritic spike propagation	AF121104

[a] Figure 2 gives the positions of all these genes.
[b] Accession numbers are given for the nucleotide sequence of each gene.

completed sequence and the size of YAC contig P (fig. 2). Furthermore, the BRD translocation maps ⩾1.45 Mb distal to the newly refined *SPCH1* interval, as established from the most recent linkage analyses of the KE family reported above (fig. 2).

Candidate Gene CAGH44, *in the Vicinity of the CS Translocation*

CAGH44 mRNA was originally isolated from human brain in an attempt to identify novel CAG-repeat-containing cDNAs whose expansion might be implicated in the etiology of neuropsychiatric disorders (Margolis et al. 1997). The *CAGH44* product contains a stretch of 40 consecutive glutamine residues, which is the longest polyglutamine tract found thus far in an unexpanded protein. This is encoded by a combination of CAG and CAA codons, such that there are never more than five consecutive CAGs. There is a second polyglutamine

stretch, containing only 10 glutamines, encoded by $(CAG)_7(CAA)(CAG)(CAA)$, which is separated from the first stretch by eight amino acids. Currently, there is only partial cDNA sequence reported for this gene, covering 912 bases of the coding region, and no information about its genomic structure.

Although Margolis et al. (1997) previously localized *CAGH44* to 6q14-15 by radiation-hybrid mapping, the chromosome 7 physical map and sequence data indicates that, in fact, it resides in 7q31, close to the CS translocation breakpoint (fig. 2). Specifically, bioinformatic analyses found that RG250D13, the BAC hybridizing to der(7) adjacent to the CS translocation breakpoint, includes the first 258 bases of the partial *CAGH44* coding sequence, organized into two exons. Analysis of sequence data from clones linking RG250D13 to RG308B22 enabled us to orient the completed RG250D13 sequence (including the two 5′ *CAGH44*
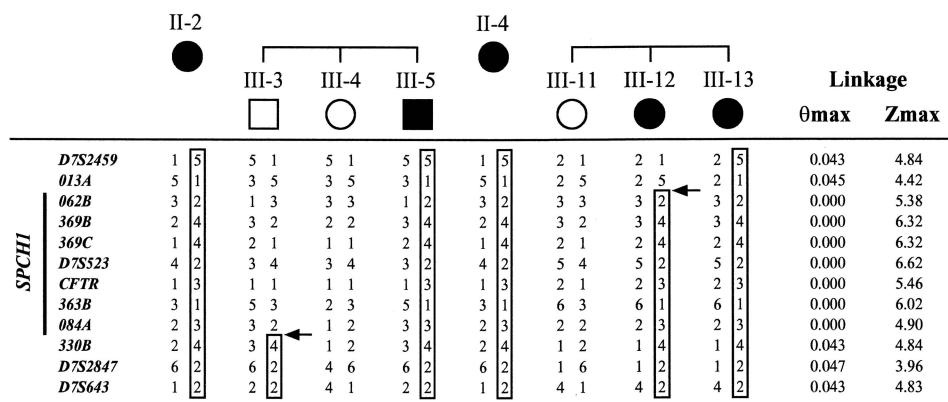
Figure 3 — Haplotype analysis (SPCH1 region)

Paternal haplotypes are on the left, maternal on the right. Boxed areas represent the haplotype that cosegregates with the disorder.

Marker (SPCH1)	III-3		III-4		III-5		III-11		III-12		III-13		θmax	Zmax
D7S2459	1	[5]	5	1	5	[5]	1	[5]	2	1	2	[5]	0.043	4.84
013A	5	[1]	3	5	3	1	5	[1]	2	5	2	1	0.045	4.42
062B	3	[2]	1	3	1	2	3	[2]	3	[2]	3	2	0.000	5.38
369B	2	[4]	3	2	3	4	3	2	3	[4]	3	4	0.000	6.32
369C	1	[4]	2	1	1	1	1	[4]	2	[4]	2	4	0.000	6.32
D7S523	4	[2]	3	4	3	2	4	[2]	5	[2]	5	2	0.000	6.62
CFTR	1	[3]	1	1	1	3	2	1	2	[3]	2	3	0.000	5.46
363B	3	[1]	5	3	2	3	5	1	3	6	[1]	6	1	0.000	6.02
084A	2	[3]	3	2	1	2	3	3	2	3	2	[3]	2	3	0.000	4.90
330B	2	[4]	3	[4]	1	2	3	[4]	2	[4]	1	2	1	[4]	1	4	0.043	4.84
D7S2847	6	[2]	6	[2]	4	6	6	[2]	1	6	1	[2]	1	2	0.047	3.96
D7S643	1	[2]	2	[2]	4	1	2	[2]	1	6	4	1	4	[2]	4	2	0.043	4.83

Figure 3 Haplotype analysis of new markers from the D7S2459–D7S643 region limits the *SPCH1* interval between 013A and 330B. This figure shows critical recombinants in unaffected individual III-3 and affected individual III-12. In addition, haplotypes are shown for the affected parent and for two sibs (one affected, one unaffected) of each of these critical individuals. Numbers used to identify individuals correspond to those in figure 1. Results from markers D7S2459, D7S523, *CFTR*, and D7S643 are taken from our previous report (Fisher et al. 1998); the remaining markers were genotyped in the present study. Haplotypes were inferred from genotype data as described (Fisher et al. 1998). Paternal haplotypes are on the left, maternal on the right. Boxed areas are used to represent the haplotype that cosegregates with the disorder. Arrows show the positions of recombination events involving the disease chromosome. All affected members not shown in this figure have inherited the nonrecombinant disorder-associated haplotype for this region. Two-point LOD score results from linkage analysis of the entire KE pedigree are given on the right side of this figure. Note that, for this localization of *SPCH1*, we are assuming that the disorder is fully penetrant.

exons) relative to the sequence contigs. This indicated that the more 3′ exons of *CAGH44* should lie on the distal side of RG250D13. Using long-range PCR and the vectorette method on BAC DNA, we were able to confirm this and to determine the genomic organization of the *CAGH44* 3′ coding region (table 3), thereby demonstrating the presence of four 3′ exons (3, 4, 5, and 6) in clone NH0563O05, in the interval between sWSS2794 and sWSS1765. This coincides exactly with the region containing the CS breakpoint (fig. 2). Bioinformatic analysis of genomic sequence from BAC RG250D13, upstream of the proposed *CAGH44* coding region, confirms that the first ATG in the reported cDNA sequence is very likely to correspond to the start of the open reading frame. Therefore, exon 1 is indeed the first coding exon of this gene, although there may be additional exons upstream of this containing 5′ untranslated sequence. However, the ORF probably extends beyond the 3′ end of the currently reported cDNA sequence, since no in-frame stop codon has yet been reached. In addition, our PCR and vectorette analyses indicate that the last 43 bases (870–912) in the cDNA sequence reported by Margolis et al. (1997) do not come from the *CAGH44* gene, but are likely to be an artifact of the original cloning procedure. This hypothesis is supported by the presence of an *Eco*RI site at 871–876 and by the observation that bases 878–912 are highly homologous (94% identity) to human ribosomal protein S16 mRNA.

Given that *CAGH44* maps to the same region as the CS breakpoint and encodes a brain-expressed polyglutamine repeat, we decided to analyze this gene in family

KE. We did not detect any expansion of the regions encoding the polyglutamine stretches in members of the family. Furthermore, sequencing of the currently known *CAGH44* coding region did not reveal any variant cosegregating with the disorder.

Discussion

Our previous investigation of family KE provided the first clues to the chromosomal location of a gene involved in speech and language disorder. In the present work, we have used the extensive bioinformatic resources available to characterize the region of 7q31 that is likely to contain this gene. A significant proportion of this interval is represented by sequenced genomic clones; 87% of the framework STSs that map between D7S2459 and D7S643 are present within complete BAC or PAC sequence entries in Genbank. When these separate entries are assembled into large blocks of contiguous sequence and are integrated with other mapping data, the resulting sequence map is a powerful tool for investigating the region of interest, as illustrated here. For example, by searching for tandem dinucleotide repeats in specific intervals of the 7q31 sequence map, we were able to generate novel polymorphic markers for refined linkage analyses, allowing us to narrow the *SPCH1* interval in family KE by several Mb.

In the absence of additional large families segregating verbal dyspraxia that might aid the fine mapping of *SPCH1*, the study of chromosomal rearrangements such as those described here provides an alternative means

Table 2

Novel Polymorphic Markers Generated from Sequence Contigs

		BAC	Primer (5′→3′)		Size	
Marker	Repeat	Clone	Forward	Reverse	BAC[a]	KE[b]
013A	$(GT)_{19}$	RG013F03	CATACTCTCCCGGCCTCAC	TGGTCCCACCTTGGTTAAAA	142	138–150
062B	$(TA)_7(CA)_{21}$	RG062N11	AGCTTTGAATACTACTGCTGCC	TGTATTCAGTGAAGTTGCCATG	191	195–221
369B	$(CA)_{21}$	RG369K23	TGGAAGAGTTTGTGATTTTCAG	AGGGTTGTTTATTCAGAGGAGG	281	275–283
369C	$(CA)_{18}$	RG369K23	TAATGTGGTTGAGCTAGGTTGG	ACCGAAGAGCCTGAAAACTG	259	259–265
363B	$(CA)_{30}$	RG363I12	GGGACTGCCCAGAGATGAC	CCTCTCCAACTTTGCTGACC	308	294–314
084A	$(CA)_{20}$	RG084D04	ACTAGAGTGCTCCCTTCAGCC	AAAATAAATTCCCACCCCTATG	302	304–308
084B	$(CA)_{25}$	RG084D04	CTGCTCAAGGCCCATCTTC	TTTTTCCATCCGTTTTCTGC	241	233–243
330B	$(TG)_{16}$	RG330P16	CTACCATAATTTCCTCCCTCCC	ACCTTCATTCAACTTCCCCC	279	281–293

[a] Size of the PCR product in bases, as predicted from the BAC sequence.

[b] Actual size range of the products amplified from the KE family, as determined by fluorescent genotyping.

of narrowing the search for the gene. However, we note that observation of a breakpoint mapping to the critical interval in a patient affected with speech and language disorder does not in itself provide sufficient evidence of a causal role for the chromosomal abnormality. In addition, the breakpoint of a chromosomal rearrangement can sometimes map outside the transcription and promoter regions of the gene implicated in the etiology of the disease but still disrupt expression of this gene via a "position effect." Some studies have demonstrated position effects acting as far as 900 kb from the gene responsible for the disorder (see Kleinjan and van Heyningen 1998). Therefore, drawing conclusions from the mapping of a chromosomal rearrangement can be difficult, unless they are substantiated by converging data from multiple cases or other kinds of analysis.

We have demonstrated that, although both the CS and BRD breakpoints map to 7q31, they are separated by >3.75 Mb of DNA. Thus, even allowing for the possibility of position effects, it seems unlikely that these breakpoints are influencing a single locus and that the patients have the same disease etiology. The developmental delay of BRD appears to be less selective than that of CS, involving behavioral problems and some general cognitive deficit, in addition to his speech and language difficulties. Warburton et al. (2000) previously mapped the BRD breakpoint to the *CFTR*–D7S643 interval and suggested that it may disrupt *SPCH1*. Our fine mapping of the BRD translocation, in combination with the refinement of the critical region from new linkage analyses of the KE family, indicates that this breakpoint in fact maps >1.45 Mb outside the current *SPCH1* interval. The clone spanning the BRD breakpoint has been fully sequenced, but no transcripts have been found by electronic analyses of this sequence. Although it is possible that the BRD translocation could alter expression of the brain-specific *KCND2* gene, whose promoter lies a few hundred kb distal to the breakpoint, our fine-mapping linkage results have excluded that lo-

cus as a candidate for SPCH1. It is worth noting that BRD has a right temporal lobe tumor, which, by virtue of its developmental origin, could interfere with the emergence of cognitive abilities, including those of speech and language.

In contrast, our analysis of the CS translocation places it within a single BAC (NH0563O05) inside the refined *SPCH1* critical interval and suggests that it may disrupt the locus for *CAGH44,* a brain-expressed gene encoding a long polyglutamine tract. NH0563O05 is currently being sequenced and we are in the process of identifying the exact position of the breakpoint in order to confirm that the CS translocation does indeed interrupt the *CAGH44* coding region. As stated above, further evidence will be required to properly establish a causal role of this breakpoint and the *CAGH44* gene in the speech and language difficulties of CS.

The *CAGH44* gene was first identified in a search for loci encoding polymorphic glutamine repeats whose expansion might cause neurological diseases (Margolis et al. 1997). Analysis of the polyglutamine-rich region showed that it is completely stable in normal individuals (Margolis et al. 1997), most likely because the CAG repeats that encode it are frequently interrupted by CAA codons, suggesting that they are unlikely to undergo the multistep gradual expansion typically associated with CAG-repeat disorders. Nevertheless, a recent study has shown that expansion of an impure CAG repeat (that of TATA-binding protein) can occur and can lead to a neurological disease (Koide et al. 1999). Therefore, on the basis of this and the results from the mapping of the CS translocation, *CAGH44* appears to be a good candidate for *SPCH1*. Our mutation analysis of the currently known coding region for this gene did not detect expansion of the polyglutamine stretches or, indeed, any other variants cosegregating with the disorder in family KE. Since the most 3′ portion of the open reading frame was not isolated in the Margolis et al. (1997) study, we are at present identifying and fully characterizing the
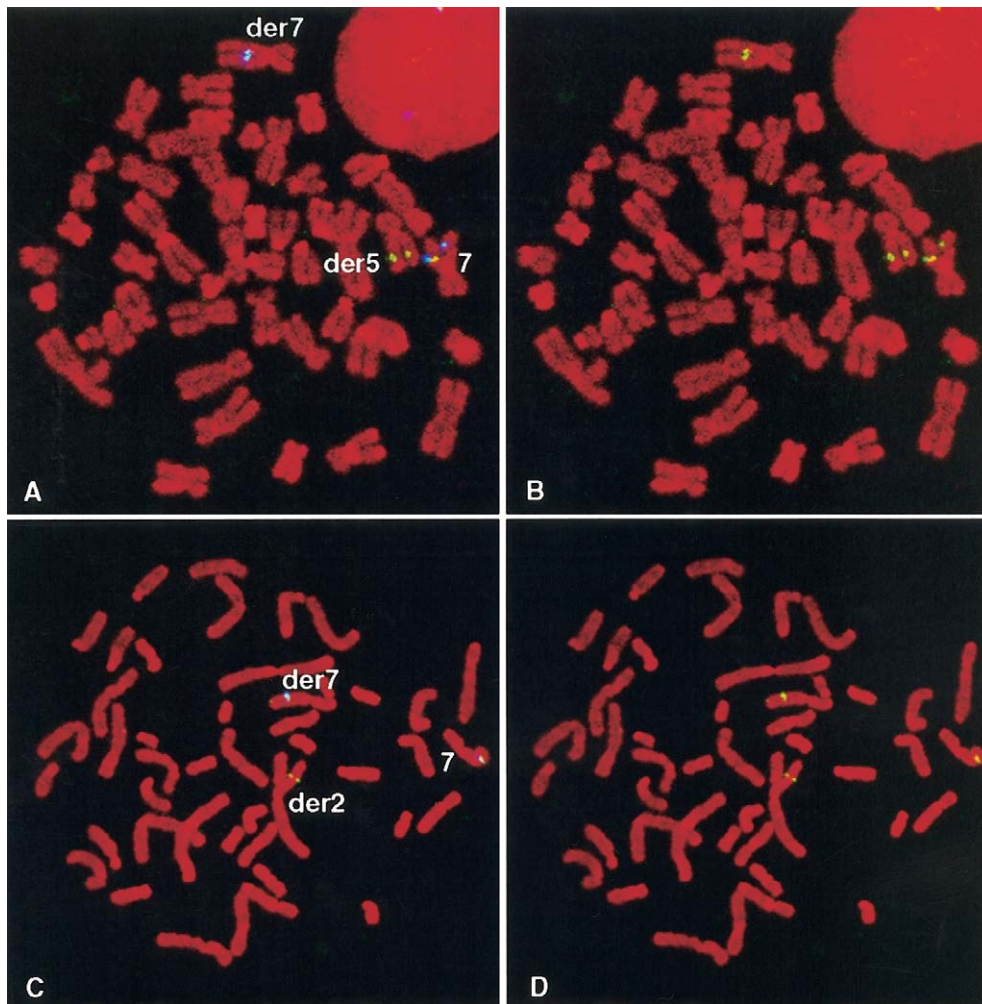
Figure 4 Two-color FISH analyses of translocations in patients with speech and language disorder, using BACs from 7q31. *A,* Patient CS; RG250D13 (blue) hybridizes to normal 7 and der(7) and is proximal to the translocation breakpoint, while NH0563O05 (green) hybridizes to normal 7, der(7) and der(5), spanning the breakpoint. *B,* Same metaphase as in 4A, with RG250D13 (blue) removed computationally, leaving only signal from NH0563O05 (green). *C,* Patient BRD; RG330P16 (blue) hybridizes to normal 7 and der(7), proximal to the breakpoint, while GS180J15 (green) hybridizes to normal 7, der(7) and der(2), spanning the breakpoint. *D,* Same metaphase as in 4C with RG330P16 removed computationally.

CAGH44 gene so that we may investigate it further in the KE family and in the patient CS.

We previously commented on the fact that the *SPCH1* interval overlaps with a ∼40-cM region identified in a genome screen for susceptibility to autism, a disorder which is often associated with speech and language abnormalities (Fisher et al. 1998; International Molecular Genetic Study of Autism Consortium 1998). Further support for a gene influencing autistic disorder on 7q has been found in additional linkage studies, but regions of linkage vary between data sets; whereas some show overlap with *SPCH1,* others implicate a region distal to this (Ashley-Koch et al. 1999; Barrett et al. 1999; International Molecular Genetic Study of Autism Consortium 1999; Philippe et al. 1999). We, and others,

have identified chromosomal rearrangements (including translocations, inversions, and a duplication) involving 7q31 in patients with autism (Ashley-Koch et al. 1999; Vincent et al. 1999; Warburton et al. 2000; C. S. L. Lai, S. E. Fisher, P. A. Jacobs, E. R. Levy, C. G. Woods, and A. P. Monaco, unpublished data). However, breakpoints for these rearrangements map to a number of different sites, some within the *SPCH1* interval and others distal to it, in a broad region of 7q31. Therefore, no simple relationship has yet emerged between the positions of chromosomal breakpoints in 7q and autism or language disorder. In addition, we note that the severe speech and language difficulties experienced by family KE are distinct from those normally associated with autistic disorder and that no members of this family have ever

Table 3

Genomic Organization of *CAGH44*

Exon	3′ Splice Site[a]	Exon Size[b]	5′ Splice Site[a]	BAC Clone	Method Used[c]
1		168	ACAACAGCAGgtaagttttg	RG250D13	GS
2	ttacttctagGCTCTCCAGG	90	ACCACTGCAGgttagtaaag	RG250D13	GS
3	tctgtgcaagGTGCCTGTGT	138	GCTGCAGCAGgtaatgtgg	NH0563O05	3′: VP / 5′: LP
4	gtttattcagCAACATCTAC	201	AGCGAAAGAGgtaggatccg	NH0563O05	LP
5	ctgataccagCAGCAGCAGC	178	CTGCCTCAAGgtacatacaa	NH0563O05	3′: LP / 5′: VP
6	cattttatagCTGGCTTAAG	>94		NH0563O05	VP

NOTE.—For mutational analysis, primers were designed to flank each of the CAGH44 exons using the genomic sequence determined here. These were used for PCR amplification of DNA from affected and unaffected individuals of the KE family, and from the hybrid cell lines containing the affected chromosome 7.

[a] Exonic sequence is in uppercase, intronic in lowercase.

[b] Size of exon 1 is estimated from ATG at start of *CAGH44* mRNA reported sequence (U80741). Exon 6 is likely to continue beyond the end of U80741.

[c] GS = comparison to BAC genomic sequence obtained from Genbank; VP = vectorette PCR followed by sequencing; LP = Long-range PCR followed by sequencing.

shown any autistic features. Thus, there is insufficient evidence from current clinical, cytogenetic, and linkage data to resolve the issue of whether there is a single 7q31 locus responsible for *SPCH1* and autism susceptibility or two separate, adjacent loci contributing independently to these disorders.

In conclusion, our genomic characterization of the *SPCH1* region has provided a framework for the investigations of family KE and the translocation patients presented in this report. These studies represent further steps towards the isolation of the first gene to be implicated in the development of speech and language.

Acknowledgments

Electronic-Database Information

URLs for data in this article are as follows:

Electronic PCR screening, http://www.ncbi.nlm.nih.gov/STS
GeneMap '99, http://www.ncbi.nlm.nih.gov/genemap (for radiation-hybrid map data)
HGMP, http://www.hgmp.mrc.ac.uk (for PRIMER program)
NCBI BLAST, http://www.ncbi.nlm.nih.gov/BLAST/ (for homology searches of sequence data)
NHGRI chromosome 7–mapping data, http://genome.nhgri.nih.gov/chr7
UniGene, http://www.ncbi.nlm.nih.gov/UniGene/ (for clustering of ESTs)
WU-GSC chromosome 7 sequencing data and BLAST searches, http://genome.wustl.edu/gsc

References

Altschul, SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Ashley-Koch A, Wolpert CM, Menold MM, Zaeem L, Basu S, Donnelly SL, Ravan SA, et al (1999) Genetic studies of autistic disorder and chromosome 7. Genomics 61:227–236

Barrett S, Beck JC, Bernier R, Bisson E, Braun TA, Casavant TL, Childress D, et al (1999) An autosomal genomic screen for autism. Am J Med Genet 88:609–615

Bishop DVM, North T, Donlan C (1995) Genetic basis for specific language impairment: evidence from a twin study. Dev Med Child Neurol 37:56–71

Bouffard GG, Idol JR, Braden VV, Iyer LM, Cunningham AF, Weintraub LA, Touchman JW, et al (1997) A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. Genome Res 7:673–692

Fisher SE, Vargha-Khadem F, Watkins KE, Monaco AP, Pembrey ME (1998) Localisation of a gene implicated in a severe speech and language disorder. Nat Genet 18:168–170

Gopnik M (1990) Feature-blind grammar and dysphasia. Nature 344:715

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

International Molecular Genetic Study of Autism Consortium (1998) A full genome screen for autism with evidence for linkage to a region on chromosome 7q. Hum Mol Genet 7:571–578

International Molecular Genetic Study of Autism Consortium (1999) Linkage disequilibrium mapping and genome screen follow-up for autism susceptibility loci. Mol Psych 4:S14

Kleinjan DJ, van Heyningen V (1998) Position effect in human genetic disease. Hum Mol Genet 7:1611–1618

Koide R, Kobayashi S, Shimohata T, Ikeuchi T, Maruyama M, Saito M, Yamada M, et al (1999) A neurological disease

caused by an expanded CAG trinucleotide repeat in the TATA-binding protein gene: a new polyglutamine disease? Hum Mol Genet 8:2047–2053

Margolis RL, Abraham MR, Gatchell SB, Li SH, Kidwai AS, Breschel TS, Stine OC, et al (1997) cDNAs with long CAG trinucleotide repeats from human brain. Hum Genet 100: 114–122

Millwood IY, Bihoreau MT, Gauguier D, Hyne G, Levy ER, Kreutz R, Lathrop M, et al (1997) A gene-based genetic linkage and comparative map of the rat X chromosome. Genomics 40:253–261

Munroe DJ, Haas M, Bric E, Whitton T, Aburatani H, Hunter K, Ward D, et al (1994) IRE-bubble PCR: a rapid method for efficient and representative amplification of human genomic DNA sequences from complex sources. Genomics 19: 506–514

Philippe A, Martinez M, Guilloud-Bataille M, Gillberg C, Rastam M, Sponheim E, Coleman M, et al (1999) Genome-wide scan for autism susceptibility genes. Hum Mol Genet 8:805–812

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, et al (1996) A gene map of the human genome. Science 274:540–546

Vargha-Khadem F, Passingham RE (1990) Speech and language defects. Nature 346:226

Vargha-Khadem F, Watkins K, Alcock K, Fletcher P, Passingham R (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, et al (1998) Neural basis of an inherited speech and language disorder. Proc Natl Acad Sci USA 95:12695–12700

Vincent JB, Herbrick J-A, Gurling HMD, Scherer SW (1999) Identification of genes at translocation breakpoints on chromosome 7q31 in autistic individuals. Mol Psych 4:S65

Warburton P, Baird G, Chen W, Morris K, Jacobs BW, Hodgson S, Docherty Z (2000) Support for linkage of autism and specific language impairment to 7q3 from two chromosome rearrangements involving band 7q31. Am J Med Genet 96: 228–234

Weber JL (1990) Informativeness of human (dC-dA)n•(dG-dT)n polymorphisms. Genomics 7:524–530