Human Molecular Genetics, 1998, Vol. 7, No. 7 **1071–1082**

ARTICLE

WHSC1, a 90 kb SET domain-containing gene, expressed in early development and homologous to a *Drosophila* dysmorphy gene maps in the Wolf–Hirschhorn syndrome critical region and is fused to *IgH* in t(4;14) multiple myeloma

Ingrid Stec[1], Tracy J. Wright[2], Gert-Jan B. van Ommen[1], Piet A. J. de Boer[3], Arie van Haeringen[4], Antoon F. M. Moorman[3], Michael R. Altherr[2] and Johan T. den Dunnen[1,4,*]

[1]MGC-Department of Human Genetics and [4]Department of Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands, [2]Genomics Group, Los Alamos National Laboratory, Los Alamos, NM, USA and [3]Institute of Anatomy and Embryology, Academic Medical Center, Amsterdam, The Netherlands

Received March 3, 1998; Revised and Accepted April 21, 1998

Wolf–Hirschhorn syndrome (WHS) is a malformation syndrome associated with a hemizygous deletion of the distal short arm of chromosome 4 (4p16.3). The smallest region of overlap between WHS patients, the WHS critical region, has been confined to 165 kb, of which the complete sequence is known. We have identified and studied a 90 kb gene, designated as *WHSC1*, mapping to the 165 kb WHS critical region. This 25 exon gene is expressed ubiquitously in early development and undergoes complex alternative splicing and differential polyadenylation. It encodes a 136 kDa protein containing four domains present in other developmental proteins: a PWWP domain, an HMG box, a SET domain also found in the *Drosophila* dysmorphy gene *ash*-encoded protein, and a PHD-type zinc finger. It is expressed preferentially in rapidly growing embryonic tissues, in a pattern corresponding to affected organs in WHS patients. The nature of the protein motifs, the expression pattern and its mapping to the critical region led us to propose *WHSC1* as a good candidate gene to be responsible for many of the phenotypic features of WHS. Finally, as a serendipitous finding, of the t(4;14) (p16.3;q32.3) translocations recently described in multiple myelomas, at least three breakpoints merge the *IgH* and *WHSC1* genes, potentially causing fusion proteins replacing *WHSC1* exons 1–4 by the IgH 5′-VDJ moiety.

INTRODUCTION

Wolf–Hirschhorn syndrome (WHS; OMIM 194190) (1,2) is a multiple malformation syndrome caused by the loss of one copy of a distal segment of chromosome 4p. The main features of WHS patients are a prominent forehead with widely spaced eyes (known as ‘Greek helmet’) and a divergent strabism, brain anomalies causing mental retardation and growth retardation. Frequently, patients suffer from heart defects and several deficiencies of midline fusion e.g. cleft lip/palate, colobomata (iris defects), hernia diaphragmatica, omphalocele and hypospadia (1–4). WHS is allelic with the milder Pitt–Rogers–Danks syndrome (PRDS; OMIM 262350) (5–8). Patients affected by PRDS suffer from multiple congenital anomalies, overlapping in part with the symptoms of WHS.

Up to now, only patients with a phenotype of WHS who have a microscopically detectable 4p deletion are considered to be affected by this syndrome, hence the cytogenetic description 4p– is used synonymously. Patients without an obvious deletion often remain ambiguous or undiagnosed. To date, no cases of typical WHS phenotype without deletion have been described. One undeleted phenotypic PRDS patient has been described by Donnai (9).

Due to the large size of the deletions, which usually span several megabases, and the high variability of involvement of inner organs, WHS may well be a contiguous gene syndrome

*To whom correspondence should be adressed. Tel: +31 71 5276105; Fax: +31 71 5276075; Email: ddunnen@ruly46.medfac.leidenuniv.nl

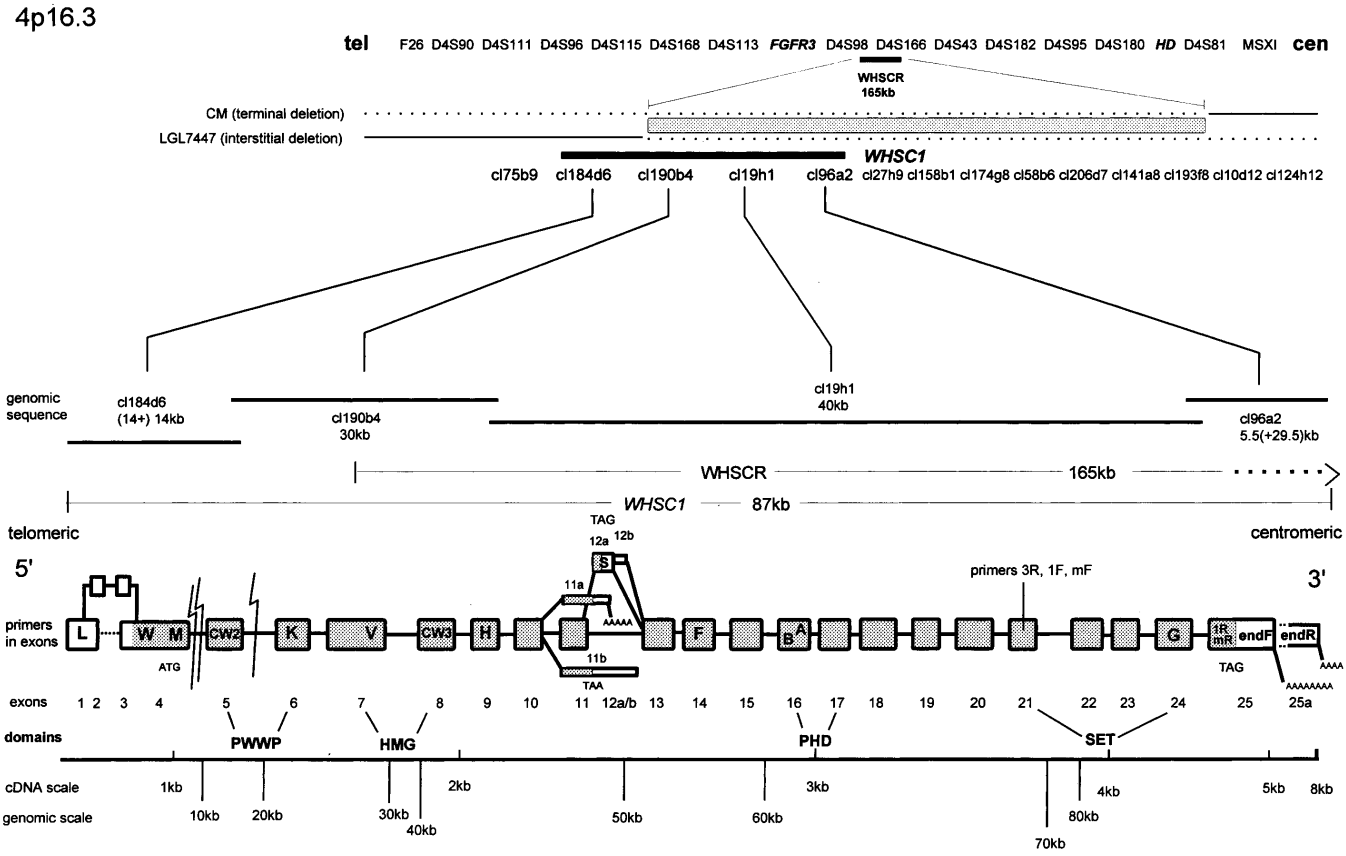

Figure 1. Gene structure of *WHSC1*. Top: 4p16.3 markers are shown from telomere to centromere. The WHSCR is located between the *FGFR3* gene locus and the Huntington disease (HD) gene. Middle: cosmid contig overlapping the WHSCR. Smallest region of overlap determined by two deleted WHS patients [cell lines CM and LGL7447 (13)] (deleted region dotted line, WHSCR shadowed, *WHSC1* bold line). *WHSC1* covers cosmids cl96a2, cl19h1, cl190b4 and cl184d6. Bottom: exon–intron organization of *WHSC1*. Boxes represent exons, lines represent introns, primers in exons are given in upper case letters. Exon 4 contains the translational start codon ATG, exon 25 contains the translational stop codon TAG. The ORF is shadowed in light grey. Lightning symbols inidicate the breakpoints of the (4;14) translocations in the multiple myeloma cell lines in *WHSC1*. Alternative splicing is found in at least three different regions indicated by superpositioned exons. Different polyadenylation sites ending in a poly(A) tail are indicated by poly(A) stretches. Four domains are detected, demonstrated below the exon numbering: PWWP domain, HMG box, PHD-type zinc finger domain and SET domain. cDNA length scale and the corresponding length of genomic sequence are shown at the very bottom of the figure.

caused by the loss of more than one gene. In ~85% of WHS cases, a *de novo* deletion of 4p16 in one of the parents' germ cells has occurred. In ~15%, the deletion is the result of an unbalanced translocation in the affected child transmitted by one of the parents carrying a balanced 4p translocation. During embryonic development, haploinsufficiency of genes influencing morphogenesis often has profound effects and may lead to a broad range of phenotypic features such as found in a number of (micro) deletion syndromes, e.g. Rubinstein–Taybi syndrome (OMIM 180849; deletion of 16p13.3) (10), Smith–Magenis syndrome (SMS; OMIM 182290; deletion of 17p11.2) (11) or velo-cardio-facial syndrome (VCFS; OMIM 192430; deletion of 22q.11.2) (12). This may apply to WHS as well. If locus control regions are deleted, gene expression may be disturbed over very large distances, i.e. position effects on the boundaries of coordinately expressed genes.

The gene defect(s) underlying both WHS and PRDS is (are) unknown. Based on the minimal overlap of deletions of different WHS patients (Fig. 1), the WHS critical gene region (WHSCR) has been confined to only 165 kb (13). This region has been sequenced completely during the search for the Huntington disease gene (14). Wright *et al*. reported different transcription units within this candidate gene region. In the present report, we describe a novel developmental gene two-thirds of which maps in the distal part of the WHSCR. We have designated this gene as *WHSC1* for Wolf–Hirschhorn syndrome candidate 1. *WHSC1* was identified initially due to the high similarity of the translation product of an expressed sequence tag (EST) contig Hs.110457, located in the 165 kb WHSCR and previously not reported, with the so-called SET domain of the *Drosophila* protein ASH1 (absent, small or homeotic discs) (15). The SET domain [for suppressor of variegation, enhancer of zeste and Trithorax (16), see Discussion] is found in proteins which are involved in embryonal development (15). *WHSC1* merges two of the transcripts in the 165 kb region, HFBEP10 and 194164, reported previously by Wright *et al*. (13). Its expression profile, especially in the tissues affected in WHS and PRDS, and its deduced function make it an excellent candidate gene for causal involvement in the phenotype of both syndromes.

RESULTS

Our strategy to complete *WHSC1* (Fig. 1) utilized database comparisons, primarily dbEST and gene prediction programs, to

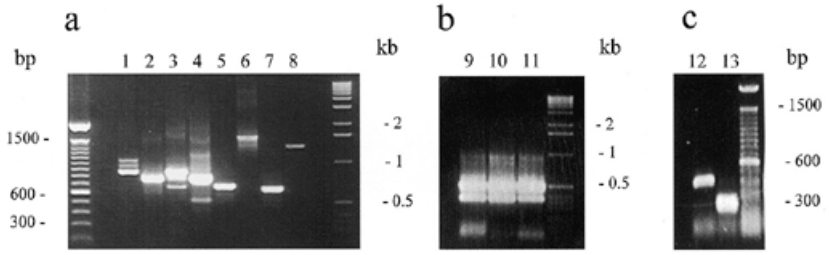

Figure 2. RT-PCR analysis in *WHSC1* on RNA. (**a**) Lane 1, 850 bp, primers L/K spanning exons 1, 4, 5 and 6 (on blood-derived RNA); lane 2, 780 bp, primers CW2/CW3 spanning exons 5–8 (blood); lane 3, 880 bp, primers H/B spanning exons 9, 10, 11 and 13–16 (testis); lane 4, 820 bp, primer combination V/F spanning exons 7–11 and 13 and 14 (testis); lane 5, 720 bp, primers A/3R spanning exons 16–20 (testis); lane 6, 1600 bp, primer combination A/mR spanning exons 16–25 (blood); lane 7, 550 bp, primers G/1R spanning exons 24 and 25 (blood); lane 8, 1320 bp, primers 1F/1R spanning exons 20–25 (blood). (**b**) Alternative splicing. Lanes 9 (blood), 10 (heart) and 11 (brain), primers S/F; 550/370 bp spanning exons 12b and 13 and exons 12a and 13 respectively. (**c**) Lanes 12 and 13, primers L/W: lane 12, brain, 400 bp spanning exons 1, 2, 3 and 4; lane 13, heart, 230 bp spanning exons 1 and 4. The 100 bp and 1 kb ladders are used as size markers (left and/or right of lanes).

design primers for RT-PCR analysis. Using overlapping RT-PCR sets, starting from the 3′ end, we extended the gene in the 5′ direction until we failed to obtain 4p16.3-derived amplification products. Transcription covering the entire gene could be detected in adult blood, brain and heart. After a nested PCR reaction, transcription could also be detected in adult testis and pancreas. All RT-PCR products were sequenced and compared with the genomic sequence to identify the exon–intron boundaries (Fig. 1). Finally, we used 5′ and 3′ RACE, northern blotting and *in situ* hybridization on mouse and human embryo sections to characterize the transcription and expression profile of *WHSC1*.

Characterization of *WHSC1*

The entire *WHSC1* gene measures 90 kb, is transcribed from telomere to centromere and extends for 60 kb into the telomeric end of the 165 kb critical region. It contains 25 exons, together encoding an 8 kb cDNA. The ATG translation start codon lies in exon 4 preceded by a 5′-untranslated region (UTR) of >400 bp (Fig. 1, Table 1). The exons vary between 82 and 3565 bp, the introns between 132 and 13 718 bp. Exon 14 does not fulfil the GT rule at the 5′-splice donor site. It contains a GC dinucleotide, a donor site which occasionally has been found in other genes, e.g. in exon 30 of the Duchenne muscular dystrophy gene (17). The splice donor site of exon 14 is functional, since RT-PCR covering this exon consistently yields unique products including this exon (Fig. 2a, lane 4).

The most 5′ dbEST match of *WHSC1* is EST 27266 (Table 2). Using a forward primer designed on more upstream EST matches (k3378, k3397 and j3435 in cl75b9a) in combination with several reverse primers located in exons 4 (Fig. 1, W and M) and 6 (Fig. 1, K), we have not been able to extend *WHSC1* further in the 5′ direction. Some amplification products were obtained, but their sequence showed that they were artefactual products, not derived from 4p16.3. Similarly, 5′ RACE using human fetal brain and adult testis cDNA (Marathon Ready kits; Clontech) with a range of primers spread throughout *WHSC1* failed to extend the cDNA sequence upstream of exon 1. Only after nested PCR were some PCR products obtained, but these where unspliced genomic products, not extending *WHSC1* in the 5′ direction.

Alternative splicing was found for exons 2 and 3 (Figs 1 and 2c). RT-PCR between exons 1 and 4 (primers L/W) produces a major 280 bp product on blood, pancreas, testis and heart RNA, and a 400 bp product on brain material (Fig. 2c, lanes 12 and 13). The sequence of the 400 bp product contains exons 1–4, while the 280 bp product lacks exons 2 and 3. The EST database contains one clone, EST 27266, which only lacks exon 2 (Table 2). Thus, at least three different splice forms of *WHSC1* exist in this region.

WHSC1 shows dbEST matches for the region spanning exons 1–12a and 20–25. Exons 13–19 are not represented in dbEST (Table 2). The exon 1–12a matches are with sequences derived mostly from random-primed cDNA libraries of human, murine and rat origin, and they do not form a single contig. Another group of EST matches is found with genomic sequences starting and/or ending in intronic sequences, often flanking oligo(dT) or oligo(dA) stretches. Some of these are derived from the opposite transcriptional strand (Table 2). These ESTs probably represent unspliced RNAs or DNA-primed products.

Transcription in the exon 10–13 region is very complex. Some transcripts contain either one of two alternatively spliced exons 12, while others terminate here using either of two alternative poly(A) addition sites. Exon 12, identified through EST zr01a04, uses two splice donor sites yielding exons of 94 and 227 bp respectively (12a and 12b, Fig. 1, Table 1). Expression of exon 12 is only detectable using a primer located in this exon (Fig. 2b, lanes 9–11). In contrast, PCR across exon 12, e.g. from exon 7 to 14 [V/F (Fig. 2a, lane 4)], yields only products without this exon.

An alternatively spliced exon was found by sequence analysis of cDNA clones HFBEP10 and zv63h03 which terminate in ‘intron 11’. HFBEP10 contains exons 7 (in part), 8, 9 and 10 spliced to 11b, and has a total size of 3613 bp. Another clone, EST zv63h03, ends in a poly(A) tail 646 bp downstream of exon 11, preceded 37 bp upstream by a perfect AATAAA poly(A) addition signal (Fig. 1, Tables 1 and 2). This polyadenylation site could be confirmed by 3′ RACE PCR on heart and brain mRNA. To find out whether the ‘intronic’ exon 11b is evolutionary conserved, a Southern blot with DNA from a range of species was hybridized with HFBEP10. Human DNA shows two hybridizing bands of 4.8 kb (prominent) for exon 11b and 3.8 kb (weak) for the exon 7 segment (345 bp). Cross-hybridization was detected in cow, dog, mouse, rat and primate (Fig. 3).

Table 1. Sequences and positions of exon–intron boundaries in *WHSC1*

Exon no.	Intron (3′)	Exon (5′)	Exon (3′)	Intron (5′)	Size (bp)	Cosmid	Position
1			TAACTG	gtaatt	(≥40)	cl184d6	143 239–14 200
2	ttgcag	AGACAA	CACGAG	gtgggc	180	cl184d6	14 938–14 759
3	cctcag	CTGTCT	TCAGAG	gtcagg	175	cl184d6	7888–7714
4	ccatag	TGTTCT	AAAAAG	gtattt	626	cl184d6	6485–5858
5	tttcag	ATTCCA	TTAAAG	gtattg	163	cl184d6	2893–2731
6	aactag	GTCAGA	ATTAAG	gtgata	167	cl190b4	19 885–19 719
7	tcccag	CTATTG	GATGAG	gtcagt	483	cl190b4	18 615–18 133
8	ccacag	GTGGTA	ACTCTG	gtaaac	145	cl190b4	6129–5985
9	tccaag	GTAATG	CGGAAG	gtaatt	119	cl190b4	1610–1492
10	taatag	AGAGAC	AGGCAG	gtaatg	82	cl19h1	38 351–38 270
11	ttacag	CAACGA	AATGAG	gtaaaa	125	cl19h1	37 148–37 024
11a	ttacag	CAACGA			227	cl19h1	37 148–36 922
11b	ttacag	CAACGA			2922	cl19h1	37 148–34 227
12a	taacag	CTTTTG	CTGCAG	gtggcg	94	cl19h1	34 463–34 370
12b	taacag	CTTTTG	ACCAAG	gtaaga	227	cl19h1	34 463–34 237
13	ctctag	GTCTCG	TGCCAG	gtgagg	132	cl19h1	25 731–25 600
14	ccgcag	CTGTGT	CCTCAG	gcaagt	124	cl19h1	24 719–24 596
15	tgttag	GGATTC	CAAAAG	gtacag	200	cl19h1	23 502–23 303
16	cttcag	GTAAAA	CCAAAG	gtgagg	180	cl19h1	21 666–21 487
17	ttgcag	GGGGGA	CTACAG	gtgtga	157	cl19h1	21 134–20 978
18	taatag	ATGGTG	AAAACG	gtacgg	206	cl19h1	20 844–20 639
19	ttttag	CACTGC	ATCAAG	gtggcg	104	cl19h1	18 894–18 791
20	ctgcag	GTGAAT	AGAAAG	gtatgt	270	cl19h1	17 356–17 087
21	ttccag	GGAGAA	GACAAG	gtaatg	117	cl19h1	15 792–15 676
22	ttacag	GACCGT	CTGCAG	gtacaa	142	cl19h1	1956–1815
23	tcccag	GGACGG	CCAAAG	gtaagg	107	cl19h1	1525–1419
24	ctctag	ACCTCG	CCTTCG	gtgggt	205	cl19h1	344–140
25	ttgcag	GGAAGT			1159	cl96a2	33 362–32 204
25a	ttgcag	GGAAGT			3565	cl96a2	33 362–29 797

For every exon–intron boundary, six nucleotides each are given.

Also at its extreme 3′ end, *WHSC1* appears to utilize two alternative poly(A) addition sites (Table 2), represented by two EST contigs. The first EST contig (Hs.110457) spans exons 20–25, and ends at a polyadenylation site located at bp 32 204 in cl96a2. This site, not preceded by an obvious consensus poly(A) addition signal, is found in five human and five murine ESTs (Table 2). The second EST contig [Hs.19416, containing transcript 194164 of Wright *et al.* (13)] ends 2.4 kb further downstream, at bp 29 796 in cl96a2, and is preceded by an AATAAA poly(A) signal 31 bp upstream. This contig contains 22 human and three murine transcripts and spans 2.3 kb. It is not interrupted by introns, has no coding potential, detects no similarities in the various databases and misses overlapping with the first contig by 684 bp. Since EST contigs rarely span >2 kb of cDNA, it is not surprising that no clone connecting the two contigs was present. Only 500 bp 3′ of *WHSC1*, at bp 29 277 of cl96a2, maps the 3′ end of another gene, transcribed from the opposite strand [EST contig Hs.21771, containing transcripts 53282 and 267784 of Wright *et al.* (13)].

Between exons 4 and 25, *WHSC1* contains a 4095 bp open reading frame (ORF) encoding a putative protein of 136 kDa followed by a 3′ UTR of 3294 bp. Alternative splicing of exons 2 and 3 does not result in additional translation products. Transcripts using the alternative poly(A) addition sites in exon 11a and 11b (Tables 1 and 2) would contain a translational stop at cDNA bp 1887 (6 bp of ‘intron 11’), yielding a 62 kDa protein, transcripts containing exon 12 would end at bp 1941 (60 bp of exon 12), encoding a 64 kDa protein with 20 novel amino acids.

Table 2. EST matches in *WHSC1*

Exons	ESTs	Remarks
1, 3, 4	EST27266	
4	HBMSF1D9	
4–11	zv63h03 (758357)	internal deletion exon 4–7, 3′ poly(A) 646 bp downstream of exon 11
5–7	EST111365	rat (reverse)
i5	ys11f05	27616–26811, ends in genomic dA stretch
6–7	45h5 (W28349)	starts in intron 5, ends in intron 7
6–7	zr01a04	unspliced, 5′ end in intron 5, 3′ in intron 7 dA stretch
7	EST05750	HMG homology?
7–8	EST28510	
7–9	ztd05d09.r1	
i7	yv67d10	intronic, ends in genomic dA stretch
9–12a	mw14h01.r1	murine, exon 2561 bp downstream in intron 11
10	zs25f05	antisense, primed at intron 9 dT and intron 10 dA stretch
11–12a	HUMEST6H1	
20–25	zo77g11, zm24f01, zm13b05, zo78d05, zo65g02.r1, zf12c11	ending in poly(A) tail at 32 205, 31 bp 5′ preceded by AATATA
20–22	mh09c05.r1	murine
22–23	ml55h10.r1	murine
25	mg09f12.r1, mi51h05.r1	murine
25	zs47g04.s1, aa61g04.s1	3′ end antisense, 5′ end not on chromosome 4[a]
25a	EST17737, EST36779, EST60207, EST64437, EST70993, nf65b08.s, nf68c05.s, nh82c06.s, nk39a02.s, nn40a11.s, zd48f12.s,zn34b11, ye76b09, yg46e03.r, ym59a12.r, yp61c04.r, yp84g03, yr11g06, yr13c09.r, yr25c04, yr72b04.r, yr76b04	covers 32 203–29 796: ending in poly(A) tail, 31 bp 5′p receded by ATTAAA

[a]On the opposite strand, bp 32530–33021 of cl96a2, exon 25 has an exact match with the 3′ end of ESTs zs47g04 and aa61g04. These ESTs, part of a large EST contig, seem not to originate from chromosome 4 since their 5′ sequences show no homology with any of the 2 Mb chromosome 4 sequences known from this region.

Northern blot analysis

Northern blot analysis with different RT-PCR probes dispersed over *WHSC1* using human adult and fetal multiple tissue blots (Clontech) revealed a complex transcription pattern (Fig. 4). Main transcripts of 9 and 6 kb were recognized with all probes and in most tissues. The size of the 9 kb transcript indicates that the 8 kb *WHSC1* cDNA sequence described here misses some 0.5–1 kb of untranslated sequence, probably 5′. Similarly, the transcripts initiated from the same promoter but ending in 'intron 11' (i.e. using exons 11a or 11b) would have sizes of 5.2 and 2.6 kb respectively. Using the exon 1–4 probe (L/M, 500 bp product derived from blood RNA), transcripts of ~2.5 kb were detected in all tissues, whereas the 9 kb transcript was observed only in heart.

Expression is most diverse in fetal tissues (Fig. 4b). At least five transcripts were detected, ranging in size from 3.7 to >10 kb using an exon 7–11b cDNA probe. Fetal brain contained the most complex pattern, with transcripts of 3.7, 4.5, 9, 10 kb and a unique >10 kb transcript. Adult brain transcripts appear to be a subset of those in fetal brain. The major 9 kb transcript was detected in all fetal and adult tissues examined, the 6 kb transcript in fetal liver and kidney only (Fig. 4c).

Hybridization of the exon 20–25 probe (1F/1R, 1300 bp product derived from testis RNA) uniquely revealed a 3.5 kb RNA in all tissues examined (Fig. 4a). Expression of all transcripts was very prominent in skeletal muscle and heart and clearly detectable in tissues of the inner organs such as kidney, lung, pancreas and liver.

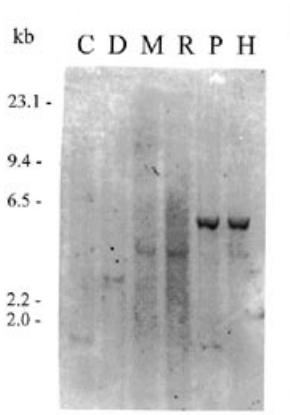

Figure 3. Zoo blot analysis with HFBEP10. Zoo blot containing 4 µg of DNA digested with *Eco*RI. The filter is hybridized non-radioactively with the 3.6 kb HFBEP10 cDNA clone, the signal was detected using an anti-fluorescein alkaline phosphatase conjugate. C, cow; D, dog; M, mouse; R, rat; P, primate; H, human.

To exclude that transcripts were derived from other, homologous genes, a chromosomal hybrid mapping panel was hybridized with the exon 7–11b probe. The results show a 100% concordance, with a unique localization on chromosome 4.

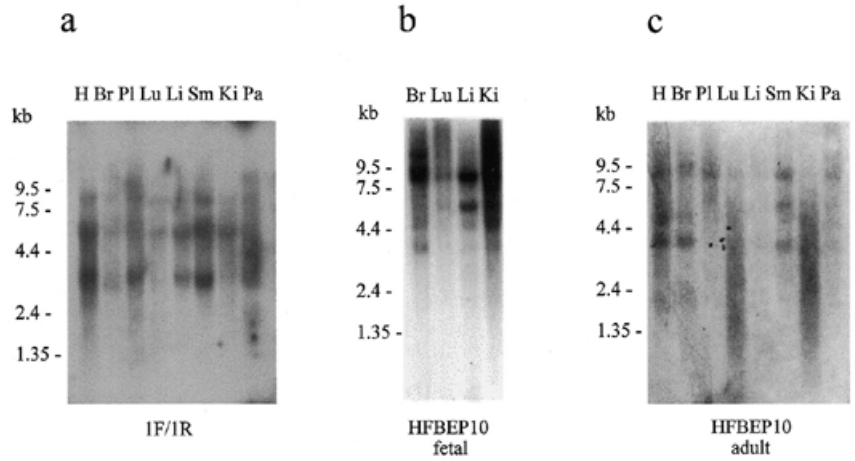

Figure 4. Expression of *WHSC1*: northern blot analysis. Northern blots containing 2 µg of poly(A)$^+$ RNA from various human adult or fetal tissues were hybridized with different RT-PCR products and the cDNA probe indicated. (**a**) Adult tissue blot hybridized with RT-PCR probe 1F/1R spanning exons 20–25. (**b**) Fetal tissue blot hybridized with cDNA clone HFPEB10 spanning exons 7–11b. (**c**) Adult tissue blot hybridized with cDNA clone HFPEB10 spanning exons 7–11b. Br, brain; H, heart; Ki, kidney; Li, liver; Lu, lung; Pa, pancreas; Pl, placenta; Sm, skeletal muscle.

Furthermore, we can exclude the existence of homologous genes elsewhere on chromosome 4 since only the known *WHSC1*-hybridizing fragments were detected (Fig. 3, lane H).

RNA *in situ* hybridization on embryonic mouse sections

To analyse the developmental pattern of expression of *WHSC1*, we performed *in situ* hybridization on mouse embryo sections of different developmental stages (E10.5, 12.5, 13.5 and 16.5 p.c.) and human embryos (developmental days 52–56). We used ^{35}S-labelled antisense and sense RNA probes covering different regions of *WHSC1*. Identical expression patterns were obtained using a human probe containing exons 1, 4, 5 and 6 (L/K, 850 bp), murine exon 9–12a probe (H/S, 280 bp) and murine exon 20–25a probe (mF/mR, 1 kb) on mouse sections (Figs 1 and 5). Hybridization of the human probe L/K to human embryonic sections displayed an essentially similar but weaker signal. The negative control with human probe L/K to human sections yielded a weak signal in the outflow region of the heart.

Figure 5a1 shows the expression pattern using the murine exon 9–12a RNA probe at developmental day 13.5. It strongly hybridizes specifically to brain, ganglia and neural tube, to the anlage region of the jaw, to the frontal face region including the developing upper and lower lip, to intestinal and lung epithelium, to liver and to the adrenals and the urogenital system. Figure 5c–e shows detailed pictures of the expression pattern in brain (c), ganglia, adrenal and epithelium of the intestine (d), and in the liver and the lung epithelium (e). At day E10.5, expression is highly specific throughout the developing nervous system (Fig. 5b). The negative (sense) RNA controls showed no signal with any probe–section combination (Fig. 5a2). Some tissues, such as the liver, the dorsal ganglia and the fifth ganglion in the brain, show relatively even expression, while other tissues, e.g. epithelium of the gut and lung, show a marked inward–outward gradient, suggesting that rapidly growing layers in these tissues have a higher expression than their surroundings.

DISCUSSION

WHS is typically caused by large deletions of the distal short arm of one of the chromosomes 4. The critical region previously has been confined to 165 kb, and nine potential transcriptional units have been described in this interval (13). Several genes may jointly contribute to the WHS phenotype. We have characterized a novel gene (*WHSC1*) covering 60 kb of the 165 kb critical region and merging two of the transcription units reported recently in the WHSCR, HFBEP10 and 194164. Its location, the nature of the motifs in the encoded protein, and the expression profile in human and mouse embryos imply that hemizygous deletions of this gene might be responsible for many of the physical and neurological features in WHS.

Our results indicate that the expression of the *WHSC1* gene is complex, showing, apart from the major consistent transcripts of 9 and 6 kb, many different transcripts in different tissues. Alternative splicing, both between and within tissues, affects exons 2, 3, 11 and 12. Since the length of exons 2, 3 and 12 is rather small, 100–230 bp, it is unlikely that their alternative splicing yields transcripts of lengths which can be discriminated on northern blots. However, alternative polyadenylation of exons 11a and 11b should yield transcripts differing some 2.2 kb in size. Similarly, the two polyadenylation sites used for exon 25 should yield differently sized transcripts depending on the respective poly(A) addition site used. Finally, based on the complex expression patterns observed on fetal and adult northern blots, *WHSC1* expression is most likely also driven by more than one promoter.

The *WHSC1*-encoded protein reveals homologies to members of several protein families. This suggests that *WHSC1* product is a DNA-binding protein, most likely a transcription factor or co-regulator. At amino acid position 211–295, the *WHSC1*-encoded protein contains a highly conserved region which we have designated the PWWP domain (for the proline–tryptophan–tryptophan–proline motif in the consensus amino acid sequence) (see Fig. 6). This domain has been noted but not described (ProDom database, ID14260). It is also present in the G/T DNA mismatch repair gene *MSH6* product causing non-polyposis

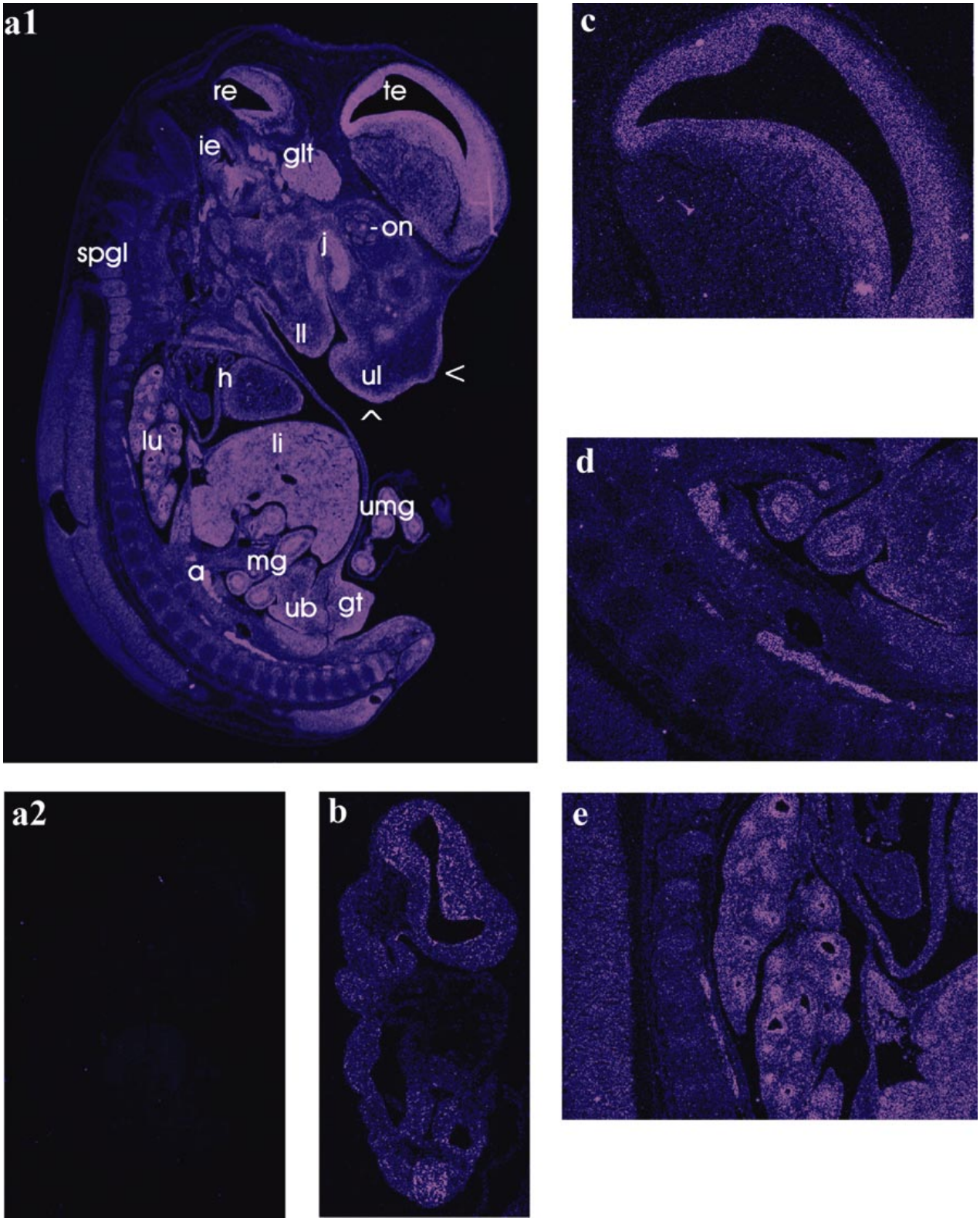

Figure 5. RNA *in situ* hybridization on sagittal mouse embryonic sections. (**a**) Hybridization of murine antisense (a1) and sense (a2) RNA probes generated with primers H/S (exons 9–12a, Fig. 1) to mouse sections of developmental day 13.5 p.c. Expression is very prominent in neuronal cells as in the brain, in the fifth ganglion (ganglion trigeminale) and optical nerve. Enhanced expression is also detected in the margin of the frontal facial region, the region forming the lip, the region of the developing jaw, the epithelium of lung and gut, the liver, the genitourinary system and the adrenals. (**b**) Hybridization of a murine antisense RNA probe generated with primers mF/mR (exons 21–25) to mouse sections of developmental day 10.5 p.c. It demonstrates specific expression in brain and the neural tube. In detail: specific expression in the brain (**c**) and epithelium of the intestine and lung (**d** and **e**). a, adrenal; ie, inner ear; glt, ganglion trigeminale; gt, genital tubercle; h, heart; j, jaw; lu, lung; li, liver; ll, lower lip; lu, lung; mg, midgut; on, optical nerve; re, rhombencephalon; spgl, spinal ganglia; te, telencephalon; ub, urinary bladder; ul, upper lip; umg, umbilical midgut; arrows indicate the developing region of the frontal face.

colorectal cancer (HNPCC) (18), in the hepatoma-derived growth factor (HDGF) (19) and in a 17q-derived EST expressed in lymphocytes (20).

Amino acids 459–503 show homology with the HMG box 2. HMG proteins are grouped in three families, HMG-1/-2, HMG-14/-17 and HMG-I(Y). HMG boxes facilitate DNA interactions and have been found in many proteins including RNA polymerase I transcription factor UBF (upstream binding factor), the mammalian testis-determining factor SRY and two mitochondrial transcription factors (21). HMG boxes are DNA-binding domains, causing a 130° bend of the double helix. HMG proteins are positively correlated to cell proliferation. High homology (36% identity, 52% similarity) is observed between the *WHSC1*-encoded protein and the HMG2 box in BF1_human nucleolar transcription factor 1 (UBF-1, autoantigen NOR-90) (22), TETPY (*Tetrahymena pyriformis*) HMG non-histone chromosomal protein (23) and a gene product of *Caenorhabditis elegans*, C26C6.1 (24).

The PHD-type zinc finger (plant-homeodomain) is located at position 833–875 of the WHSC1 protein. This domain has been found in a series of proteins of different species: human ATRX protein (alpha-thalassaemia mental retardation syndrome, X-linked, with genital abnormalities and facial dysmorphism), human FAC-fetal protein (fetal Alz-50-reactive clone 1 (25), human KRIP-1(KRAB-A interacting protein; KRAB-A: Krüppel-associated box A) (26), human Mi-2 immune autoantigen, a nuclear protein (27), AIRE (autoimmune regulator), a recently cloned gene responsible for APECED syndrome (28,29), human TIF1 (transcriptional intermediary factor 1) and *C.elegans* gene C44B9.4. ATRX is thought to influence gene expression by affecting chromatin (30). KRIP-1 is a member of a group of proteins which are known to play important roles in differentiation, oncogenesis and signal transduction. It probably interacts, like TIF1, with proteins essential for transcription regulation. The function of the TIF proteins appears to be as a mediator of ligand-dependent activation which acts on the basal transcription machinery (31).

The SET domain (16) at position 1074–1180 consists of ~140 amino acids and has been identified in a large range of transcriptional/developmental proteins from diverse organisms (15): human ALL1 (acute lymphocytic leukaemia) protein (32), in ORFKG1T, a hypothetical human protein (24), in three ASH1-related *Drosophila* proteins TRX (Trithorax) (33), SUVAR 9 (suppressor of variegation 9) and enhancer of zeste/polycomb [E(z)] (16,34,35), in HRX (human homologue of TRX, human and murine) (36,37), in two predicted *C.elegans* genes (24) and two yeast ORFs (38,39). Highest homology (44% identity, 64% similarity) of the WHSC1 SET domain occurs with ASH1 of *Drosophila melanogaster*. Mutations in the *ash1* gene cause misdevelopment in a number of organs in *Drosophila*, e.g. transformation of the halters to wings, genitalia to legs, first leg to second leg, etc. (40). Nilsow *et al*. (41) created a null mutant in the fission yeast *SET1* gene displaying morphological abnormalities, cell growth and sporulation defects. In *Drosophila*, the '*enhancer of zeste*' gene has a negative regulator function of segment identity genes of the Antennapedia and Bithorax complexes (34,35). The SET domain genes are a highly conserved family encoding proteins which influence transcription by changing chromatin-mediated regulating mechanisms leading to secondary effects on developmental programmes. In combination with the expression profile of *WHSC1*, the protein domain homologies, in particular with the SET domain, strongly suggest that *WHSC1* encodes a protein which plays a substantial role in transcription regulation of genes involved in embryonic development.

We have collected a set of seven patients with many features characteristic for WHS or PRDS. To date, we have not been able to detect any gross deficiencies nor microdeletions or point mutations focused on the domains recognized in the protein sequence. A silent third base change from C to T was found in one of the patients in exon 22. Since most WHS/PRDS patients carry extensive 4p16 deletions and lack this gene completely, the question arises if small mutations in this single gene would already yield the characteristic WHS phenotypical spectrum or whether the deletion of more genes is required.

Apart from the facial hallmarks, WHS patients suffer from multiple malformations which arise during the first trimester of development. These malformations affect the CNS (microcephaly, hypotonia, seizures), eyes (widely spaced, round), ears (low set), mouth (cleft lip/palate, fused teeth), extremities (overlapping toes), genitourinary system (hypospadia, cryptorchism) and heart. The results of the *in situ* hybridization on murine embryonic sections demonstrate expression in many of these tissues as early as day 10.5. The temporal and spatial multi-organic expression pattern provides independent support for the involvement of *WHSC1* in the development of the WHS phenotype, besides its location. The brain, jaw and genitourinary system show gene expression corresponding to affected organs in WHS. For example, brain anomalies cause developmental delay, while midline fusion defects of the developing jaw and genitourinary region may lead to cleft lip/palate and hypospadia. Also the distinct inward–outward gradients observed in several tissues and the overall abundance in rapidly growing, not yet differentiated structures suggest that *WHSC1* may well be involved in determining as yet unknown morphogenetic gradients, as are many other transcription factors. Typically, the actions and effects of such proteins are highly dosage sensitive, which would be consistent with a haploinsufficiency model for the protein(s) affected in WHS. The complex pattern of tissue-specific transcription initiation, splicing and polyadenylation would present a significant potential source of further inter-individual genetic diversity at the transcript and protein level. Indeed, this variability would present an attractive mechanistic explanation for the generally observed phenotypic variation in haploinsufficiency (deletion) syndromes. This would link the natural genetic diversity of the expression levels of genes on the non-affected chromosome to clinical variation caused by deletion on the homologous chromosome.

Finally, an apparently unrelated recent finding is well worth noting. Patients with multiple myeloma (MM) suffer from tumours which destroy bone marrow structures especially in the skull, clavicles, sternum and vertebrae. Recent reports describe t(4;14)(p16.3;q32.3) translocations in a significant fraction of primary tumours and tumour-derived cell lines of MM patients located <100 kb centromeric of the *FGFR3* gene in 4p16.3. Several of these map in cosmids cl75b9, cl184d6 and cl190b4 (42,43). As deduced from the cloned and sequenced breakpoints, we conclude that three of the breakpoints disrupt the *WHSC1* gene: two in intron 4 (i.e. cell lines NCI-H929 at bp 4127 and JIM3 at bp 3491 in cosmid cl184d6) and one in intron 5 (cell line OPM2 at bp 23 665 in cosmid cl190b4). In these cases, transcripts are expected which fuse either the 5′ region of the *IgH* locus (14q32, switch region of the heavy chain gene) with nearly the entire *WHSC1* gene [der(4)] or the 5′ *WHSC1* gene region

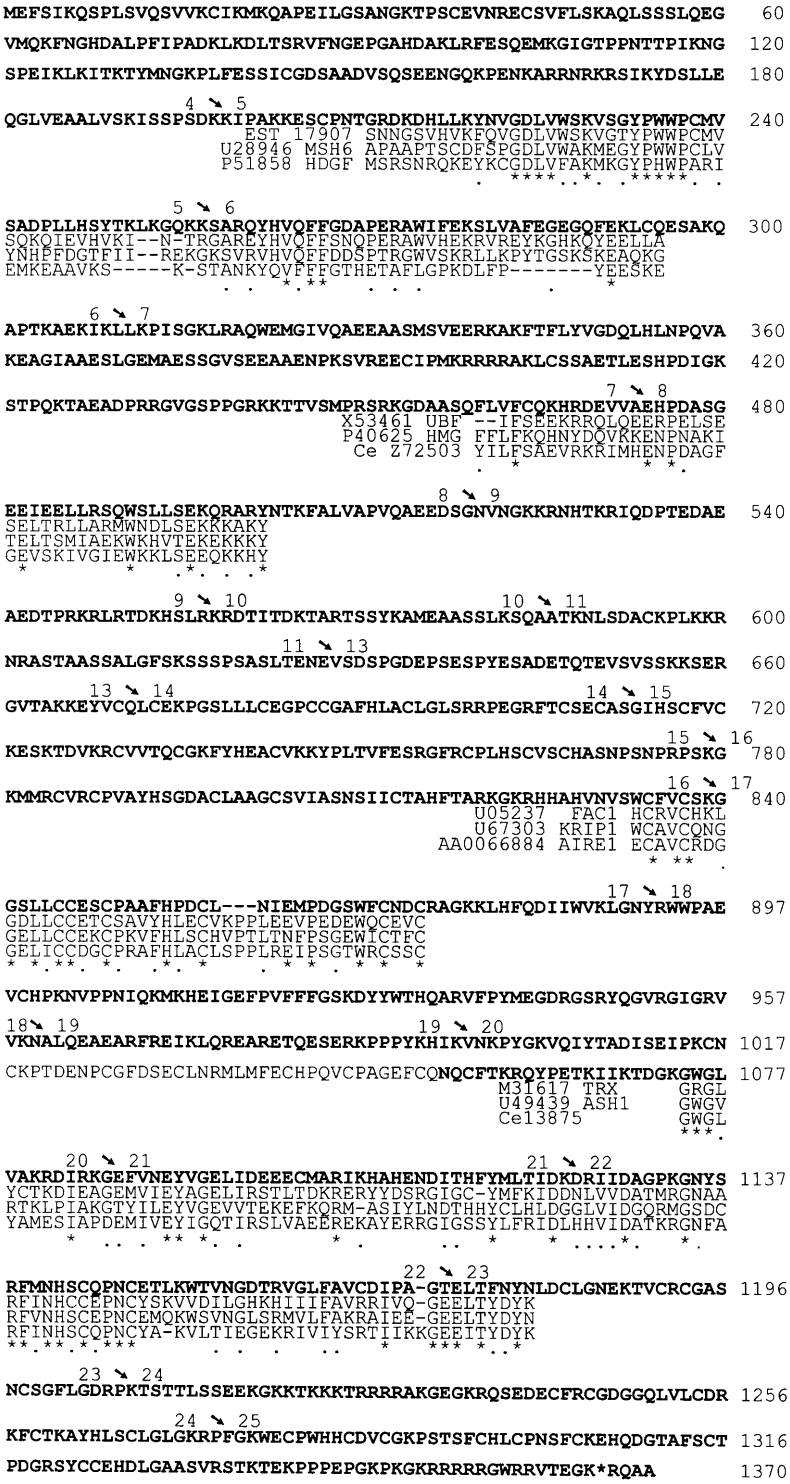

Figure 6. *WHSC1*-encoded amino acid sequence, starting with methionine at position 1, and alignment. Evolutionarily conserved regions are indicated by asterisks, evolutionary conservation for three amino acids and amino acids which are functionally equivalent are indicated by dots. Exon–exon boundaries are given by arrows. Abbreviations are explained elsewhere. Domains of the following proteins are shown. PWWP domain: WHSC1, position 211–295; U28946 (MSH6), position 81–162; P51858 (human HDGF), position 1–78; U17907 EST 17q, position 96–180; HMG box: WHSC1, position 459–503; X53461 (human UBF-1), position 416–456; P40625 (HMG, *T.pyriformis*, SGC5), position 18–60; Z72503 (*C.elegans* cosmid C26C6.1), position 1196–1238. PHD finger: WHSC1, position 833–873; U05237 (FAC1), position 239–297; U67303 (KRIP-1), position 627–669; AB006684 (AIRE1), position 298–340. SET domain: WHSC1, position 1074–1180; M31617 (TRX), position 3632–3759; U49439 (ASH1), position 1385–1490; U13875 (*C.elegans* cosmid C26E6.9), position 1674–1780.

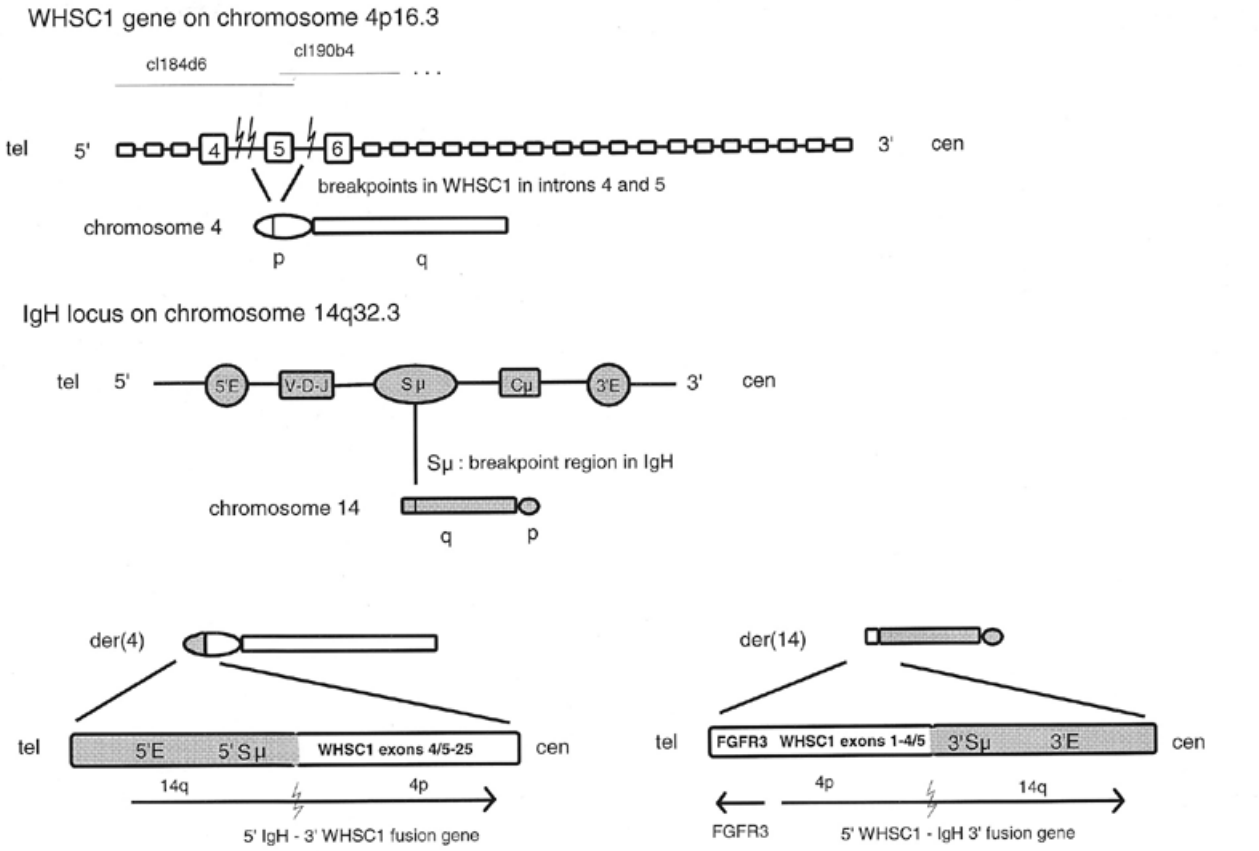

Figure 7. (4;14) (p16.3;q32.3) translocation in multiple myelomas. Top: *WHSC1* gene on chromosome 4p. Middle: *IgH* locus on chromosome 14q. Bottom: at the left derivative chromosome 4; at the right derivative chromosome 14. 3′ E, 3′ enhancer; 5′ E, 5′ enhancer; V-D-J, variable region, diversity genes, joining segments; Cμ, isoform of the constant region; Sμ, switch region on the *IgH* locus.

with the 3′ *IgH* locus [der(14)] (Fig. 7). Depending on the translational consequences, these rearrangements are capable of generating fusion proteins. Most of the other breakpoints lie 10–30 kb upstream, probably disrupting an as yet undetected 5′ end of *WHSC1* and/or its transcription regulatory region. Up-regulation of the *FGFR3* gene, present 70 kb distally on the der(14), was observed in seven out of the eight (4;14) MM translocations. This led to the proposal that the *FGFR3* gene should be a candidate oncogene, controlled by the 3′ Cμ enhancer. Our data indicate that fusion of *IgH* and *WHSC1* genes and their untimely expression in the myeloid lineage driven from the 5′ IgH enhancer might be another event with similar profound consequences, e.g. involvement of *WHSC1*-encoded proteins in the clinical heterogeneity of MM.

MATERIALS AND METHODS

RT-PCR

Based on database homologies and computer prediction, we designed >40 primers spread over *WHSC1*. RNA of adult brain, blood, heart, testis, spleen and/or kidney was used for first strand cDNA synthesis by incubating 1–3 μg of total RNA and 1 μg of random primer at 65°C for 10 min. MMLV reverse transcriptase, 1× MMLV reverse transcriptase buffer, 10 mM dithiothreitol (DTT), 1 mM dNTPs (Pharmacia) and 40 U of RNAsin (Promega) were added and incubated at 42°C for 1 h.

For PCR, we used 2 μl of cDNA in an PCR end volume of 16.6 μl, 1× Amplitaq buffer, 200 μM dNTPs, 1.3 mM $MgCl_2$ and 0.7 U of Amplitaq *Taq* polymerase. Briefly, 30 cycles each of 30 s at 94°C, 400 s at an annealing temperature between 53 and 60°C and an extension time depending on the size of the expected PCR fragment between 30 s and 3 min were carried out. First delay at 94°C for 1 min, last delay at 72°C for 5 min was used. RACE PCR with Marathon ready cDNA (Clontech) was carried out according to the protocol of Lung *et al*. (44).

The following primers were used (F, f: forward; R, r: reverse, given in the 5′ to 3′ direction. T_m: annealing temperature; ISH: *in situ* hybridization): LF, CCTAGAACCACTGGTAACTC, T_m 51.3°C; LF(ISH), [T3f]-ACCTAGAACCACTGGTAACTCTT; WR, CATCCAGCCCAGATGCTTCCGTTC, T_m 71.8°C; MR, CAGCAGCACTGTCACCACAAATGG, T_m 69°C; KR, CTGTACGTGATACTGGCGTGCACTC, T_m 67.1°C; K(ISH), [T7r]-CTGTACGTGATACTGGCGTGCACTC; VF, [M13f]-ATGCAGCATCCCAGTTTTTG, T_m 61°C; HF, GGACGGACAAGCACAGTCTT, T_m 60.3°C; H(ISH), [T7r]-AGRRCRGAYAAGCACAGTCTT; SF(ISH), [T3f]-ACTTGACTGRTGTGGGCTCC; SF, [M13f]-ACTTGACTGRTGTGGGCTCC; SR, [M13r]-ACTTGACTGGTGTGGGCTCC, T_m 62.5°C; FR, [M13r]-CGCTGCAGGTGAACCTCCCTTCTGG, T_m 70°C; BR, GAAGCACCAGCTCACGTTGA, T_m 60°C; AF, AACAGCATCATCTGCACTGC, T_m 60°C; 3R, TCCTGTTCAGA-

CACTCCGAA, T_m 60°C; 1F,CAGATGAGAATCCTTGTGGC, T_m 60°C; 1R, AAATTCAAGTGTGGCGGTAA, T_m 60°C; mF, GAGTGTCTGAACAGGA, T_m: 60 ºC; mR, GTCATGCTCA-CAGCAGTAG, T_m 60°C; GF, GGCAGCTGGTGCTGTGTG-AC, T_m 60°C.

Primer extensions: M13f, GACGTTGTAAAACGACGGC-CAGT; M13r, CAGGAAACAGCTATGACCATGA; T3f, CA-CAATTAACCCTCACTAAAGGG; T7r, CACTAATACGACT-CACTATAGGG.

RNA *in situ* hybridization

RNA probes were labelled as described by Moorman *et al*. (45). Fixation of freshly isolated embryos was done overnight at 4°C in 4% (para) formaldehyde in phosphate-buffered saline (PBS) with agitation. The embryos were embedded in paraffin to prepare serial 7 µm sections. Hybridization was done with ^{35}S-labelled *WHSC1* cRNAs as described in detail previously (45,46). Image analysis software (Adobe Photoshop) was used to change the black–white picture into pink–blue.

Zoo blot analysis

Zoo blots containing DNA digested with *Eco*RI from a number of species (human, primate, rat, mouse, dog and cow) were purchased from Bios. The exon 7–11b probe was labelled non-radioactively using fluorescein-11-dUTP (Vistra; Molecular Dynamics/Amersham). Following hybridization, washing and detection using an anti-fluorescein alkaline phosphatase conjugate, the blot was analysed using a Storm 840 (Molecular Dynamics).

Northern blotting

Northern blots containing poly(A)$^+$ RNA from adult and fetal human tissues were purchased from Clontech (fetal blot: no. 7765-1, adult blots: nos 7760-1 and 7759-1). Northern blots were hybridized with labelled RT-PCR probes and DNA from cDNA clone HFBEP10 spanning different exonic regions of *WHSC1*. Blots were washed at 0.1–0.3× SSC, 0.05–0.1% SDS at 65°C and exposed for 1–5 days at –70°C.

Database searches/exon prediction programs

The BLAST algorithm (47) was used to search for homologies of nucleic acids and protein sequences. Exon prediction was performed using the X-Grail program (48).

Accession numbers (Genbank/Swissprot/EMBL)

Cell line NCI-H929, U73662; cell line JIM3, U73660; cell line OPM2, AF006657; G/T mismatch-binding protein, U28946; HDGF, P51858; 17q-derived EST: U17907-1; human ATRX, U72937; human FAC-fetal protein, U05237; human KRIP-1, U67303; human Mi-2 immune autoantigen, X86691; AIRE1, AA006684; human TIF1, S78219 and S78221; gene product C44B9.4 of *C.elegans*, Z73424; human ALL1, Q03164; ORFKG1T, D31891; TRX, M31617; SUVAR 9, EMBLX80070, P45975; [E(z)], P42124; human HRX, L04284; murine HRX, L17069; two predicted *C.elegans* genes: U13875 and U37430; two yeast ORFs, P38827 and P46995; ASH1 protein, U49439; human BF_1 UBF-1, autoantigen NOR, X53461; HMG_TETPY HMG protein, P40625; and gene product C26C6.1 of *C.elegans*, Z72503. Further numbers refer to Figure 6.

ACKNOWLEDGEMENTS

We wish to thank Dian Donnai, Marie Croquette and Annick Toutain for their readiness and efforts to provide us with patient material. We are especially grateful for the willingness of all parents and patients to cooperate in this study. We further wish to thank Cees Hersbach and Jaco Hagoort for their expert technical support in scanning of mouse embryo sections after hybridization, Johan van Triest for his expert photographical assistence, Tom Peat for his valuable comments during gene analysis, and the Genome Technology Center (University Leiden) for help in sequencing and Ilo. I.S. is supported by the Deutsche Forschungsgemeinschaft (DFG) (Ste802/1-1). T.J.W. and M.R.A. are supported by the US Department of Energy under contract no. W-7405-ENG-36 and the Los Alamos National Laboratory Directed Research and Development Fund.

REFERENCES

1. Wolf, U., Reinwein, H., Porsch, R., Schroter, R. and Baitsch, H. (1965) Defizienz an den kurzen Armen eines Chromosoms nr. 4. *Humangenetik*, **1**, 397–413.
2. Hirschhorn, K., Cooper, H.L. and Firschein, I.L. (1965) Deletion of short arms of chromosome 4–5 in a child with defects of midline fusion. *Humangenetik*, **1**, 479–482.
3. Estabrooks, L.L., Rao, K.W., Driscoll, D.A., Crandall, B.F., Dean, J.C.S., Ikonen, E., Korf, B. and Aylsworth, A.S. (1995) Preliminary phenotypic map of chromosome 4p16 based on 4p deletions. *Am. J. Med. Genet.*, **57**, 581–586.
4. Wilson, M.G., Towner, J.W., Coffin, G.S., Ebbin, A.J., Siris, E. and Brager, P. (1981) Genetic and clinical studies in 13 patients with the Wolf–Hirschhorn syndrome. *Hum. Genet.*, **59**, 297–307.
5. Pitt, D.B., Rogers, J.G. and Danks, D.M. (1984) Mental retardation, unusual face, and intrauterine growth retardation: a new recessive syndrome? *Am. J. Med. Genet.*, **19**, 307–313.
6. Clemens, M., Martsolf, J.T., Rogers, J.G., Mowery-Rushton, P., Surti, U. and McPherson, E. (1996) Pitt–Rogers–Danks syndrome: the result of a 4p microdeletion. *Am. J. Med. Genet.*, **66**, 95–100.
7. Kant, S.G., Van Haeringen, A., Bakker, E., Stec, I., Donnai, D., Mollevanger, P., Beverstock, G.C., Lindemann-Kusse, M.C. and Van Ommen, G.J.B. (1997) Pitt–Rogers–Danks syndrome and Wolf–Hirschhorn syndrome are caused by a deletion in the same region on chromosome 4p16.3. *J. Med. Genet.*, **34**, 569–572.
8. Altherr, M.R., Denison, K., Clemens, M., Quarrel, O. and Wright, T.J. (1996) Molecular overlap in Wolf–Hirschhorn and Pitt–Rogers–Danks syndromes. *Am. J. Hum. Genet.*, **59**, Suppl., A23.
9. Donnai, D. (1986) Brief clinical report: a further patient with the Pitt–Rogers–Danks syndrome of mental retardation, unusual face, and intrauterine growth retardation. *Am. J. Med. Genet.*, **24**, 29–32.
10. Rubinstein, J.H. and Taybi, H. (1963) Broad thumbs and toes and facial abnormalities. *Am. J. Dis. Child.*, **105**, 588–608.
11. Smith, A.C.M., McGavran, L., Robinson, J., Waldstein, G., Macfarlane, J., Zonona, J., Reiss, J., Lahr, M., Allen, L. and Magenis, E. (1986) Interstitial deletion of (17)(p11.2p11.2) in nine patients. *Am. J. Med. Genet.*, **24**, 393–414.
12. Shprintzen, R.J., Goldberg, R.B., Young, D. and Wolford, L. (1981) The velo-cardio-facial syndrome: a clinical and genetic analysis. *Pediatrics*, **67**, 167–172.
13. Wright, T.J., Ricke, D.O., Denison, K., Abmayr, S., Cotter, P.D., Hirschhorn, K., Keinaenen, M., McDonald-McGinn, Somer, M., Spinner, N., Yang-Feng, T., Zackai, E. and Altherr, M.R. (1997) A transcript map of the newly defined 165 kb Wolf–Hirschhorn syndrome critical region. *Hum. Mol. Genet.*, **6**, 317–324.
14. Baxendale, S., MacDonald, M.E., Mott, R., Francis, F., Lin, C., Kirby, S.F., James, M., Zehetner, G., Hummerich, H., Valdes, J., Collins, F.S., Deaven, L.J., Gusella, J.F., Lehrach, H. and Bates, G.P. (1993) A cosmid contig and high resolution restriction map of the 2 megabase region containing the Huntington's disease gene. *Nature Genet.*, **4**, 181–186.

15. Tripoulas, N., LaJeunesse, D., Gildea, J. and Shearn, A. (1996) The *Drosophila ash1* gene product, which is localized at specific sites on polytene chromosomes, contains a SET domain and a PHD finger. *Genetics*, **143**, 913–928.
16. Tschiersch, B., Hofmann, A., Krauss, V., Dorn, R., Korge, G. and Reuter, G. (1994) The protein encoded by the *Drosophila* position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene complexes. *EMBO J.*, **13**, 3822–3831.
17. Roberts, R.G., Coffey, A.J., Borrow, M. and Bentley, D.R. (1993) Exon structure of the human dystrophin gene. *Genomics*, **16**, 536–538.
18. Miyaki, M., Konishi, M., Tanaka, K., Kikuchi-Yanoshita, R., Muraoka, M., Tasuno, M., Igari, T., Koike, M., Chiba, M. and Mori, T. (1997) Germline mutation of *MSH6* as the cause of hereditary nonpolyposis colorectal cancer. *Nature Genet.*, **17**, 271–272.
19. Nakamura, H., Izumoto, Y., Kambe, H., Kuroda, T., Mori, T., Kawamura, K., Yamamoto, H. and Kishimoto, T. (1994) Molecular cloning of complementary DNA for a novel human hepatoma-derived growth factor. Its homology with high mobility group-1 protein. *J. Biol. Chem.*, **269**, 25143–25149.
20. Friedman, L.S., Ostermeyer, E.A., Lynch, E.D., Welcsh, P., Szabo, C.I., Meza, J.E., Anderson, L.A., Dowd, P., Lee, M.K. and Rowell, S.E. (1995) 22 genes from chromosome 17q21: cloning, sequencing, and characterization of mutations in breast cancer families and tumors. *Genomics*, **25**, 256–263.
21. Baxevanis, A.D. and Landsman, D. (1995) The HMG-1 box protein family: classification and functional relationships. *Nucleic Acids Res.*, **11**, 1604–1613.
22. Jantzen, H.M., Admon, A., Bell, S.P. and Tjian, R. (1990) Nucleolar transcription factor hUBF contains a DNA-binding motif with homology to HMG proteins. *Nature*, **344**, 830–836.
23. Hayashi, T., Hayashi, H. and Iwai, K. (1989) *Tetrahymena* HMG nonhistone chromosomal protein. Isolation and amino acid sequence lacking the N- and C-terminal domains of vertebrate HMG 1. *J. Biochem.*, **105**, 577–581.
24. Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J., Burton, J., Connell, M., Copsey, T., Cooper, J., Coulson, A., Craxton, M., Dear, S., Du, Z., Durbin, R., Favello, A., Fulton, L., Gardner, A., Green, P., Hawkins, T., Hillier, L., Jier, M., Johnston, L., Jones, M., Kershaw, J., Kirsten, J., Laister, N., Latreille, P., Lightning, J., Lloyd, C., McMurray, A., Mortimore, B., O'Callaghan, M., Parsons, J., Percy, C., Rifken, L., Roopra, A., Saunders, D., Shownkeen, R., Smaldon, N., Smith, A., Sonnhammer, E., Staden, R., Sulston, J., Thierry-Mieg, J., Thomas, K., Vaudin, M., Vaughan, K., Waterston, R., Watson, A., Weinstock, L., Wilkinson-Sproat, J. and Wohldman, P. (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C.elegans*. *Nature*, **368**, 32–38.
25. Bowser, R. (1996) Assignment of the human *FAC1* gene to chromosome 17q24 by fluorescence *in situ* hybridization. *Genomics*, **38**, 455–457.
26. Kim, S.S., Chen, Y.M., O'Leary, E., Witzgall, R., Vidal, M. and Bonventre, J.V. (1996) A novel member of the RING finger family, KRIP-1, associates with the KRAB-A transcriptional repressor domain of zinc finger proteins. *Proc. Natl Acad. Sci. USA*, **24**, 15299–15304.
27. Seelig, H.P., Moosbrugger, I., Ehrfeld, H., Renz, M. and Genth, E. (1995) The major dermatomyositis-specific Mi-2 autoantigen is a presumed helicase involved in transcriptional activation. *Arthritis Rheum.*, **38**, 1389–1399.
28. Nagamine, K., Peterson, P., Scott, H.S., Kudoh, J., Minoshima, S., Heino, M., Krohn, K.J.E., Lalioti, M.D., Mullis, P.E., Antonarakis, S.E., Kawasaki, K., Asakawa, S., Ito, F. and Shimuzu (1997) Positional cloning of the *APECED* gene. *Nature Genet.*, **17**, 393–398.
29. Finnish–German APECED consortium (1997) An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc finger domains. *Nature Genet.*, **17**, 399–403.
30. Gibbons, R.J., Bachoo, S., Picketts, D.J., Aftimos, S., Asenbauer, B., Bergoffen, J.A., Berry, S.A., Dahl, N., Fryer, A., Keppler, K., Kurosawa, K., Levin, M.L., Masuno, M., Neri, G., Pierpont, M.A., Slaney, S.F. and Higgs, D.R. (1997) Mutations in transcription regulator ATRX establish the functional significance of a PHD-like domain. *Nature Genet.*, **17**, 147–148.
31. Le Douarin, B., Zechel, C., Garnier, J.-M., Lutz, Y., Tora, L., Pierrat, B., Heery, D., Gronemeyer, H., Chambon, P. and Losson, R. (1995) The N-terminal part of TIF1, a putative mediator of the ligand dependent activation function (AF-2) of nuclear receptors, is fused to B-raf in the oncogenic protein T18. *EMBO J.*, **14**, 2020–2033.
32. Gu, Y., Nakamura, T., Adler, H., Prasad, R., Canaani, O., Cimino, G., Croce, C.M. and Canaani, E. (1992) The t(4;11) chromosome translocation of human acute leukemias fuses the *ALL-1* gene, related to *Drosophila trithorax*, to the *AF-4* gene. *Cell*, **71**, 701–708.
33. Mazo, A.M., Huang, D.-H., Mozer, B.A. and David, I.B. (1990) The *trithorax* gene, a *trans*-acting regulator of the bithorax complex in *Drosophila*, encodes a protein with zinc-binding domains. *Proc. Natl Acad. Sci. USA*, **87**, 2112–2116.
34. Abel, K.J., Brody, L.C., Valdes, J.M., Erdos, M.R., Mc Kinley, D.R., Castilla, R.H., Merajver, S.D., Couch, F.J., Friedman, L.S., Ostermeyer, E.A., Lynch, E.D., King, M.-C., Welcsh, P.L., Osborne-Lawrence, S., Spillman, M., Bowcock, A.M., Collins, F.S. and Weber, B.L. (1996) Characterization of *EZH1*, a human homolog of *Drosophila enhancer of zeste* near *BRCA1*. *Genomics*, **37**, 161–171.
35. Jones, R.S. and Gelbart, W.M. (1993) The *Drosophila* Polycomb-group gene *Enhancer of zeste* contains a region with sequence similarity to *trithorax*. *Mol. Cell. Biol.*, **13**, 6357–6366.
36. Tkachuk, D.C., Kohler, S. and Cleary, M.L. (1992) Involvement of a homolog of *Drosophila trithorax* by 11q23 chromosomal translocations in acute leukemias. *Cell*, **13**, 691–700.
37. Djabali, M., Selleri, P., Parry, P., Bower, M., Young, B. and Evans, G.A. (1992) A *trithorax*-like gene is interrupted by chromosome 11q23 translocations in acute leukaemias. *Nature Genet.*, **1**, 113–118.
38. Johnston, M., Andrews, S., Brinkman, R., Cooper, J., Ding, H., Dover, J., Du, Z., Favello, A., Fulton, L., Gattung, S., Geisel, C., Kirsten, J., Kucaba, T., Hillier, L., Jier, M., Johnston, L., Langston, Y., Latreille, P., Louis, E.J., Macri, C., Mardis, E., Menezes, S., Mouser, L., Nhan, M., Rifkin, L., Riles, L., St Peter, H., Trevaskis, K., Vaughan, K., Vignati, D., Wilcox, L., Wohldman, P., Waterston, R., Wilson, R. and Vaudin, M. (1994) Complete nucleotide sequence of *Saccharomyces cerevisiae* chromosome VIII. *Science*, **265**, 2077–2082.
39. Obermaier, B., Gassenhuber, J., Piravandi, E. and Domdey, H. (1995) Sequence analysis of a 78.6 kb segment of the left of *Saccharomyces* chromosome II. *Yeast*, **15**, 1103–1112.
40. Shearn, A., Hersberger, E. and Hersbeger, G. (1987) Genetic studies of mutations at two loci of *Drosophila melanogaster* which cause a wide variety of homeotic transformations. *Wilhelm Roux's Arch. Dev. Biol.*, **196**, 231–242.
41. Nislow, C., Ray, E. and Pillus, L. (1997) SET1, a yeast member of the Trithorax family, functions in transcriptional silencing and diverse cellular processes. *Mol. Biol. Cell*, **8**, 2421–2436.
42. Richelda, R., Ronchetti, D., Baldini, L., Cro, L., Viggiano, L., Marzella, R., Rocchi, M., Otsuki, T., Lombardi, L., Maiolo, A.T. and Neri, A. (1997) A novel chromosomal translocation t(4;14) (p16.3;q32) in multiple myeloma involves the fibroblast growth-factor receptor 3 gene. *Blood*, **90**, 4062–4070.
43. Chesi, M., Nardini, E., Brents, L.A., Schrock, E., Ried, T., Kuehl, W.M. and Bergsagel, P.L. (1997) Frequent translocation t(4;14) (p16.3;q32) in multiple myeloma is associated with increased expression and activating mutations of fibroblast growth factor receptor 3. *Nature Genet.*, **16**, 260–264.
44. Lung, C.-C. and Chan, E.K.L. (1996) Simplifying 5′ RACE in the hunt for full-length cDNAs. *Trends Genet.*, **12**, 389–391.
45. Moorman, A.F.M., Vermeulen, J.L.M., Koban, M.U., Schwartz, K., Lamers, W.H. and Boheler, K.R. (1995) Patterns of expression of sarcoplasmic reticulum calcium ATPase and phospholamba mRNAs during rat heart development. *Circ. Res.*, **76**, 616–625.
46. Moorman, A.F.M., de Boer, P.A.J, Vermeulen, J.L.M. and Lamers, W.H. (1993) Practical aspects of radio-isotopic *in situ* hybridization on RNA. *Histochem. J.*, **25**, 251–266.
47. Altschul, S.F., Gish, W., Miller, W., Meyers, E.W. and Lipman, D. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
48. Uberbacher, E.C. and Mural, R.J. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor–neural network approach. *Proc. Natl Acad. Sci. USA*, **88**, 11261–11265.

NOTE ADDED IN PROOF

Recent publications (Cui *et al.*, *Nature Genet.*, **18**, 331–337; Cardoso *et al.*, *Hum. Mol. Genet.*, **7**, 679–684) implicate the SET domain as the key interaction site with dual-specific protein phosphatases and antiphosphatases. The phosphorylation status of SET domain proteins is proposed to be involved in regulating the balance between differentiation and transformation, through stable chromatin modification, while their involvement in fusion genes leads to oncogenic transformation. These observations greatly strengthen the case for direct involvement of *WHSC1* in development and of *IgH/WHSC1* fusion gene products in t(4:14)-based multiple myeloma.