Published online 17 March 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 13* ***5369–5378***
doi:10.1093/nar/gkr091

Discovery of *SMAD4* promoters, transcription factor binding sites and deletions in juvenile polyposis patients

Daniel Calva, Fadi S. Dahdaleh, George Woodfield, Ronald J. Weigel, Jennifer C. Carr, Sathivel Chinnathambi and James R. Howe*

Department of Surgery, Carver College of Medicine, University of Iowa Hospitals and Clinics, Iowa City, 52242-1086 IA, USA

Received November 2, 2010; Revised February 3, 2011; Accepted February 4, 2011

ABSTRACT

Inactivation of *SMAD4* has been linked to several cancers and germline mutations cause juvenile polyposis (JP). We set out to identify the promoter(s) of *SMAD4,* evaluate their activity in cell lines and define possible transcription factor binding sites (TFBS). 5′-rapid amplification of cDNA ends (5′-RACE) and computational analyses were used to identify candidate promoters and corresponding TFBS and the activity of each was assessed by luciferase vectors in different cell lines. TFBS were disrupted by site-directed mutagenesis (SDM) to evaluate the effect on promoter activity. Four promoters were identified, two of which had significant activity in several cell lines, while two others had minimal activity. *In silico* analysis revealed multiple potentially important TFBS for each promoter. One promoter was deleted in the germline of two JP patients and SDM of several sites led to significant reduction in promoter activity. No mutations were found by sequencing this promoter in 65 JP probands. The predicted TFBS profiles for each of the four promoters shared few transcription factors in common, but were conserved across several species. The elucidation of these promoters and identification of TFBS has important implications for future studies in sporadic tumors from multiple sites, and in JP patients.

INTRODUCTION

SMAD4 is a tumor suppressor gene that is essential for transforming growth factor β (TGFβ) signalling (1), which plays important roles in cell differentiation, growth and apoptosis. It is the human ortholog of the *Drosophila Mad* (mothers against decapentaplegic) and *Caenorhabditis elegans sma-4* genes. Originally called *DPC4* (deleted in pancreatic cancer 4) due to the finding that the majority of pancreatic cancers have 18q allelic loss (2), it was later renamed *SMAD4* to better reflect its orthology to its worm and fly gene counterparts (3). It is the common intracellular mediator for the TGFβ, bone morphogenetic protein (BMP), activin and inhibin pathways. Its role is to form oligomers with receptor regulated SMAD proteins (SMAD1, 2, 3, 5 and 8) phosphorylated after the binding of ligand to the Types II and I cell surface receptors, then these complexes migrate to the nucleus to regulate transcription of target genes (4,5).

A variety of human cancers have been shown to have loss of heterozygosity at the *SMAD4* locus on chromosome 18q21, including 50% of pancreatic cancers (6,7), 41% of cervical cancers (8,9), >60% of colorectal cancers (10), 25% of small intestinal cancers (11), 27% of thyroid cancers (12) and 60% of gastric carcinomas (13). Furthermore, up to 21% of juvenile polyposis (JP) patients have germline mutations in *SMAD4* (14,15).

The promoter regions of genes are important regulatory regions for RNA and protein expression and may play a role in many diseases. Studies of *SMAD4* gene regulation have been limited thus far. Minami *et al.* (16) suggested that the region immediately upstream from the 5′-untranslated region (5′-UTR) and first coding exon had promoter activity. A later report found two substitutions in endometrial cancers within this proposed promoter region (17). Roth *et al.* (18) screened a region about 14 kb upstream from this putative promoter, but did not find evidence of methylation in colorectal cancer specimens. Other groups did find methylation within the region examined by Roth *et al.* (18) in tumors from patients with esophageal adenocarcinoma (19) and in prostate cancers (20). Kloth *et al.* (9) screened

*To whom correspondence should be addressed. Tel: +319 356 1727; Fax: +319 356 1218; Email: james-howe@uiowa.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

approximately 230 bases further upstream from this region in cervical cancer specimens, but did not find methylation.

Although Minami *et al.* (16) and Zhou *et al.* (17) performed limited functional assays of one potential promoter, no studies have systematically examined *SMAD4* mRNA isoforms to identify additional transcriptional start sites (TSS) and their corresponding promoters. It is becoming increasingly apparent that genes are commonly regulated by multiple promoters, allowing for flexibility of gene expression in different tissues and environments (21). The purpose of this study was to fully characterize the promoter regions located upstream of the 5′-UTR by 5′-rapid amplification of cDNA ends (RACE), computational analysis and functional studies with luciferase reporter assays and to further study these regions for potential TFBS that could be altered by germline mutations or epigenetic modifications leading to the genesis of human cancer. Furthermore, we wanted to screen JP probands that did not have mutations in the coding regions of *SMAD4* and *BMPR1A*, the two genes known to cause JP (14,22), to find out if mutations in a *SMAD4* promoter might account for additional cases of JP.

MATERIALS AND METHODS

RNA extraction and 5′-RACE

RNA was extracted from lymphoblastoid cell lines (LCLs; created from peripheral blood leukocytes from our JP patients at the Baylor College of Medicine, Department of Molecular and Human Genetics Tissue Culture Core Laboratory), normal colon tissue and colon polyps from a JP patient, using RNeasy miniprep columns (Qiagen, Valencia, CA, USA). The cDNA was created using gene-specific primers (GSP) using the Invitrogen 5′-RACE kit (Carlsbad, CA, USA) as per the manufacturer's instructions. Successive rounds of amplification were performed using GSP1 chosen in coding *SMAD4* exon 4 (3′-CCAAGTAATCGTGCATCG-5′), GSP2 in coding exon 2 (3′-GATCTATGCCCGTCTCTGGA-5′), GSP3 in coding exon 1 (3′-CCTGAATACATGTCTAAC AA-5′) and GSP4 in the 5′-UTR (3′-CCTGAATACATG TCTAACAATTTTCCT-5′). The cDNA products were then cloned into the Topo 2.1 vector (Invitrogen) and recombinant clones were sequenced.

In Silico analysis

The Genomatix software suite (www.genomatix.de) was used for computational analysis. Gene2Promoter was first used to identify putative promoter regions, then MatInspector identified all TFBS matching a database of pre-defined matrix descriptions. The comparative genomics feature of Eldorado allowed the analysis of a group of *SMAD4* orthologous genes across species. Common TFBS were then processed through FrameWorker to define groups of sites that occur in a specific order and are separated by a certain distance across the orthologous sequences. The genomic sequences upstream from the non-coding (NC) exons 1, 3, 4 and 5′-UTR of *SMAD4* were obtained using the UCSC Genome Browser (www.genome.ucsc.edu) assembly GRCh37/hg18 from March 2006.

Plasmid construction and site-directed mutagenesis

Primers were designed using Primer3 v. 0.4.0 (http://frodo.wi.mit.edu/) to clone the four potential promoter regions. To create the deletion constructs, successively shorter PCR products were amplified from genomic DNA using increasingly closer 5′ primers and the same 3′ primer. All primers used had the *Mlu1* endonuclease site incorporated at the 5′-end and the *BglII* site at the 3′-end in order to clone the sequence into the pGL3 luciferase basic reporter vector (Promega, Madison, WI, USA). Deletion constructs were cloned into the pGL3 vector and transformed into *Escherichia coli* and recombinant plasmids were verified by direct sequencing.

Luciferase assays

The normal human colon fibroblast cell line CRL-1459, human embryonic kidney cell line HEK-293, breast cancer cell line MCF-7 and human colon cancer cell line CCL-247 were obtained from the American Type Culture Collection (Manassas, VA, USA) and cultured according to the ATCC recommendations. The transfection mix consisted of 5 μg of pGL3/deletion construct insert, 1 μg of *Renilla* control vector with CMV promoter (pRL-CMV; Promega), 36 μl of Transfast reagent and minimum essential media (MEM) without serum for a total volume of 1 ml. Prior to adding the transfection mix, 5×10^5 cells from each cell line were added to 6-well plates (Corning, NY, USA) in MEM with 10% serum and penicillin/streptomycin/amphotericin (PSA) and grown to 80–90% confluence. One milliliter of transfection mix was added to each well and allowed to incubate for 4 h. After the incubation, 2 ml of MEM with 10% serum and PSA were added to each well and the cells allowed to grow for 72 h. All deletion constructs from each promoter plus the pGL3-basic vector (without insert) were co-transfected with the pRL-CMV control vector in triplicate. Cells were then harvested and assayed using the Dual-Luciferase reporter assay kit (Promega). The final amount of firefly luciferase activity for each construct was determined by subtracting the background firefly luciferase activity from the control pGL3 basic vector without construct and then normalizing to the *Renilla* luciferase activity for each individual reaction.

Screening for promoter genetic alterations

The promoter region with the greatest luciferase activity and upstream from the greatest number of transcripts discovered by 5′-RACE was then sequenced in 65 JP probands that did not have coding mutations of *SMAD4* or *BMPR1A* or large exonic deletions by multiplex ligation-dependent probe amplification (MLPA; MRC Holland, Amsterdam). Two individuals were previously described to have a large deletion by MLPA at the 5′-end of *SMAD4* (15). One of these deletions was further characterized to identify the breakpoints using an

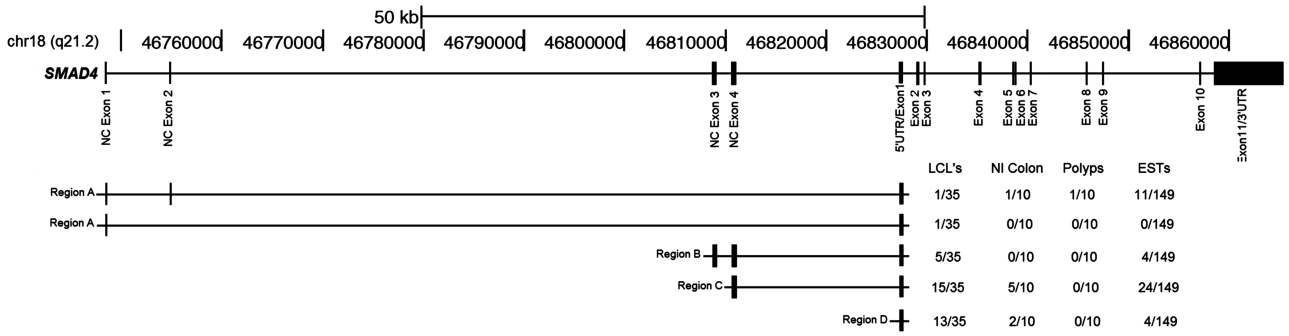

Figure 1. Map of chromosome 18 near *SMAD4*, including NC and coding exons (top line) and five different splice variants found by 5′-RACE below. The number of 5′-RACE clones found in various tissues and ESTs described in the UCSC genome browser (www.genome.ucsc.edu) are shown at the right of each isoform. From the 149 ESTs described, only 43 of them contained the NC exons and 5′-UTR, while the rest consisted only of coding exons. The map positions of NC exons are: NC-exon 1:46 748 387–46 748 482; NC-exon 2:46 754 765–46 754 933; NC-exon 3:46 808 762–46 809 033; NC-exon 4:46 810 582–46 810 991 and 5′-UTR 46 827 288–46 827 414. Regions A, B, C, D are predicted to be the potential promoters for these isoforms.

oligonucleotide array comparative genomic hybridization (CGH) chip (HG18_WG_CGH_7 of 8 array; NimbleGen Systems, Madison, WI, USA), which has probes spaced approximately every 700 bp.

Chromatin immunoprecipitation sequencing and DNaseI hypersensitivity sites

MCF-7 cells were grown in MEM supplemented with 10% FBS and were harvested at 80–90% confluence to obtain 8×10^7 cells. Cells were then cross-linked with 1% formaldehyde for 10 min at room temperature, the cell pellet was then resuspended in ChIP lysis buffer with complete protease inhibitor tablet (Roche), then sonicated to produce chromatin fragments approximately 350 bases in length. Five micro grams of Anti-RNA Pol II (mouse monoclonal, Cat No: 17-620, Millipore) were added and incubated overnight at 4°C. The antibody/protein/DNA complexes were eluted and complexes treated with 10 µg RNase A and 0.3 M NaCl at 67°C for 4 h to reverse the cross-links. DNA/proteins were precipitated, the proteins were digested in proteinase K, then the DNA purified with a QIAGEN PCR purification column. The ChIP-enriched DNA was then prepared for Chromatin immunoprecipitation sequencing (ChIP-Seq) and short-read sequencing performed using the Illumina GA2 sequencing system. The raw sequencing images were analyzed using the Illumina analysis pipeline and reads were aligned to the human reference genome (NCBI v36, hg18). ChiP-seq data for other cell lines and Digital DNaseI-Seq were obtained from the ENCODE Project Consortium (23).

Determining relative abundance of each NC isoform

Total RNA was extracted from CRL-1459, MCF-7 and HEK-293 cells using the RNeasy kit (Qiagen), then transferred to a nitrocellulose membrane and UV cross-linked. Oligos were designed for each of five *SMAD4* NC exon splice variants, which spanned across adjacent NC exons, adjacent NC exon and 5′-UTR, or 5′-UTR (Probe 1: 5′-GTATTCAGGATAACTAACCTG CTTTAAGTTGGC-3′; Probe 2: 5′-GACATGTATTCA GGATAACCCTCTCCCCG-3′; Probe 3: 5′-ATTCAGG ATAACAGATTCTCTGAGTCAGGATTC-3′; Probe 4: 5′-ATGTATTCAGGATAACCTGGGCTCGGGCGG-3′; Probe 5: 5′-TTGGTGTATTCGTAATAGACATATTGT CCAT-3′). Each oligo was end-labelled with [γ32P]-ATP using T4 Polynucleotide Kinase, then hybridized to the membranes in conjunction with denatured salmon sperm DNA. Blots were washed, exposed to film and relative concentrations determined using a Typhoon FLA 7000 (General Electric).

RESULTS

NC splice variants of *SMAD4*

5′-RACE revealed that there were four NC exons upstream of the 5′-UTR and the first coding exon (exon 1), which starts at 46 827 288 (Figure 1). NC exon 1 began at 78 901 bp upstream from the 5′-UTR and NC exon 4 was 16 706 bp upstream from the 5′-UTR. There were a total of five different splice variants found and the relative abundance of each from LCls, normal colon, JP and expressed sequence tags (ESTs) in the UCSC genome browser (www.genome.ucsc.edu) is shown in the figure.

Defining promoter regions

There were four areas that by 5′-RACE and EST data were candidates for being potential promoters (regions A, B, C and D, Figure 1). Using Gene2Promoter software (Genomatix), there were six regions that were predicted to be promoters, including all four of the regions predicted by 5′-RACE. To identify the TSS, we analysed published libraries of capped analysis gene expression (CAGE) tags (Genomatix). CAGE tags help identify transcriptional start sites on a genome-wide scale and are short sequences that originate from the 5′-end of mRNA transcripts. For this study, we designated the TSS for each isoform as being the site of the most 5′-CAGE tag. In the case of Promoter A, there were four different potential TSS distributed over 56 bp, supported by 28 CAGE tags. For Promoter C, there were three TSS over a 30 bp range supported by 63 CAGE tags. The TSS for Promoter D was designated as the beginning of the 5′-UTR, which was confirmed by the

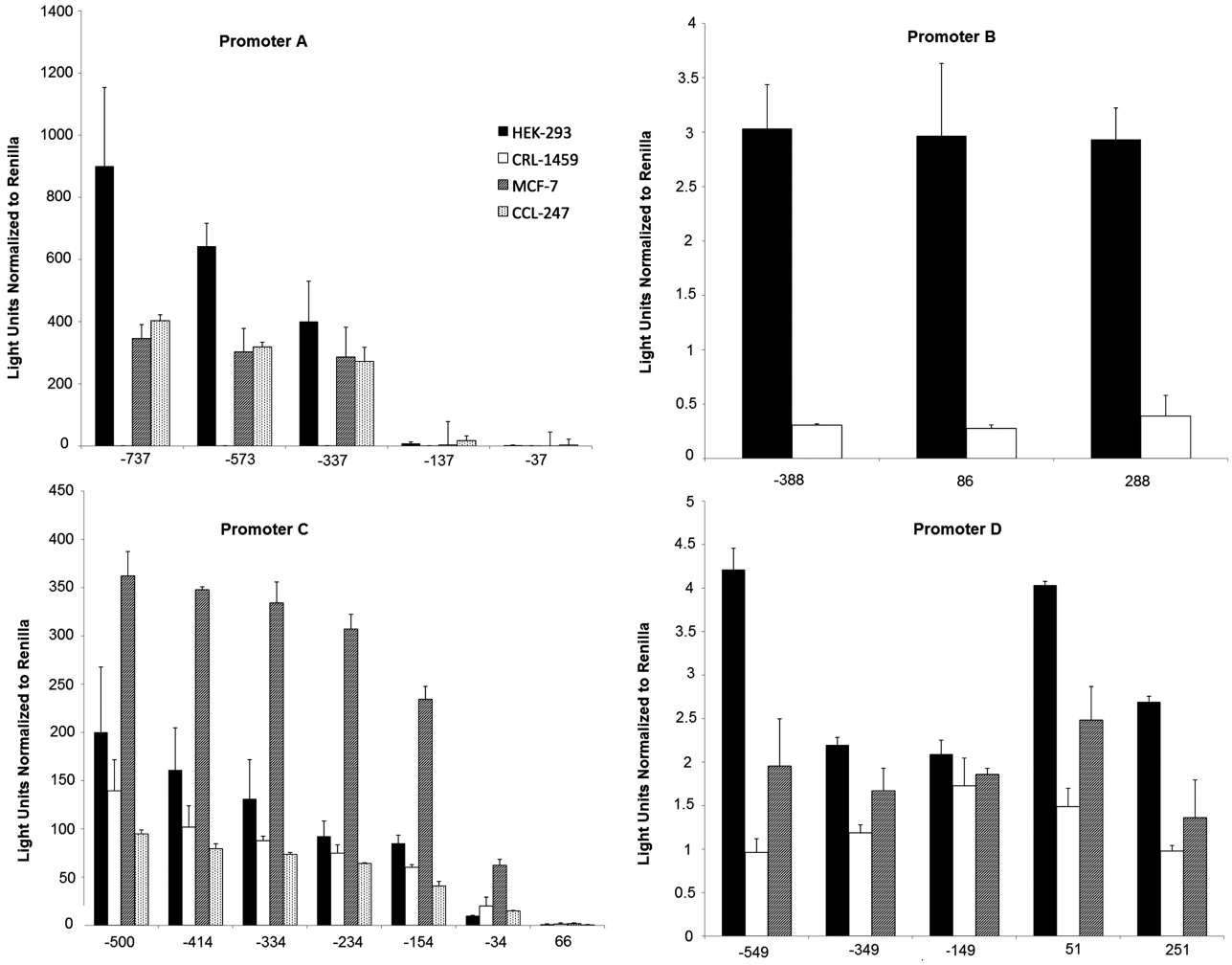

Figure 2. Luciferase expression from the various deletion constructs of promoter A, B, C and D from the 4 cell lines, HEK-293, CRL-1459, MCF-7 and CCL-247. Deletion construct numbering is relative to the TSS; the last constructs of promoter A end at +19 relative to the TSS, for B at +473, C at +156 from the TSS and for D at +449 into the first coding exon. Note that there is no activity seen for promoter A in CRL-1459; that promoter B was only evaluated in HEK-293 and CRL-1459 cells and that Promoter D was not evaluated in CCL-247 cells.

finding of many 5′-RACE clones splicing into this site from upstream NC exons. The TSS of Promoter B was less certain, as it was based upon the most 5′ transcript found in our 5′-RACE clones and ESTs in the UCSC browser; there were no corresponding CAGE tags.

Activity of *SMAD4* promoters in cell lines

Two of the four putative promoters had significant luciferase activity when transfected into different cell lines. The region cloned as promoter A was the 736 bases immediately 5′ to the TSS of NC exon 1 and the first 19 bases of NC exon 1, which was found to have significant activity in three cell lines (HEK-293, MCF-7 and CCL-247), but not in CRL-1459 (Figure 2a). Promoter C was the 500 base region upstream from the TSS of NC exon 4, plus 156 bp of NC exon 4, which showed significant luciferase activity in all four cell lines (Figure 2c).

The two other promoters (B and D; Figure 2b and d) just showed modest activity in HEK-293 cells and even less in CRL-1459 cells. The full construct for promoter D spanned from 46 826 739 to 46 827 736 in the genomic sequence and included the 5′-UTR and a part of coding exon 1.

Transcription factor binding sites

Comparative genomics using ElDorado software (Genomatix), revealed several TFBS within each promoter. Eight different species were analysed for conservation (monkey, chimpanzee, human, mouse, rat, cow, pig and opossum). Using ModelInspector (Genomatix), promoter A had eight TFBS which fit into three-element models that are conserved across species and had 42 CpGs (Figure 3a). These TFBS include: homeodomain transcription factors (HOMF); GATA binding factors (GATA); human and murine EST factors (ETSF); activator-, mediator- and TBP-dependent core promoter element for RNA polymerase II transcription from TATA-less promoters (XCPE); CTCF and BORIS gene family transcriptional regulators (CTCF); nuclear

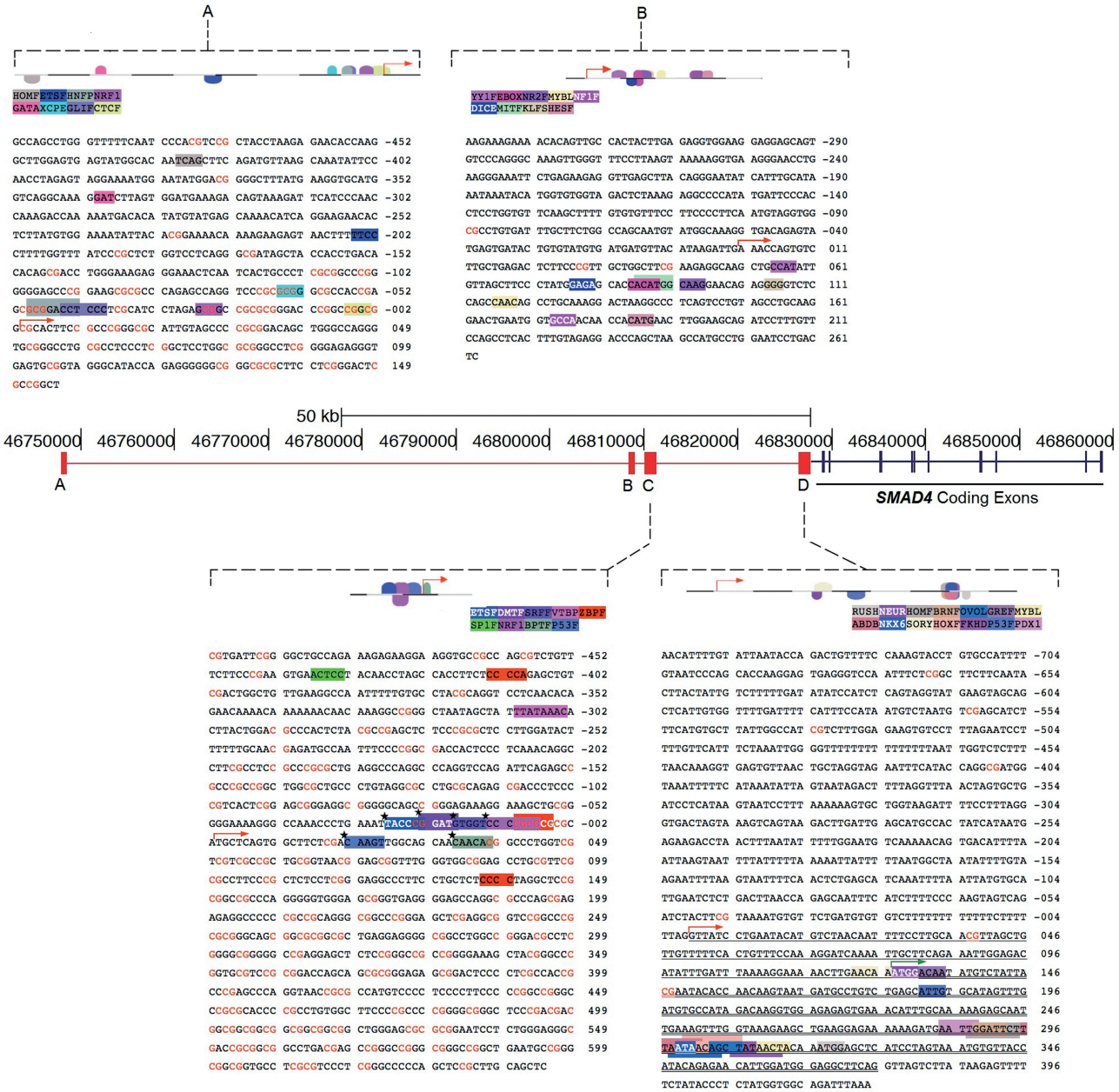

Figure 3. Sequences of each promoter region, including phylogenetically conserved TFBS as predicted by comparative genomics, where highlighted nucleotides exhibit the highest conservation within the TFBS consensus sequence. The center line is the Chromosome 18 map, with the red vertical lines representing the four promoters relative to *SMAD4* coding exons. Red arrows designate TSS in each sequence, the numbers on the right of each sequence are relative to the TSS; CG dinucleotides are shown in red. (**A**) Gene2promoter ID: GXP_57502, which overlaps with promoter A. There were eight TFBS predicted to be phylogenetically conserved (with spacing demonstrated by the line with colored beads at top), represented by colored highlighted regions at their sequences and names given to each color above: (HOMF, XCPE, ETSF, HNFP, GLIF, NRF1 and CTCF. (**B**) GXP_1261977, which overlaps with promoter B. TFBS were predicted to be: YY1F, EBOX, NR2F, MYBL, NF1F, DICE, MITF, KLFS and HESF. (**C**) GXP_57777, which overlaps with promoter C; six TFBS were predicted: ETSF, DMTF, SRFF, NRF1, BPTF and P53F. The sequences with stars are predicted to be phylogenetically conserved between species, while the colored sequences without stars were predicted as TFBS by the Genomatix software (but were not conserved) and were also evaluated by SDM. (**D**) GXP_1484970, which overlaps with promoter D. In promoter D, 14 TFBS were predicted: RUSH, NEUR, HOMF, BRNF, OVOL, GREF, MYBL, ABDB, NKX6, SORY, HOXF, FKHD, P53F and PDX1. The sequence within promoter D that has a single underline is the 5′-UTR and the doubly underlined represents coding exon 1. The green arrow is the translation start site.

respiratory factor 1 (NRF1); GLI zinc finger family (GLIF) and histone nuclear factor P (HNFP).

Promoter B had nine TFBS that were phylogenetically conserved in the species examined: Activator/repressor binding to transcription initiation site (YY1F); E-Box binding factors (EBOX); Nuclear receptor subfamily 2 factors (NR2F); cellular and viral myb-like transcriptional regulators (MYBL); nuclear factor 1 (NF1F); downstream immunoglobulin control element (DICE); microphthalmia transcription factor (MITF); Krueppel-like transcription

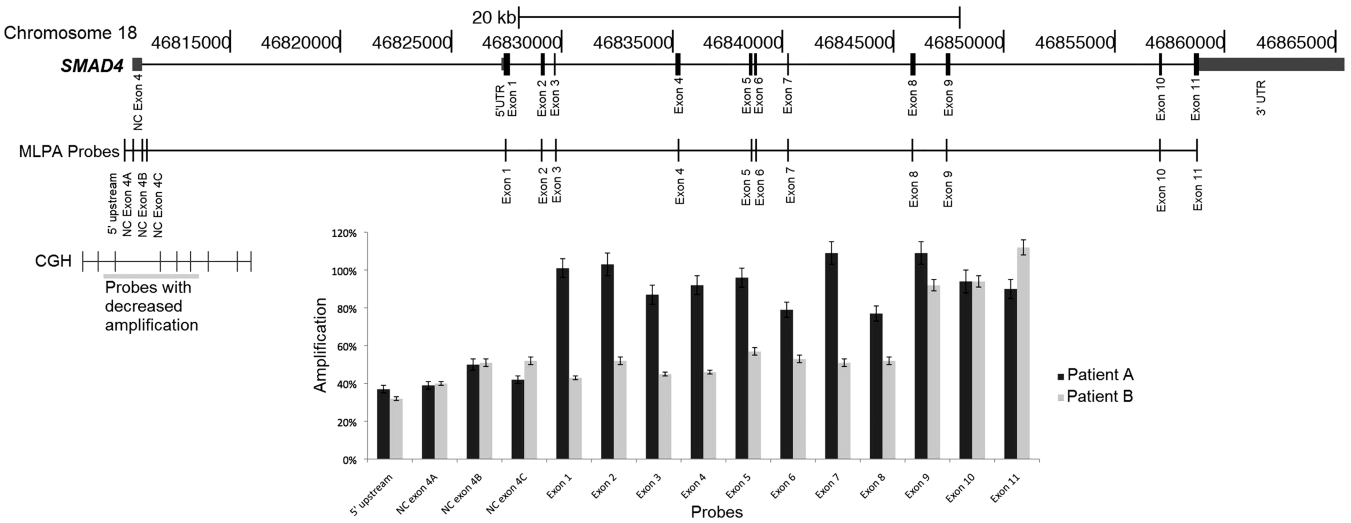

Figure 4. Map of chromosome 18 in the vicinity of the *SMAD4* gene. The top line shows the map positions and exons of *SMAD4*; the second line the location of MLPA probes and the smaller line at left beneath, the location of CGH probes. The bar graph below shows MLPA results of two patients, showing a 50% decrease in amplification of the three probes within NC exon 4 and the probe located just 5′ upstream of NC exon 4 (Patient A) and additionally up through coding exon 8 in Patient B.

factors (KLFS) and vertebrate homologues of enhancer of split complex (HESF).

Promoter C had six different TFBS that fit into distinct four-element models that are conserved across species and 55 CpGs (Figure 3c). These TFBS include: human and murine EST1 factors (ETSF); serum response element binding factor (SRFF); nuclear respiratory factor 1 (NRF1); cyclin D binding myb-like transcription factor (DMTF); bromodomain and PHD domain TF (BPTF) and the p53 tumor suppressor (P53F).

For Promoter D, there were 14 phylogenetically conserved TFBS: SWI/SNF related nucleophosphoproteins with a RING finger DNA binding motif (RUSH); NeuroD, Beta2, HLH domain (NEUR), homeodomain transcription factors (HOMF); Brn POU domain factors (BRNF); OVO homolog-like transcription factors (OVOL); glucocorticoid responsive and related elements (GREF); cellular and viral myb-like transcriptional regulators (MYBL); abdominal B-type homeodomain transcription factors (ABDB); NK6 homeobox transcription factors (NKX6); SOX/SRY-sex/testis determining and related HMG box factors (SORY); paralog hox genes 1–8 from the four hox clusters A–D (HOXF); fork head domain factors (FKHD); p53 tumor suppressor (P53F) and pancreatic and intestinal homeodomain transcription factor (PDX1). Note that all of these factors except for MYBL fell within the first coding exon of *SMAD4*.

Promoter sequencing, screening for deletions, mutagenesis of TFBS

There were 65 JP probands meeting the clinical criteria for JP (24) who had no mutations identified by sequencing of all coding exons of both *SMAD4* and *BMPR1A*. Two patients were found to have larger exonic deletions of *SMAD4* by MLPA, showing 50% decrease in amplification of the probes in the 5′ region of the gene. Patient A had a heterozygous deletion of a probe upstream of NC exon 4 and all probes located within NC exon 4 and Patient B, in addition to these same probes, also had loss of all probes from coding exons 1–8 (Figure 4). Using the CGH chip, the deletion of Patient A was confirmed and further characterized. Probes from position 46 809 737 through 46 813 229 had 50% decreased amplification, which meant the deletion included all of Promoter C, NC exon 4 and at least 2 kb of downstream intron. Sequencing of Promoter C in the 65 JP probands and 100 control patients revealed no mutations or polymorphisms.

Due to the deletions found in two JP patients and its functionality in our luciferase models, several potential TFBS predicted by Genomatix and PromoterScan software were evaluated by site-directed mutagenesis (SDM) in Promoter C. When both thymine residues of the vertebrate TATA binding protein factor of Promoter C were changed to guanines by SDM, this resulted in only 10% of the luciferase activity compared to the wild-type promoter (Figure 5). When other TFBS were changed by 2 nt by SDM, the activity with mutation of one SP-1 site was 38% and one zinc binding protein factor (ZBPF) site was 37% (74 and 80% of the wild-type for two other ZBPF sites). With respect to the phylogenetically conserved sites, the luciferase activity with mutation of the ETSF1 site was 36% of that seen for the wild-type vector, with mutation of DMTF was 52% and minimal change was observed with mutation of NRF1 (99% of wild-type), SRFF (112%) and BPTF sites (99%). Interestingly, when one base of the p53 site was changed, the luciferase activity increased to 162% of the wild-type sequence.

ChIP-Seq, DNAseI and abundance of RNA isoforms

RNA Polymerase II ChIP-Seq results in MCF-7 cells and ENCODE data (23) from HEK-293 cells revealed that promoters A and C were the most active areas of RNA

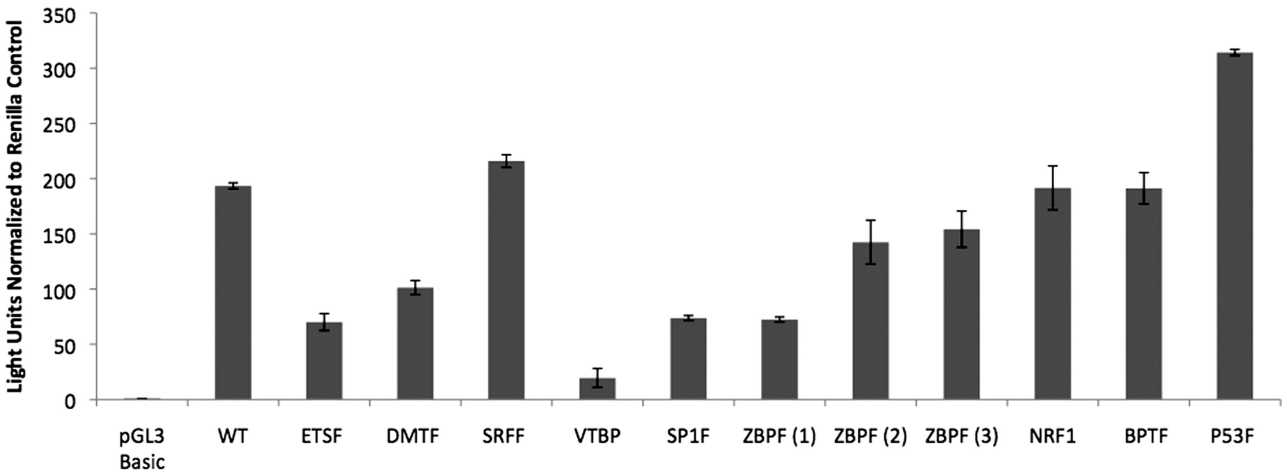

Figure 5. Luciferase activities in HEK-293 cells after mutating selected TFBS in promoter C. The wild-type luciferase activity is compared to 11 different mutant constructs: ETSF; DMTF; SRFF; VTBP; SP1F: Sp/KLF family of transcription factors; ZBPF: Zinc finger transcription factors (three different sites); NRF1; BPTF; P53F: p53 tumor suppressor.

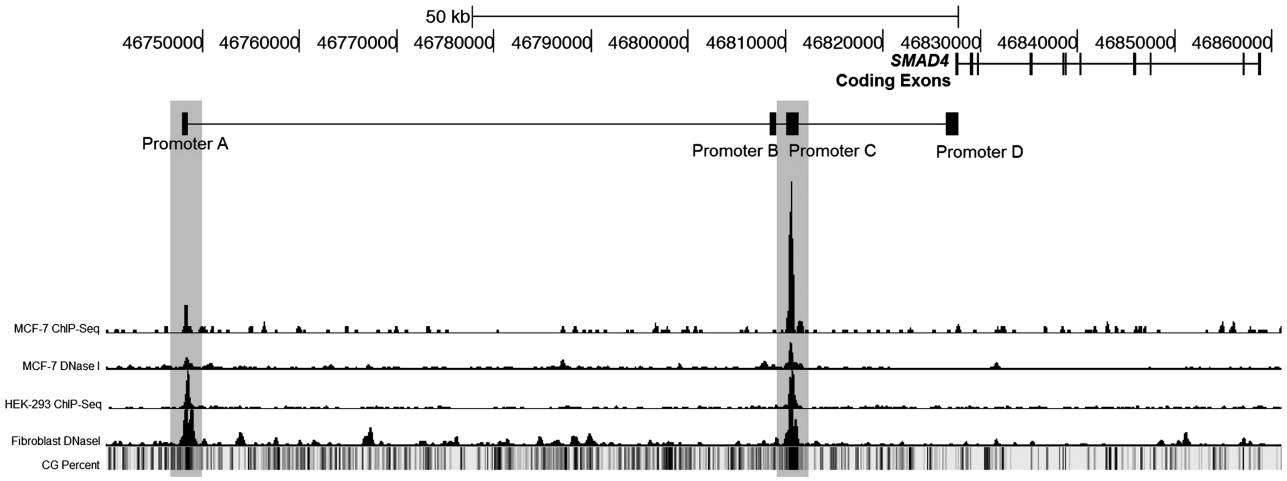

Figure 6. Map of Chromosome 18, with (lines from top to bottom): the coding exons of *SMAD4*; promoter regions; results of ChIP-Seq for RNA Polymerase II in MCF-7 cells; DNase I hypersensitivity sites in MCF-7 cells (from ENCODE data); ChIP-Seq for RNA Polymerase II in HEK-293 cells (from ENCODE data); DNase I hypersensitivity sites in fibroblasts (from ENCODE data) and the density of CG dinucleotides. The areas under promoters A and C are highlighted in gray.

polymerase II binding, with promoter C being much more active than A and negligible activity was seen for promoters B and D (Figure 6). ENCODE DNaseI data in MCF-7 cells and fibroblasts demonstrated the same pattern. These data support the luciferase results and further validates the importance of these two promoters. The relative abundance of different NC RNA isoforms were evaluated by slot blot hybridization to RNA derived from HEK-293, CRL-1459 and MCF-7 cells using probes spanning contiguous NC segments (except the last isoform). The most common isoform found by this method in all three lines was that of NC exon 1 splicing into the 5′-UTR, which is presumed to be under the control of Promoter A. The second most common was NC exon 4 splicing into the 5′-UTR, which would be just downstream of Promoter C (Supplementary Table S1).

DISCUSSION

The complexity of the promoters for *SMAD4* has been under-appreciated until now. Minami *et al.* (16) evaluated 1285 bases immediately upstream of the 5′-UTR of *SMAD4* plus 45 in the 5′-UTR (comparable to Promoter D in this study) and described a peak luciferase activity of 38 light units compared to the control of 2.2 light units. However, this region lacked typical promoter region characteristics, such as high CG content (CpG islands) or a TATA-box, but did have some TATA-like structures (TA AAAT) and other potential transcriptional binding sites. Current evidence suggests that low CG content promoters are more likely to be cell specific, while CG rich are more likely to be ubiquitously expressed (25).

Zhou *et al.* (17) found that six endometrial carcinoma specimens had LOH of chromosome 18q21 markers and altered transcription of *SMAD4* and sequencing for

mutations of a part of promoter D revealed substitutions in two patients. Functional assays (chloramphenicol acetyltransferase) of promoter activity revealed these substitutions led to significantly reduced activity relative to the wild-type sequence. We found that promoter D had minimal activity in the four different cell lines we tested, despite the inclusion of the 5′-UTR and part of coding exon 1 where the majority of predicted TFBS were concentrated and similar results were obtained when the 5′-UTR and this portion of exon 1 were left out (data not shown). The lack of RNA polymerase II binding sites and DNaseI sites around promoter D support the notion that this is not a very active site of transcription, at least in the cell lines examined.

Promoter C was found to have significant luciferase activity in all four cell lines evaluated and this region has sparked interest from investigators in the past. This began with Roth *et al.* (18) who screened an overlapping region of 700 bp from position 46 810 540 to 46 811 239 for methylation in colorectal cancer specimens. They selected this region because it was CG-rich and an unpublished manuscript by Hagiwara *et al.* apparently examined the sequence immediately upstream from a newly discovered NC exon. The region they looked at is primarily within and downstream of NC exon 4 (which starts at position 46 810 582 and ends at 46 810 991). However, it did contain 42 bases of promoter C upstream from the TSS and while these bases were found to have negligible luciferase activity in our deletion constructs, the sequence examined would have contained four of six phylogenetically conserved sites. Roth *et al.* did not find methylation in 42 colorectal cancer samples, but these studies may have been inconclusive since this promoter appears to include a larger sequence upstream, as evident by our luciferase models.

Onwuegbusi *et al.* (19) also screened the segment studied by Roth *et al.* and found that 70% of esophageal adenocarcinomas had methylation. The same region was evaluated in prostate cancer specimens and although no evidence of methylation was found in benign prostatic hypertrophy samples, 45% of prostate cancers had methylation. Furthermore, they found that patients with lymph node metastasis had a higher incidence of methylation (63%). They also looked for mutations, but none were identified within the 40 bases of promoter C that were screened, or in the sequences of NC exon 4 (20). Again, these studies examined primarily NC exon 4 and the intron downstream and therefore, the consequence of finding methylated CpGs here upon tumor formation is unclear.

Ando *et al.* screened sequences between 46 810 524 and 46 810 769 for methylation in CRC specimens and found none. The area they studied is mostly within NC exon 4, with only 58 bases upstream of the TSS included (26). Wang *et al.* (27) examined a larger area of promoter C by looking at a region of 200 bases upstream from position 46 810 611, which was believed to be the TSS of NC exon 4 (although current evidence suggests the TSS is now at position 46 810 581). They screened gastric carcinoma specimens and found that 4 specimens out of 75 (5%) had methylation that was associated with decreased expression of SMAD4. Kloth *et al.* (9) looked at an even larger area of promoter C in cervical cancer specimens, by screening for methylation including up to 270 bases upstream from the TSS, to 155 bases downstream and they found no evidence of methylation. This study did not include the other 255 bases upstream to position 46 809 882. It remains to be determined whether the incidence of methylation in gastric and cervical cancer patients would have been higher if these additional bases had been screened. It should also be noted that prior to the current study, there have been no luciferase studies published that confirmed this region had promoter activity.

In the cell lines tested, Promoters A and C were the most functionally active. Promoter C has a greater number of mRNA isoforms that might potentially be regulated by it, a higher CpG content relative to promoter A and showed luciferase activity in all four cell lines, while promoter A had no promoter activity in a normal colon fibroblast cell line (CRL-1459). ChIP-Seq and DNaseI data revealed a greater number of sites at promoter C, but the most abundant RNA isoform found by hybridization was one predicted to be under the control of promoter A. In contrast, none of these studies suggested an important role for either promoter B or D in the cell lines we tested. However, different promoters may play distinct roles in various tissues, during stages of development, or physiologic conditions. Presumably this activity is influenced by the differing context and abundance of transcription factors which are present in each circumstance. Analyzing promoter C with MatInspector (www.genomatix.de/products/MatInspector) revealed several potentially important TFBS that could be involved in regulation of *SMAD4*. From the −500 bp to the −414 bp construct, there was a >45–50% drop in luciferase activity and this region has one zinc finger homeodomain transcription factor binding site (ZFHX), an AP-2 site, an SP-1 site, several C-abl DNA binding sites (CABL), ZBPF sites and a TATA box. This area was not examined in all of the previous studies attempting to screen the *SMAD4* promoter for methylation. Between the −500 bp and −34 bp constructs, there was a loss of >85–95% of the luciferase activity in both CRL-1459 and HEK-293 cell lines. Between the −234 bp and −34 bp deletion constructs, there is an additional SP-1 site, six ZBPF, two AP-2 sites, a CABL site and a possible core promoter element for RNA pol II transcription binding site for TATA-less promoters (Figure 3). Furthermore, phylogenetic data show that there are multiple TFBS that are conserved between species, which is not only a testament to the importance of these regions, but also provides insight into the transcriptional regulatory elements that might play a role in the expression of this important tumor suppressor gene. We focused further attention on the sequence of Promoter C because more of the mRNA isoforms are likely regulated by this region and the deletion seen in one JP patient affected this region. Further analysis by SDM showed how important the TATA box, ZBPF and the SP-1 sites could be in influencing the transcription of

SMAD4, for when these sites were mutated the promoter activity was significantly diminished (Figure 5).

Although screening of our 65 JP probands revealed no germline mutations within Promoter C, we did find two JP patients with germline deletions affecting this region, one of which involved only promoter C and NC exon 4 (and none of the coding exons). Aretz *et al*. (28) also found four JP patients with deletion of these four MLPA probes, but they also had deletions involving all the coding exons as well. These data suggest that promoter alterations play a role in the genesis of JP, as recently reported for *BMPR1A* (29), and therefore, further evaluation of promoter A in JP patients will be of interest. Whether epigenetic inactivation of the normal copy of *SMAD4* leads to polyp formation is another important question to be examined in JP patients. Now that these promoter regions have been characterized, follow-up studies in colon, gastric, cervical, pancreatic and other sporadic cancers will be imperative to define their importance in tumorigenesis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

The Roy J. Carver Charitable Trust and National Institutes of Health (RO1 CA098193). Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Zhou,S., Buckhaults,P., Zawel,L., Bunz,F., Riggins,G., Le-Dai,J., Kern,S.E., Kinzler,K.W. and Vogelstein,B. (1998) Targeted deletion of Smad4 shows it is required for transforming growth factor beta and activin signaling in colorectal cancer cells. *Proc. Natl Acad. Sci. USA*, **95**, 2412–2416.
2. Hahn,S.A., Schutte,M., Hoque,A.T.M., Moskaluk,C.A., da Costa,L.T., Rozenblum,E., Weinstein,C.L., Fischer,A., Yeo,C.J., Hruban,R.H. *et al*. (1996) *DPC4,* a candidate tumor suppressor gene at human chromosome 18q21.1. *Science*, **271**, 350–353.
3. Derynck,R., Gelbart,W.M., Harland,R.M., Heldin,C.-H., Kern,S.E., Massagué,J., Melton,D.A., Mlodzik,M., Padgett,R.W., Roberts,A.B. *et al*. (1996) Nomenclature, vertebrate mediators of TGFb family signals. *Cell*, **87**, 173.
4. Wrana,J.L. and Attisano,L. (1996) MAD-related proteins in TGF-B signaling. *Trends Genet*., **12**, 493–496.
5. Heldin,C.H., Miyazono,K. and Ten Dijki,P. (1997) TGF-β signaling from cell membrane to nucleus through SMAD proteins. *Nature*, **390**, 465–471.
6. Hahn,S.A., Shamsul,H.A.T.M., Moskaluk,C.A., da Costa,L.T., Schutte,M., Rozenblum,E., Seymour,A.B., Weinstein,C.L., Yeo,C.J., Hruban,R.H. *et al*. (1996) Homozygous deletion map at 18q21.1 in pancreatic cancer. *Cancer Res*., **56**, 490–494.
7. Goggins,M., Shekher,M., Turnacioglu,K., Yeo,C.J., Hruban,R.H. and Kern,S.E. (1998) Genetic alterations of the transforming growth factor beta receptor genes in pancreatic and biliary adenocarcinomas. *Cancer Res*., **58**, 5329–5332.
8. Kersemaekers,A.M., Kenter,G.G., Hermans,J., Fleuren,G.J. and van de Vijver,M.J. (1998) Allelic loss and prognosis in carcinoma of the uterine cervix. *Int. J. Cancer*, **79**, 411–417.
9. Kloth,J.N., Kenter,G.G., Spijker,H.S., Uljee,S., Corver,W.E., Jordanova,E.S., Flueren,G.J. and Gorter,A. (2008) Expression of Smad2 and Smad4 in cervical cancer, absent nuclear Smad4 expression correlates with poor survival. *Mod. Pathol*., **21**, 866–875.
10. Thiagalingam,S., Lebauer,C., Leach,F.S., Schutte,M., Hahn,S.A., Overhauser,J., Willson,J.K.V., Markowitz,S., Hamilton,S.R., Kern,S.E. *et al*. (1996) Evaluation of candidate tumour suppressor genes on chromosome 18 in colorectal cancers. *Nat. Genet*., **13**, 343–346.
11. Blaker,H., von Herbay,A., Penzel,R., Grob,S. and Otto,H.F. (2002) Genetics of adenocarcinomas of the small intestine, frequent deletions at chromosome 18q and mutations of the *SMAD4* gene. *Oncogene*, **21**, 158–164.
12. Lazzereschi,D., Nardi,F., Turco,A., Ottini,L., D'Amico,C., Mariani-Costantini,R., Gulino,A. and Coppa,A. (2005) A complex pattern of mutations and abnormal splicing of Smad4 is present in thyroid tumours. *Oncogene*, **24**, 5344–5354.
13. Okano,H., Shinohara,H., Miyamoto,A., Takaori,K. and Tanigawa,N. (2004) Concomitant overexpression of cyclooxygenase-2 in HER-2-positive on Smad4-reduced human gastric carcinomas is associated with a poor patient outcome. *Clin. Cancer Res*., **10**, 6938–6945.
14. Howe,J.R., Roth,S., Ringold,J.C., Summers,R.W., Jarvinen,H.J., Sistonen,P., Tomlinson,I.P.M., Houlston,R.S., Bevan,S., Mitros,F.A. *et al*. (1998) Mutations in the *SMAD4*/*DPC4* gene in juvenile polyposis. *Science*, **280**, 1086–1088.
15. Calva-Cerqueira,D., Chinnathambi,S., Pechman,B., Bair,J., Larson-Haidle,J. and Howe,J.R. (2008) The rate of germline mutations and large deletions of *SMAD4* and *BMPR1A* in juvenile polyposis. *Clin. Genet*., **75**, 79–85.
16. Minami,R., Kitazawa,R., Maeda,S. and Kitazawa,S. (1998) Analysis of 5′-flanking region of human *SMAD4* (*DPC4*) gene. *Biochim. Biophys. Acta*, **1443**, 182–185.
17. Zhou,Y., Kato,H., Shan,D., Minami,R., Kitazawa,S. and Matsuda,T. (1999) Involvement of mutations in the *DPC4* promoter in endometrial carcinoma development. *Mol. Carcinog*., **25**, 64–72.
18. Roth,S., Laiho,P., Salovaara,R., Launonen,V. and Aaltonen,L.A. (2000) No *SMAD4* hypermethylation in colorectal cancer. *Br. J. Cancer*, **83**, 1015–1019.
19. Onwuegbusi,B.A., Aitchison,A., Chin,S.F., Kranjac,T., Mills,I., Huang,Y., Lao-Sireix,P., Caldas,C. and Fitzgerald,R.C. (2006) Impaired transforming growth factor beta signaling in Barrett's carcinogenesis due to frequent *SMAD4* inactivation. *Gut*, **55**, 764–774.
20. Aitchison,A.A., Veerakumarasivam,A., Vias,M., Kumar,R., Hamdy,F.C., Neal,D.E. and Mills,I.G. (2008) Promoter methylation correlates with reduced Smad4 expression in advanced prostate cancer. *Prostate*, **68**, 661–674.
21. Hollstein,M. and Hainaut,P. (2010) Massively regulated genes, the example of TP53. *J. Pathol*., **220**, 164–173.
22. Howe,J.R., Bair,J.L., Sayed,M.G., Anderson,M.E., Mitros,F.A., Peterson,G.M., Velculescu,V.E., Traverso,G. and Vogelstein,B. (2001) Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. *Nat. Genet*., **28**, 184–187.
23. ENCODE Project Consortium, Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Marguilies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T. *et al*. (2007) Identification and Analysis of Functional Elements in 1% of The Human Genome by The ENCODE Pilot Project. *Nature*, **447**, 799–816.
24. Jass,J.R. (1994) Juvenile Polyposis. In Phillips,R.K.S., Spigelman,A.D. and Thomson,J.P.S. (eds), *Familial Adenomatous Polyposis and Other Polyposis Syndromes*. Edward Arnold Publishers, London, pp. 203–214.
25. Landolin,J.M., Johnson,D.S., Trinklein,N.D., Aldred,S.F., Medina,C., Shulha,H., Weng,Z. and Myers,R.M. (2010) Sequence features that drive human promoter function and tissue specificity. *Genome Res*., **20**, 890–898.
26. Ando,T., Sugai,T., Habano,W., Jiao,Y.F. and Suzuki,K. (2005) Analysis of *SMAD4*/*DPC4* gene alterations in multiploid colorectal carcinomas. *J. Gastroenterol*., **40**, 708–715.
27. Wang,L.H., Kim,S.H., Lee,J.H., Choi,Y.L., Chul,K.Y., Park,T.S., Hong,Y.C., Wu,C.F. and Kee,S.Y. (2007) Inactivation of *SMAD4*

tumor suppressor gene during gastric carcinoma progression. *Clin. Cancer Res.*, **13**, 102–110.
28. Aretz,S., Stienen,D., Uhlhaas,S., Stolte,M., Entius,M.M., Loff,S., Back,W., Kaufmann,A., Keller,K.M., Blaas,S.H. *et al.* (2007) High proportion of large genomic deletions and a genotype phenotype update in 80 unrelated families with juvenile polyposis syndrome. *J. Med. Genet.*, **44**, 702–709.
29. Calva-Cerqueira,D., Dahdaleh,F.S., Woodfield,G., Chinnathambi,S., Nagy,P.L., Larsen-Haidle,J., Weigel,R.J. and Howe,J.R. (2010) Discovery of the *BMPR1A* promoter and germline mutations that cause juvenile polyposis. *Hum. Mol. Genet.*, **19**, 4654–4662.