Supporting Information

SI Materials and Methods

Tissue samples and nucleic acid isolation. De-identified snap-frozen and formalin-fixed and paraffin-embedded tumor and normal samples were collected according to the protocols approved by University of Pittsburgh Health Sciences Tissue Bank (HSTB) and by the University of Pittsburgh Institutional Review Board. No patient consent was required. We used 21 PTCs with no common driver mutations (35) for RNA-Seq analysis and 4 PTCs for WGS analysis. RNA and DNA were isolation as previously reported (35). Formalin-fixed and paraffin-embedded tissue samples were collected from 192 papillary thyroid carcinomas, which represented consecutive tumors diagnosed at the Department of Pathology, University of Pittsburgh and tested by ThyroSeq targeted next-generation sequencing assay from March 2014 until November 2015. The assay includes the analysis for point mutations in 14 genes and 42 types of gene fusions known to occur in thyroid cancer (1).

RNA-Seq and data analysis. Total RNA from tumor samples were processed to remove ribosomal RNA using the Ribozero Magnetic Gold kit (Illumina, Inc., San Diego, CA), followed by library preparation for RNA sequencing using the IlluminaTruSeq RNA Sample Preparation Kit v2. The prepared libraries were assessed using a Bioanalyzer and the High Sensitivity DNA kit (Agilent). Paired-end sequencing was performed on Illumina HiSeq2000 at the High Throughput Genome Center at the Department of Pathology, University of Pittsburgh. The sequence reads obtained were investigated for gene fusion events by ChimeraScan and deFuse programs. Briefly, the predicted fusion events from the 2 programs were integrated and combined with genomic annotation to generate a list of candidate gene fusions. The trimmed sequence reads were aligned to reference human genome (National Center for Biotechnology Information, build 37.1, hg19) and gene annotation databases (Ensembl genes v69 and UCSC genes hg19) for the analysis. In each program, the tumor and its matched normal thyroid sample were analyzed separately, and all fusion events in tumors that were detected in their matched normal samples were filtered out. To reduce false positive findings, the fusion events detected by both programs were further narrowed down by excluding 1) fusion events between adjacent genes, 2) fusion events with no supporting reads spanning the predicted breakpoints, 3) fusion events predicted to have five or more fusion partners and lacking specificity of target regions. RNA-Seq data were used to show the mRNA expression level for each exon of *THADA* and *IGF2BP3* genes. Reads were aligned with Tophat v2.0.13 to GRCh37, and expression levels in RPKM (reads per kilobase per million reads) for individual exons were quantified using RSeQC v2.

Whole genome sequencing and data analysis. Whole genome sequencing libraries were generated from genomic DNA using the Nextera DNA Sample Preparation Kit (Illumina) followed by AMPure XP purification (Beckman Coulter). The concentration of prepared libraries was determined using the High Sensitivity DNA kit (Agilent) on the Bioanalyzer (Agilent). The High-Throughput Genome Center at the Department of Pathology, University of Pittsburgh preformed paired-end sequencing on the Illumina HiSeq2000. Before analysis, sequences with low quality (base quality < 13) at both ends of reads were trimmed, and trimmed reads with less than 25 bp were removed (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). Trimmed sequences were aligned to reference genomes described above, using the Burrows-Wheeler Aligner (BWA). Each alignment was assigned a mapping quality score by BWA, the Phred-scaled probability that the alignment is incorrect. The PCR duplicates were detected and removed by Picard (http://picard.sourceforge.net). After alignment, we used VarScan to call

SNV for each chromosomal position. To minimize false positives, we also set minimum coverage as 8X in normal and 15X in tumor, minimum reads of variant allele as 4 and minimum proportion of variant allele as 15% in tumor. We also filter by minimum number of 2 reads supporting the variant allele per strand. For indels called from VarScan, we further remove them with any support reads from its matched normal control. The lists of SNVs/indels were annotated by using ANNOVAR. ANNOVA is software to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes. We first filtered SNVs/indels by various public databases including dbSNP (snp132) and 1000 genomes. After filtering, these novel SNVs/indels were subject to annotation using NCBI, UCSC and Ensembl databases. For SNVs, we also used SIFT software to predict whether an amino acid substitution affects protein function so that users can prioritize SNVs for further study (http://blocks.fhcrc.org/sift/SIFT.html). Structural variants were identified by the clipping reveals structure (CREST) algorithm with default parameters from paired end Illumina sequencing data. CREST is an algorithm that uses next-generation sequencing reads with partial alignments to a reference genome to directly map structural variations at the nucleotide level of resolution. Structural variants predicted by CREST were further reviewed by visual inspection in the Integrative Genomics Viewer.

PCR and Sanger Sequencing. The RNA from all samples was reverse transcribed by High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). PCR and RT-PCR were conducted using HotStarTaq DNA Polymerase (Qiagen). Quantitative RT-PCR analysis was performed with the QuantiTect SYBR Green PCR kit (Qiagen) on a 7500 Real-Time PCR System (ABI). mRNA expression levels were calculated by relative quantitation using the ABI 7500 Real-Time PCR SDS 2.0.5 software (Applied Biosystems) and the fold expression changes were determined by $2^{-\Delta\Delta CT}$ method. Expression of GAPDH was used as the internal control.

Fluorescence in situ hybridization. In order to confirm the presence of *THADA* and *IGF2BP3* fusion, hybridization was performed on tumor touch preparations using the probes generated from BAC clones RP11-204D19 labeled with SpectrumGreen (Vysis) for *THADA* and from BAC clone RP11-354G11, RP11-367B5, RP11-405L20, RP11-451F11, RP11-469O17 labeled with SpectrumOrange for the *IGF2BP3* region. To test for the presence of any rearrangements involving the *IGF2BP3* region in paraffin sections, the break-apart probes were used generated from BAC clones RP11-469O17, RP11-354G11 labeled with SpectrumGreen and RP11-451F11 clone labeled with SpectrumOrange. Microscopy was performed using Leica SP5 TCS 4D confocal microscope.

Immunohistochemistry. Immunohistochemistry was performed on 4-µm formalin-fixed and paraffin embedded tissues sections on a BenchMark Ultra instrument (Ventana).The primary anti-IGF2BP3 rabbit mouse monoclonal antibody (clone 69.1, Dako; 1:100) was used in according to the manufacturer's instructions.

RNA ISH. Manual RNA *in situ* hybridization ISH was performed in a single-plex format using the ViewRNA™ ISH Tissue Assay Kit (Cat# QVT0012, Affymetrix, Santa Clara, CA, USA). Briefly, FFPE sections were baked for 1 hr at 60°C and incubated in 1X pretreatment buffer at 90 to 95°C for 10 minutes followed by digestion with 1:100 dilution of protease at 40°C for 20 minutes. Tissue sections were subsequently hybridized with 1:40 dilution of the human IGF2

probe (Cat# VA1-12450), human IGF2BP3 (Cat# VA1-16156) for 2 hours at 40°C, followed by a series of washes. Slides were counterstained with hematoxylin. RNA integrity in the FFPE tissues was assessed using housekeeping gene, human UBC (Cat# VA1-10203). Bacillus Subtilis probe, DapB (Cat# VF1-11712) was used to determine background staining of the assay; Rat Kidney stained with Rat Arbp (Cat# VC1-10192) and DapB (VF1-11712) was used as assay controls.

Cell culture and chemicals. HEK293, U2OS and Caco-2 cells were obtained from ATCC, FTC133 cells from University of Colorado Cancer Center and PaCa-2 cells were kindly provided by Dr. Hou Wen (UPCI). Cell lines were maintained at 37°C and 5% CO2 in a humidified atmosphere and grown in DMEM (HEK293, U2OS, and PaCa-2), DMEM/F12 (FTC133), or EMEM (Caco-2). Purified IGF2 full length protein was purchased from Abcam (ab9575). IGF1R inhibitor OSI-906 was purchased from Selleck Chemicals (S1091). To treat cells with IGF2, purified IGF2 (Abcam) full length protein was added to cell cultures and PBS was used as mock control.

Plasmid and transfections. HEK293, U2OS, FTC133 were transiently transfected with pCMV6-Entry vector (PS100001) and IGF2BP3 expression vector (RC209597) (OriGene Technologies) using X-tremeGENE HP reagent (Roche Applied Science). PaCa-2 cells were transfected with IGF2BP3 siRNA (IGF2BP3 ON-TARGET plus SMART pool, acc. no. NM_006547) and scrambled control siRNA (ON-TARGET plus non-targeting pool) which were purchased from Dharmacon. Cells were transfected with 25 nM of IGF2BP3 siRNA or scramble control siRNA using DharmaFECT-1 transfection reagent (Dharmacon), harvested 24h after transfection for RNA extraction and 48h after transfection for protein isolation. All transfection experiments were conducted in quadruplicate for each condition.

Western Blotting analysis. Cells were lysed in RIPA buffer (Boston BioProducts) containing protease inhibitors (Roche) and phosphatase inhibitor cocktail (Sigma). Total protein was isolated from snap-frozen tissue homogenized in RIPA buffer using OMNI-GLH with hard tissue tips (OMNI International). Tissue and cell lysates were cold centrifuged at 12,000rmp for 30min and then the supernatants were collected. Protein concentration was determined using the BCA Protein Assay kit (Pierce) according to the manufacturer's protocol, and 40 µg of total protein was fractionated by SDS-PAGE (Lonza) and transferred to a polyvinylidene difluoride (PVDF) membrane (Millipore). The membrane were blocked in 5% bovine serum albumin (BSA) for 1 hour and then incubated overnight at 4 ˚C with the appropriate primary antibodies. Membranes were subsequently incubated with HRP-conjugated secondary antibody and signals were detected by ECL (GE Healthcare). The following primary and secondary antibodies were used: IMP3 (N-19) (1:400, Santa Cruz Biotechnology cat# sc-47893), IGF2 (1:1500 Abcam, cat# ab9574), p42/44 MAPK (1: 1000, Cell Signaling Technology cat# 9102), phospho-p42/44 MAPK (pT202/pY204) (1: 1000, Cell Signaling Technology cat# 9101), Akt (1: 1000, Cell Signaling Technology cat# 9272), phospho-Akt (Ser473) (1:1000, Cell Signaling Technology cat# 4060), phospho-S6 ribosomal protein (Ser235/236) (1:2500, Cell Signaling Technology cat# 4858), S6 Ribosomal Protein (5G10) (1:1000, Cell Signaling Technology cat# 2217), IGF-I Receptor β (1:1000, Cell Signaling Technology cat# 3018), phosphor-IGF-I Receptor β (Tyr1131)/Insulin Receptor β (Tyr1146) (1:250, Cell Signaling Technology, cat# 3021), Actin (1:1000, Sigma, cat# A5441), donkey anti-goat (1:5000, Santa Cruz Biotechnology, cat# sc-2020), anti-rabbit (1:10000, Promega cat# W401B).

Cell proliferation assay. Cells were transfected and 24 h later plated in 96-well plates. Alamar blue (Invitrogen) was added for 4h incubation. Afterward, the culture medium was collected and reduction of the Alamar blue was determined by fluorescence (excitation 560 nm, emission 595 nm). For growth inhibition assay, not transfected or transfected cells were plated in 96 well plates, and then the medium was replaced with medium containing the inhibitor OSI-906 at doses ranging from 0.1 μM to 5 μM or equal volume of vehicle (DMSO). Cells were treated for 72h and then cell proliferation was assayed using Alamar blue reagent.

Migration and invasion assays. For a wound-healing assay, cells were transfected for 24h in 6 well plates and then scraped using a 20 μL pipette tip. For inhibitor treatment, untransfected cells were scraped and then the medium was replaced with fresh medium containing OSI-906 at concentration of 1 μM or equal volume of vehicle (DMSO). For in vitro cell invasion assay, Biocat Matrigel Invasion chambers (354480, BD Biosciences) with 8-μm pore size and BD BioCoat cell culture inserts (354578; BD Biosciences) were used following the manufacturer's instructions. Briefly, 24-48h post-transfection the cells were harvested and added to the top chamber in serum-free medium. For the inhibitor treatment, untransfected cells were added to the top chamber in serum-free medium with 1 μM of OSI-906 or equal volume of vehicle (DMSO). The bottom chamber was filled with 10% FBS medium. The invaded cells were fixed using Quick-Diff stain kit (Dade Behring, Inc.) and counted.

Soft agar colony formation assay. Soft agar colony formation assay was performed by CytoSelect™ 96-Well Cell Transformation Assay Kit (Soft Agar Colony assay, Cell Biolabs, CBA-130) according to the manufacturer's instructions. Briefly, 24h post-transfection cells were harvested and plated in soft agar in a 96-well plate and cultured for 7 days, pictures were taken and cell colonies were counted manually. Then, the cells were solubilized, lysed, detected by CyQUANT® GR dye and the fluorescence was measured using a 485/530 nm filter. For treatment with OSI-906, cells were plated in soft agar and culture medium containing 1 μM OSI-906 was added to each well and cultured for 6 days. Cells treated with DMSO were used as a control.

Tumor growth inhibition in xenograft models. 1 x 107 of PaCa-2, FTC133 or Caco-2 cells were suspended in basement membrane extract (Cultrex Basement Membrane Extract, Type 3, PathClear). Cell suspensions were injected subcutaneously on flanks of 4 to 6-week-old female nude mice. When the tumors sizes reached ~500 mm3, the mice were randomized in two groups and treated by daily oral gavage with 50mg/kg OSI-906 or with equal volume of vehicle (30% PEG400/0.5% Tween80/5% propylene glycol). Tumor size was measured daily with a digital caliper. Animal handling and procedures were approved by the Division of Laboratory Animal Resources of the University of Pittsburgh.

Statistical analysis. For in vitro studies and RT-qPCR analysis, unless otherwise noted, each sample was assayed in quadruplicate and the numerical values are reported as means ± SD. For in vivo analysis, the data are presented as mean ± SEM. Pearson correlation coefficients were calculated to examine the relationships between methylation and expression gene. Statistical analyses were performed using GraphPad Prism 6 (GraphPad Software Inc., San Diego, CA). Student's t test was used to compare two groups independent, the statistical tests were two-tailed and the differences were considered statistically significant at p value <0.05.

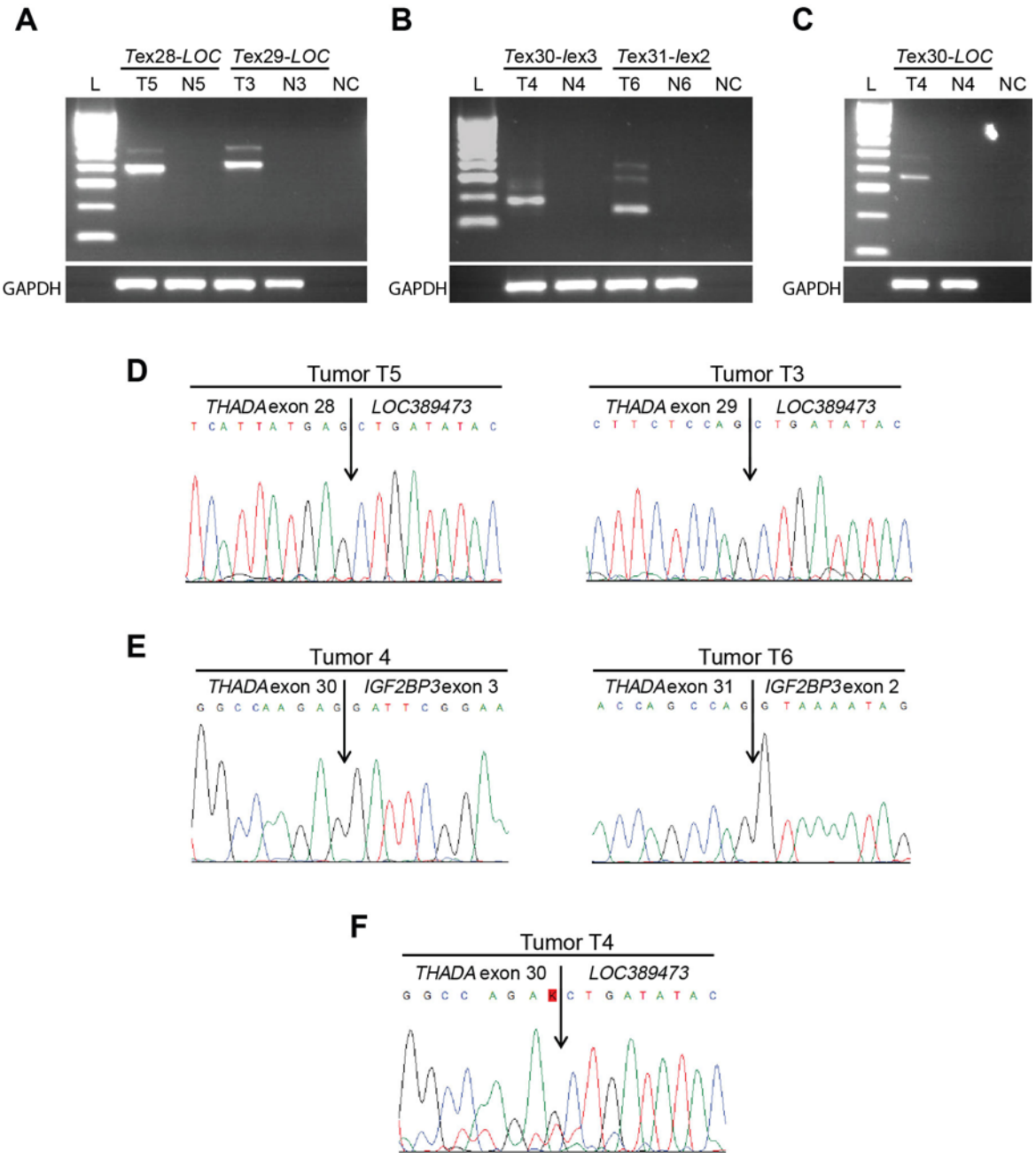

Fig. S1. Confirmation of the *THADA* gene fusions detected by RNA-Seq analysis. Results of RT-PCR and Sanger sequencing for tumors T5 and T3 positive for *THADA-LOC389473* (*A* and *D*) and T4 and T6 positive for *THADA-IGF2BP3* (*B* and *E*). In addition, manually inspection of fusion reads in tumor T4 identified several *THADA-LOC389473* reads confirmed by RT-PCR and Sanger sequencing (*C* and *F*). (L, 1-kb ladder; T, tumor tissue; N, normal tissue; NC, negative control; *T*, *THADA*; *I*, *IGF2BP3*; ex, exon; in, intron).

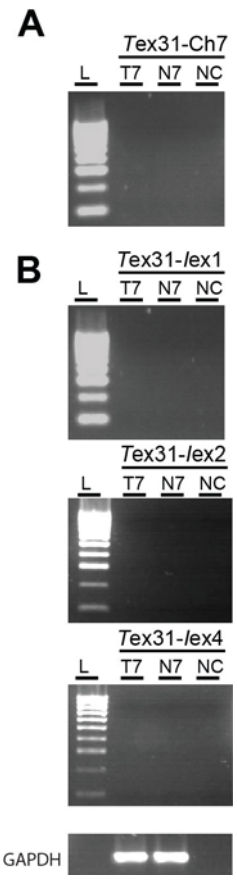

Fig. S2. Search for fusion mRNA in tumor T7. This tumor sample showed a fusion between *THADA* intron 31 and a region located 1.5 kb upstream of the *IGF2BP3* gene (chr. position 23,511,508) identified by WGS. RT-PCR analysis failed to detect fusion transcripts between (*A*) *THADA* exon 31 and the chr. position 23,511,508 or (*B*) between *THADA* exon 31 and exon 1, 2 or 3/4 of *IGF2BP3*. L, 1-kb ladder; T7, tumor T7; N7, matching normal tissue; NC, negative control; *T*, *THADA*; *I*, *IGF2BP3*; ex, exon; in, intron; Ch7, Ch7 23,511,508.

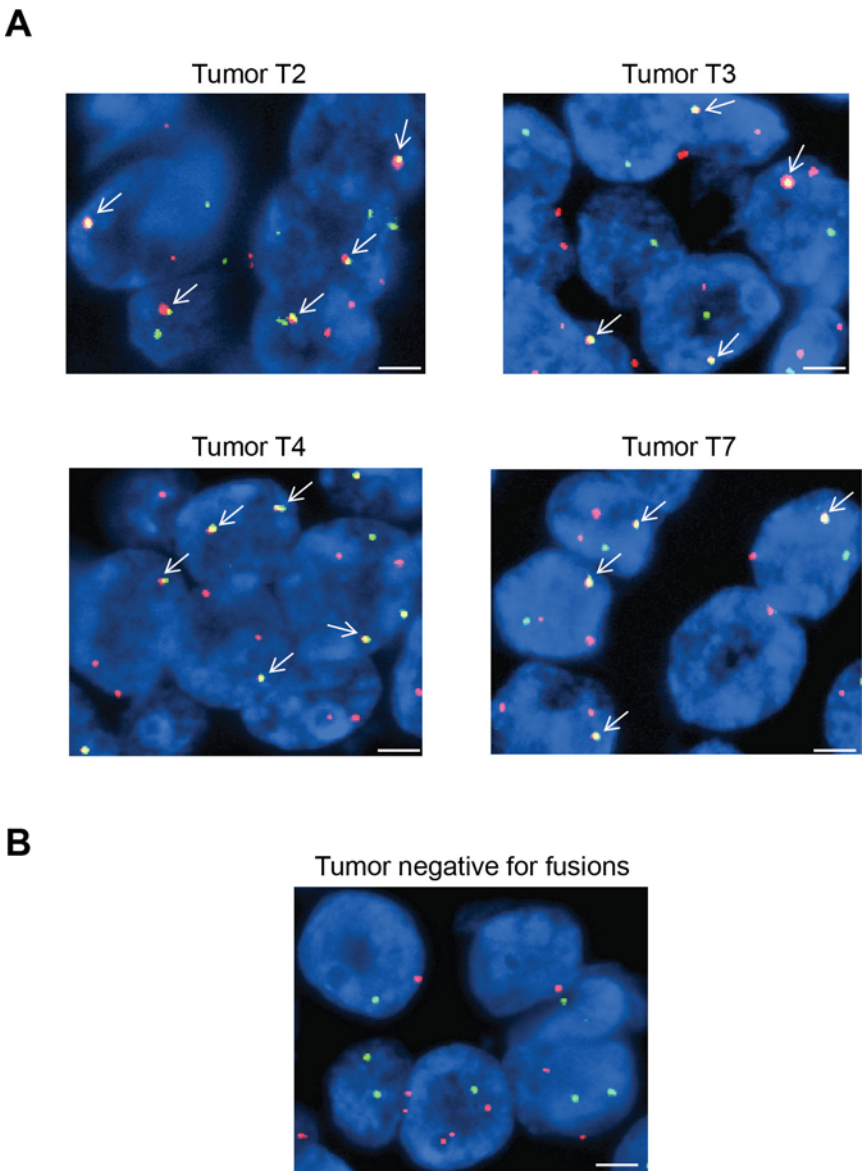

Fig. S3. Confirmation of fusions between the *THADA* region on 2p21 and *IGF2BP3* region on 7p15.3 by FISH. (*A*) Confirmation of the fusion in tumors T2, T3, T4 and T7 using tumor sections hybridized with DNA probes for *THADA* (red) and *IGF2BP3* (green). Fusion of the red and green signals is indicates by arrows. (*B*) Representative FISH image of thyroid tumor negative for *THADA* fusions. All cells show separate red and green signals. Scale bars, 5 μm.

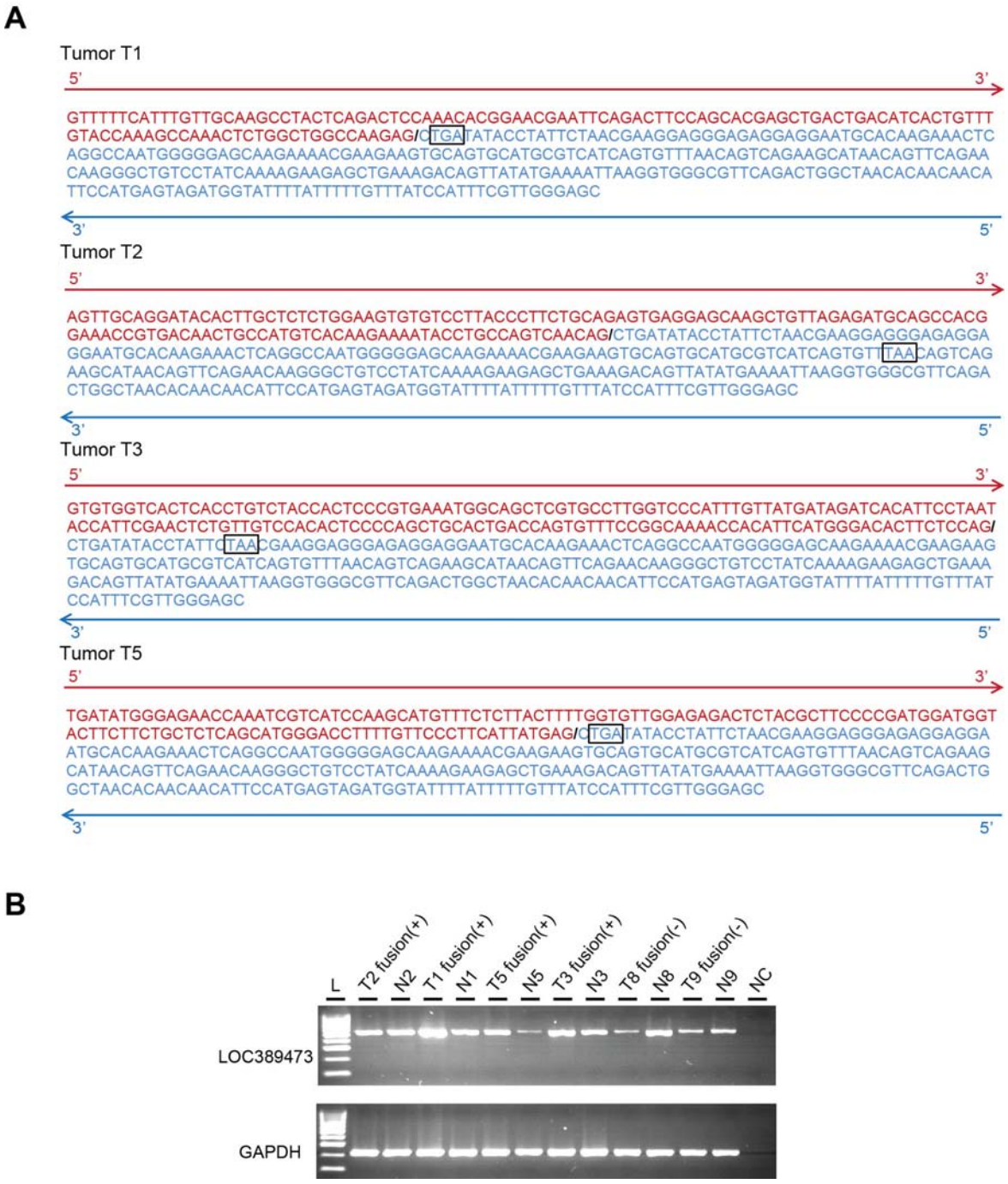

Fig. S4. Nucleotide sequence of the THADA-LOC389473 fusion transcripts and LOC389473 expression in thyroid tumors. (*A*) Sequence of the THADA-LOC389473 transcripts obtained by RT-PCR and Sanger sequencing, generated by the fusion between *THADA* (red; direction 5' to 3') and *LOC389473* (blue; direction 3' to 5') in the indicated tumor samples. Black rectangles indicate the stop codon generated downstream of the fusion. (*B*) *LOC389473* mRNA expression in tumors positive (+) and negative (-) for the fusion and corresponding normal thyroid tissues detected by RT-PCR. (L, 1-kb ladder; T, tumor tissue; N, normal tissue; NC, negative control)

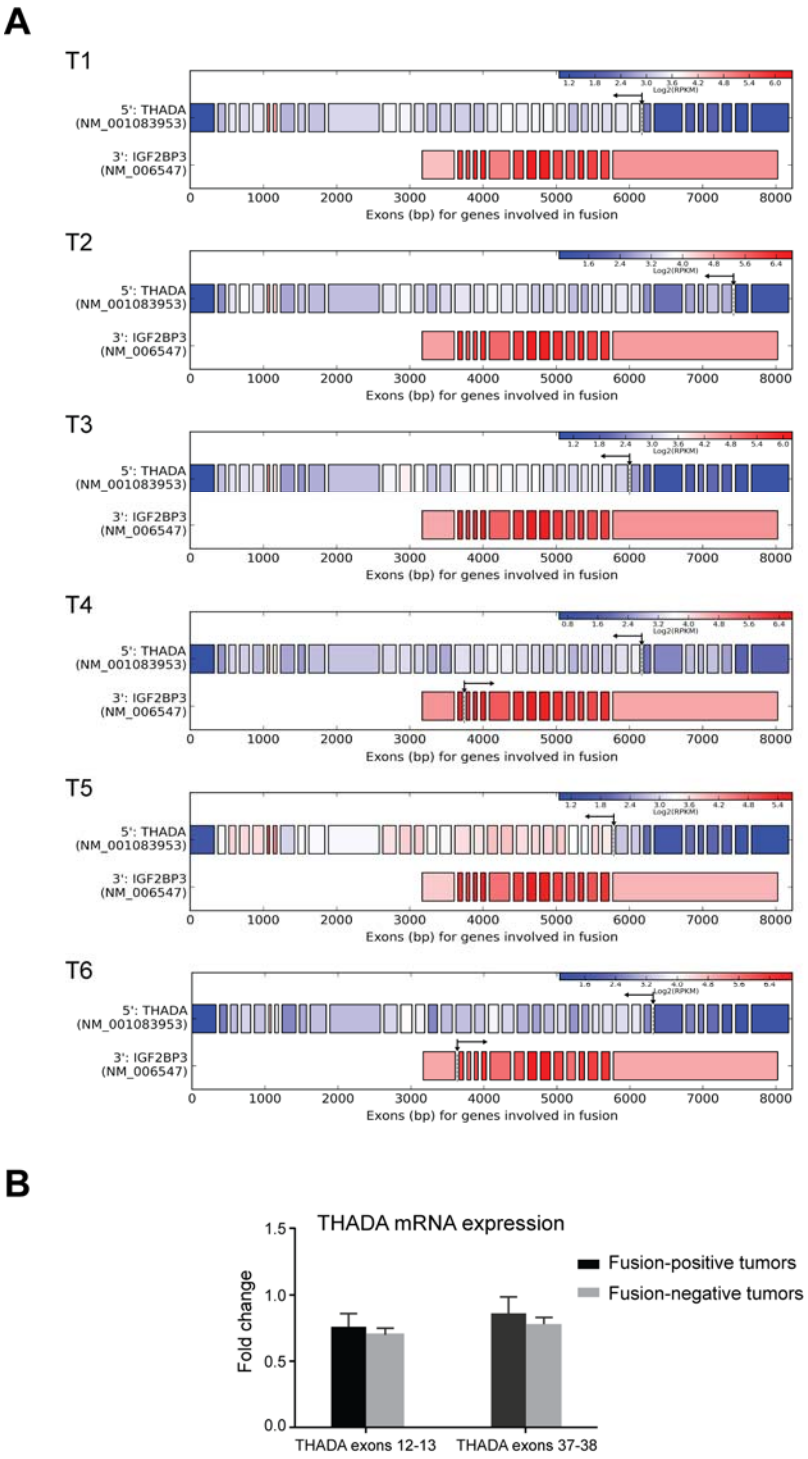

Fig. S5. Expression levels of THADA and IGF2BP3 mRNA in tumors positive for the fusion. (*A*) mRNA expression levels of individual exons of *THADA* and *IGF2BP3* genes based on the number of sequencing reads in the RNA-Seq data. No difference in the expression of *THADA* exons, before and after the break points (arrows) was seen. Similar findings were observed for *IGF2BP3*. All fusion-positive tumors showed increased levels of expression of all *IGF2BP3* exons, including tumors T4 and T6 with *THADA-IGF2BP3* fusion transcripts

identified by RNA-Seq. Each box represents an exon, and the boxes are colored according to the logarithm of their expression levels as measured in RPKM (reads per kb per million reads). (*B*) Comparison of *THADA* mRNA expression level (mean ± SD) measured by RT-qPCR for the gene region upstream (exons 12-13) and downstream (exons 31-32) of the fusion in fusion-positive tumors (n=5) and fusion-negative (n=4) tumors. Fold change of 1 represents the baseline of normal thyroid (n=5).

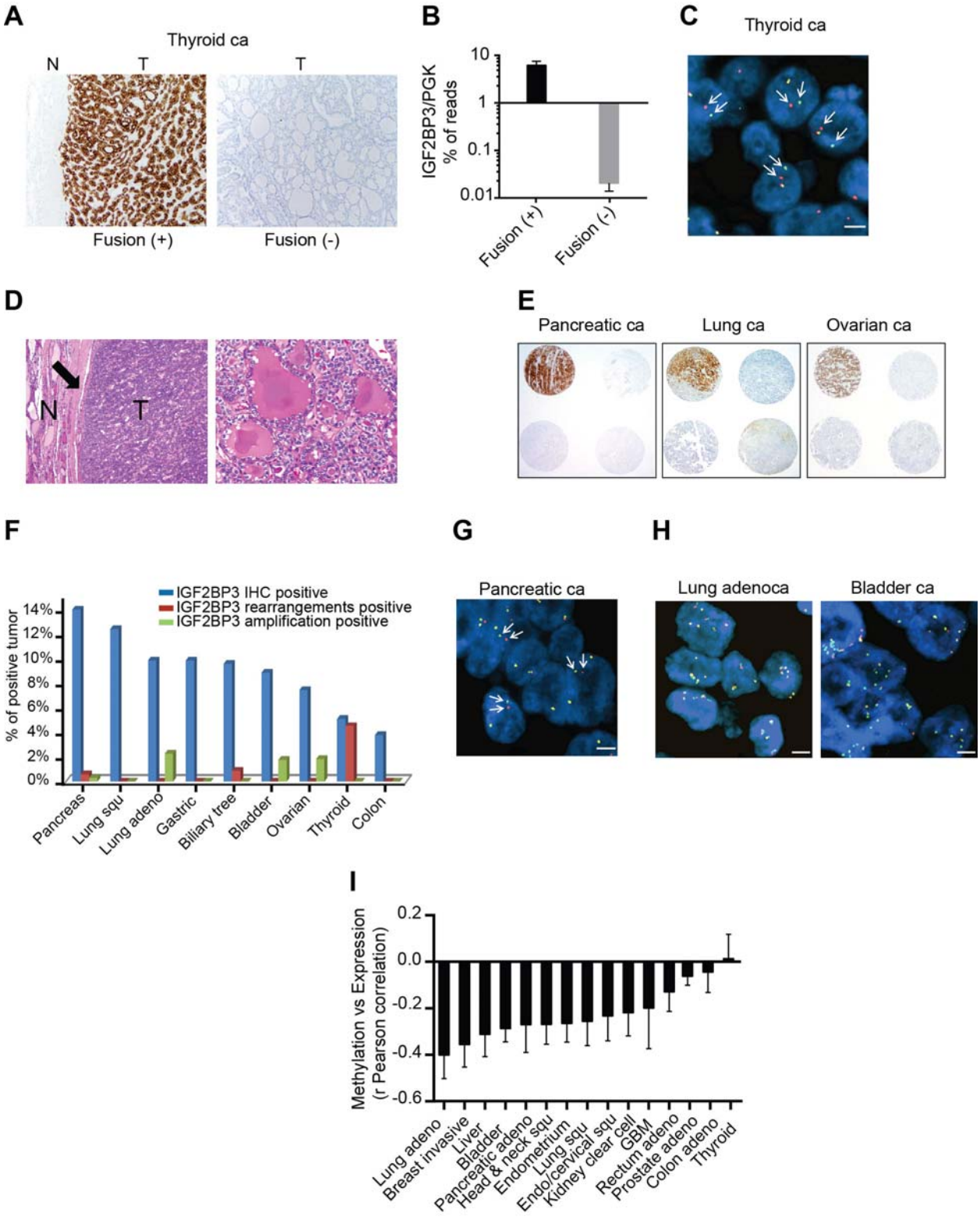

Fig. S6. Frequency of IGF2BP3 overexpression in thyroid and other cancers. (*A*) Representative images of immunostaining of thyroid cancers positive and negative for *THADA-LOC389473* fusion showing strong cytoplasmic immunoreactivity with anti-IGF2BP3 antibody

in the fusion-positive tumor. (*B*) Expression of IGF2BP3 mRNA detected by targeted NGS shown as percentage of reads of IGF2BP3/PGK in thyroid tumors positive (n=5) and negative (n=20) for *THADA* fusions. The y-axis values are shown in logarithmic scale. (*C*) Representative FISH results of thyroid tumor sections hybridized with a split-apart DNA probe for the *IGF2BP3* region. Many cells show a split of the red and green signals (arrows) indicative of a rearrangement involving the *IGF2BP3* region. Scale bars, 5 μm. (*D*) Microscopic appearance of the fusion-positive thyroid cancer. Left: low-power (40x) view showing a capsule (arrow) clearly demarcating the tumor (T) from normal thyroid (N). Right: high-power view (200x) showing the follicular architecture and tumor cells with nuclear features of papillary cancer. (E) Representative images of immunostaining with anti-IGF2BP3 antibody of indicated cancer microarrays showing in each array one tumor (*top left*) with strong diffuse immunoreactivity and 3 other tumors with complete lack of immunoreactivity. (*F*) Percentage of cancers with strong expression of IGF2BP3 by immunohistochemistry (IHC) and estimated percentage of cancers positive for rearrangement and amplification of the *IGF2BP3* locus based on the FISH analysis of tumors positive by immunohistochemistry. (*G-H*) Representative FISH results of tumor sections hybridized with a split-apart DNA probe for the *IGF2BP3* region. Scale bars, 5 μm. (*G*) Pancreatic tumor cells showing a split of the red and green signals (arrows) indicative of a rearrangement involving the *IGF2BP3* region. (*H*) Cells of indicated tumor types showing 4 or more fused red and green signals consistent with amplification of the intact (non-rearranged) *IGF2BP3* region. (*I*) The Pearson correlation (r) between *IGF2BP3* DNA methylation and mRNA expression in 15 major cancer types. The results are based on Illumina Infinium Human DNA Methylation 450 and HiSeq 2000 RNA Sequencing v.2 datasets from TCGA. Data shown as mean ± SD of the Pearson correlation (r) between individual DNA methylation sites of *IGF2BP3* promoter or first exon with gene expression. The results of the analysis showed inverse correlation between methylation and IGF2BP3 expression in many studied cancer types, whereas the methylation pattern for thyroid cancer was distinct from other cancers and lacked such correlation.

Table S1. Thyroid tumor samples with *THADA* fusions identified by RNA-Seq or WGS analysis

Tumor sample	Detection Mode	Fusion partners	
		5'	3'
T1	RNAseq	*THADA* exon 30	*LOC389473*
T2	RNAseq	*THADA* exon 36	*LOC389473*
T3	RNAseq	*THADA* exon 29	*LOC389473*
T4	RNAseq	*THADA* exon 30	*IGF2BP3* exon 3*
T5	RNAseq	*THADA* exon 28	*LOC389473*
T6	RNAseq	*THADA* exon 31	*IGF2BP3* exon 2
T7	WGS	*THADA* intron 31	1.5 kb upstream of *IGF2BP3*

*Manual inspection of RNA-Seq reads for this tumor also identified single fusion reads between *THADA* exon 30 and *LOC389473* .

Table S2. Frequency of IGF2BP3 immunoreactivity and the *IGF2BP3* gene locus rearrangement and amplification detected by FISH in different cancer types.

Cancer types (n=cases)	IGF2BP3 IHC	IGF2BP3 rearrangements (among IHC-positive tumors)	IGF2BP3 Amplification (among IHC-positive tumors)
Pancreas (n=331)	47/331 (14.2%)	2/8 (25%)	1/8 (12.5%)
Lung squamous cell (n=119)	15/119 (12.6%)	0/5	0/5
Lung adenocarcinoma (n=129)	13/129 (10.1%)	0/18	3/18 (16.7%)
Gastric (n=30)	3/30 (10%)		
Biliary tree (n=113)	11/113 (9.7%)	1/4 (25%)	0/3
Bladder (n=55)	5/55 (9.1%)	0/5	1/5 (20%)
Ovarian (n=53)	4/53 (7.5%)	0/4	1/4 (25%)
Thyroid (n=192)	10/192 (5.2%)	9/10 (90%)	0/5
Colon (=78)	3/78 (3.8%)	0/3	0/3

Percentage of tumors positive for IGF2BP3 expression, rearrangement and amplification based on the total number of cases analyzed for each tumor type.

Table S3. Frequency of the IGF2BP3 gene amplification in different cancer types in the TCGA database.

Tumor types (n=cases)	Frequency of IGF2BP3 amplification
Esophagus (n=1840)	13/1840 (7.1%)
Endometrium (n=56)	4/56 (7.1%)
Bladder (n=128)	6/128 (4.7%)
Sarcoma (n=256)	11/256 (4.3%)
Stomach (n=293)	12/293 (4.1%)
Lung adeno (n=230)	9/230 (3.9%)
ACC (n=90)	2/90 (2.2%)
Melanoma (n=366)	8/366 (2.2%)
Prostate (n=333)	7/333 (2.1%)
Testicular germ cell (n=150)	3/150 (2%)
Ovarian (n=489)	9/489 (1.8%)
Breast (n=778)	14/778 (1.8%)
Uterine (n=363)	6/363 (1.7%)
Lung squ (n=178)	3/178 (1.7%)
Pancreas (n=184)	3/184 (1.6%)
GBM (n=563)	7/563 (1.2%)
Liver (n=370)	4/370 (1.1%)
Head & neck (n=279)	3/279 (1.1%)
LGG-GBM (n=1084)	7/1084 (0.6%)
AML (n=191)	1/191 (0.5%)
Colon adeno (n=615)	3/615 (0.5%)
Cervical (n=295)	1/295 (0.3%)
ccRcc (n=436)	1/436 (0.1%)
Thyroid (n=499)	0/499 (0%)

List of the tumor types and number of cases analyzed for IGF2BP3 amplification reported by The Cancer Genome Atlas. The number of cases positive is showed as percentage frequency of IGF2BP3 amplification. ACC, Adrenocortical Carcinoma; Lung squ, Lung squamous; GBM, Glioblastoma; LGG-GBM, Merged Cohort of diffuse low grade glioma and glioblastoma; ccRcc, Kidney Renal Clear Cell Carcinoma.