Published online 22 May 2017

Nucleic Acids Research, 2017, Vol. 45, No. 14 ***8269–8281***
doi: 10.1093/nar/gkx435

A novel isoform of TET1 that lacks a CXXC domain is overexpressed in cancer

Charly R. Good, Jozef Madzo, Bela Patel, Shinji Maegawa, Nora Engel, Jaroslav Jelinek and Jean-Pierre J. Issa*

Fels Institute for Cancer Research and Molecular Biology, Lewis Katz School of Medicine at Temple University, Philadelphia, PA 19140, USA

Received January 13, 2017; Revised April 14, 2017; Editorial Decision April 26, 2017; Accepted May 03, 2017

ABSTRACT

TET1 oxidizes methylated cytosine into 5-hydroxymethylcytosine (5hmC), resulting in regulation of DNA methylation and gene expression. Full length TET1 (TET1^{FL}) has a CXXC domain that binds to unmethylated CpG islands (CGIs). This CXXC domain allows TET1 to protect CGIs from aberrant methylation, but it also limits its ability to regulate genes outside of CGIs. Here, we report a novel isoform of TET1 (TET1^{ALT}) that has a unique transcription start site from an alternate promoter in intron 2, yielding a protein with a unique translation start site. Importantly, TET1^{ALT} lacks the CXXC domain but retains the catalytic domain. TET1^{ALT} is repressed in embryonic stem cells (ESCs) but becomes activated in embryonic and adult tissues while TET1^{FL} is expressed in ESCs, but repressed in adult tissues. Overexpression of TET1^{ALT} shows production of 5hmC with distinct (and weaker) effects on DNA methylation or gene expression when compared to TET1^{FL}. TET1^{ALT} is aberrantly activated in multiple cancer types including breast, uterine and glioblastoma, and TET1 activation is associated with a worse overall survival in breast, uterine and ovarian cancers. Our data suggest that the predominantly activated isoform of TET1 in cancer cells does not protect from CGI methylation and likely mediates dynamic site-specific demethylation outside of CGIs.

INTRODUCTION

DNA methylation is a biochemical modification that occurs after DNA replication, primarily on cytosine when followed by guanosine (CpG sites). It is an important epigenetic mark that has been implicated in gene regulation, early embryonic development, genomic imprinting, X chromosome inactivation and cancer (1,2). Cancer cells often display aberrant DNA methylation, where both hypomethylation and hypermethylation can be found at important cancer-associated genes (3–6). Methylation of cytosine is initially established by DNMT3A and DNMT3B, and then maintained by the maintenance methyltransferase DNMT1 (7,8). Recently, the TET enzymes (TET1, TET2, TET3) were identified as DNA demethylases that convert 5-methylcytosine (5mC) into 5-hydroxymethylcytosine (5hmC), which can then be further oxidized or converted to unmethylated cytosine (9–12). The TET enzymes display tissue-specific expression, where TET1 is primarily expressed in embryonic stem cells (ESCs) but has been shown to be expressed at low levels in select adult tissues (13–15). Although it was reported that loss of TET1 in ESCs leads to defects in differentiation and self-renewal, these findings were shown to be an artifact of the shRNA used in the study (16,17). Indeed, loss of TET1 in ESCs does not affect pluripotency or embryonic development, and TET1 knockout (KO) mice remain viable and fertile (13). TET1 has also been implicated as a transcriptional co-activator and co-repressor, where its effects on transcription can be both dependent and independent of its demethylase activity (17–19).

TET1 and TET3 proteins have CXXC domains, which recognize and bind to unmethylated stretches of CpGs (CpG islands or CGIs) (20). We previously showed that TET1 binds to CGIs and protects them from gaining aberrant methylation, as overexpression of TET1 leads to the accumulation of 5hmC at the borders of CGIs while loss of TET1 leads to increased DNA methylation specifically at CpG sites within and around CGIs (19). However, other studies have implicated TET1 in the dynamic regulation of DNA methylation outside of CGIs (21), which is incongruous with the constraints on TET1 function imparted by the CXXC domain. In studying TET1 function, we discovered a new transcription start site (TSS) that directs expression of a truncated TET1 (TET1^{ALT}) that lacks the CXXC domain.

*To whom correspondence should be addressed. Tel: +1 215 707 4307; Fax: +1 215 707 1454; Email: jpissa@temple.edu
Present address: Jean-Pierre J. Issa, Fels Institute for Cancer Research and Molecular Biology, Lewis Katz School of Medicine at Temple University, Philadelphia, PA 19140, USA.

We found that TET1^{ALT} is overexpressed in cancer but, *in vitro*, it has minor effects on CGI methylation, consistent with the lack of the CXXC domain. TET1^{ALT} may account for dynamic regulation of DNA methylation by TET1, and its minimal effects on CGIs explains the paradoxical co-occurrence of TET1^{ALT} overexpression and CGI hypermethylation in cancer.

MATERIALS AND METHODS

Cell culture

Breast cancer cell lines (BT549, HCC2218, HCC1599, MCF-7), immortalized breast, but non-malignant (MCF10A), human embryonic kidney (HEK293T), chronic myelogenous leukemia (K562) and the PC-3 prostate cancer cell line were all obtained from American Type Culture Collection (ATCC). Normal breast epithelium (NBE) was a kind gift from Dr Xiaowei Chen at the Fox Chase Cancer Center. The immortalized human mammary epithelial cells (HMLE) were a generous gift from Dr Sendurai A. Mani at the University of Texas MD Anderson Cancer Center. The lymphoblastoid cell line (GM12878) was a kind gift from Dr Italo Tempera at the Fels Institute for Cancer Research and Molecular Biology. GM12878 were cultured in Roswell Park Memorial Institute (RPMI) 1640 media, 2 mM L-glutamine with 15% fetal bovine serum (FBS). K562 cells were cultured in Iscove's Modified Dulbecco's Medium, supplemented with 10% FBS. PC-3 cells were cultured in ATCC-formulated F-12K Medium with 10% FBS. To culture MCF10A cells, we used DMEM/F12 with 5% horse serum (treated to remove divalent cations), 20 ng/ml Epidermal Growth Factor (EGF), 100 ng/ml cholera toxin, 10 μg/ml insulin, 500 ng/ml hydrocortisone, 1.39 mM calcium and 1% penicillin–streptomycin antibiotics. HMLE cells were grown as previously described (22). MCF-7 cells were cultured in Eagle's Minimum Essential Medium with 10% FBS. HEK293T cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% FBS. HCC1599, BT549 and HCC2218 were cultured in RPMI-1640 Medium with 10% FBS. All cell lines routinely tested negative for mycoplasma contamination.

Western blot and DNA slot blot assay

Protein was extracted using a lysis buffer consisting of (50 mM Tris–HCl (pH 7.4), 5 mM ethylenediaminetetraacetic acid, 250 mM NaCl, 50 mM NaF, 0.1% Triton X-100, 0.1 mM Na_3VO_4) supplemented with 1× protease inhibitor cocktail solution (Roche). Extracts were quantified using Qubit protein assay (ThermoFisher) and were run on polyacrylamide gels. Gels were transferred to polyvinylidene difluoride (PVDF) membranes using a wet transfer method in 3-(Cyclohexylamino)-1-propanesulfonic acid (CAPS) buffer. Primary antibodies were incubated overnight at 4°C. The following primary antibodies were used in this study: anti-FLAG (A8592, Sigma), anti-TET1 (GTX124207, GeneTex), anti-TET1 (GT1462, Sigma), anti-5hmC (catalog # 39769, Active Motif) and anti-β-actin (A5316, Sigma). For the DNA slot blot analysis, we followed the protocol established previously with the exception of using a slot blot apparatus instead of a dot blot (19). Blots were imaged using FluorChem Q and unsaturated bands were quantified using the multiplex band analysis tool followed by normalizing to local background and β-actin.

TET1 overexpression

The TET1^{FL} pIRES hrGFP II expression plasmid along with the TET1^{CD} plasmid was generated in our lab previously (19). To generate the TET1^{ALT} plasmid, TET1^{FL} plasmid was digested with restriction enzymes BamHI and BglII, which removed the first 1850 bp of the TET1^{FL} cDNA sequence thus excluding the TET1^{FL} start codon. One hundred and sixty three base pairs of the TET1^{FL} cDNA is upstream of the TET1^{ALT} ATG; however, no in-frame ATG is contained within this region. The plasmids were transfected into HEK293T cells using Lipofectamine 2000 following the manufacturer's instructions.

Digital restriction enzyme analysis of methylation (DREAM)

Digital restriction enzyme analysis of methylation (DREAM) is a quantitative, deep sequencing based method for DNA methylation analysis and it was performed as previously described (19,23). Briefly, 2 μg of genomic DNA from HEK293T cells expressing empty vector, TET1^{ALT} or TET1^{FL} were digested with 20 units of SmaI (8 h at 25°C, NEB) and 20 units of XmaI (∼16 h at 37°C, NEB), resulting in a distinct DNA methylation signature at CCCGGG sites. 3′ ends of the DNA fragments were repaired using Klenow fragment (3′→5′ exo-) DNA polymerase and deoxycytidine triphosphate (dCTP), deoxyguanosine triphosphate (dGTP) and deoxyadenosine triphosphate (dATP) nucleotides. Illumina sequencing adapters were ligated to the DNA fragments and the libraries were sequenced by paired-end 40 nt sequencing on Illumina HiSeq2500. The sequencing reads were mapped to the hg19 genome and methylation values were calculated as the ratio of the number of the reads with the methylated XmaI signature over the total number of tags mapped to a given SmaI/XmaI site. The coverage threshold was set to >50 reads per sample. TET1^{ALT} and TET1^{FL} were compared to empty vector control and data were filtered for sites that change methylation by >5%. High-throughput DNA methylation data generated in this study have been deposited in the GEO database under accession number GSE93617.

RNA-seq

RNA was isolated using Qiagen's RNeasy Plus Mini kit following manufacturer's instructions from experiments done in biological triplicates. RNA was quantified by Nanodrop and purity was checked using the Agilent Bioanalyzer. Strand-specific RNA libraries were generated from 1 μg of RNA using TruSeq stranded total RNA with Ribo-Zero Gold (Illumina). Sequencing was performed using single end reads (50 bp, average 30 million reads per sample) on the HiSeq2500 platform (Illumina). Sequenced reads were aligned to the hg19 genome assembly using TopHat2 software suite (24). The expression level and fold change of each

treatment group were evaluated using Cuffdiff (25). Genes that had zero reads across all samples were excluded. Genes were considered differentially upregulated with fold change (FC) > 1.5 and downregulated with FC < 0.66 and $P < 0.05$. High-throughput RNA-sequencing data generated in this study have been deposited in the GEO database under accession number GSE93619.

RNA isolation and qPCR analysis (cell lines)

RNA isolation, reverse transcription and reverse transcriptase-polymerase chain reaction (RT-PCR) analysis were performed as described previously (19). Total RNA (in biological triplicates) was extracted using TRIzol following manufacturer's protocol. RNA was DNase-treated using TURBO DNA-free kit following manufacturer's protocol (Ambion). Complementary DNA (cDNA) was synthesized using 1 μg of RNA using the High-Capacity cDNA reverse transcription kit following manufacturer's protocol (Applied Biosystems). qPCR was performed on Applied Biosystems 7500 machine using iTaq Universal SYBR Green Super Mix following manufacturer's instructions (BioRad). qPCR was performed in technical and biological triplicates and the average C_t values were determined for each gene. Samples were normalized to the housekeeping gene HPRT. The primers used are listed in Supplementary Table S1.

Luciferase reporter assay

Luciferase reporter assays were conducted using the Dual-Luciferase Reporter Assay System (Promega) following the manufacturer's protocol. HEK293T cells were seeded in 6-well plates and co-transfected with Renilla expression plasmid and the pGL4.10 [luc2] reporter constructs containing either empty vector, intron control or the $TET1^{ALT}$ promoter region. Cells were transfected using Lipofectamine 2000. Luciferase activities were measured 48 h post transfection and normalized to Renilla and to empty vector control. The primers used are listed in Supplementary Table S2.

Mouse tissues

Adult tissues were harvested from C57BL/6 male mice 6.2 months after birth. RNA was extracted and DNase treated as described above under RNA isolation.

To obtain staged embryonic tissues from C57BL/6 mice, matings were set up and plugs checked the following morning. Noon of the day of the plug was designated E0.5. Pregnant dams were killed at the appropriate day of gestation (10.5, 12.5, 14.5, 16.5 dpc) and fetal tissues were dissected and frozen for further analysis. Neonatal tissues were obtained from day 3 pups. C57BL/6 mouse ESCs were maintained in leukemia inhibitory factor (LIF) on inactivated mouse embryonic fibroblasts (MEFs), on gelatin-coated tissue culture plates in a 5% CO_2 humid incubator at 37°C. RNA extraction was performed following two consecutive 60 min MEF-depletion steps. RNA from ESCs, fetal and neonatal tissues was extracted using the Roche High Pure RNA Isolation Kit (#11828665001), following the manufacturer's protocol. All RNA samples were subjected to DNAse treatment using Turbo DNA-free (Ambion #AM1907). Three to five biological samples were collected for each tissue.

CRISPR

To knockout TET1, we used the Lenti CRISPR V2 plasmid (Addgene) (26). CRISPR gRNAs were designed using http://crispr.mit.edu/ to target a common exon between $TET1^{FL}$ and $TET1^{ALT}$. gRNAs can be found in Supplementary Table S3. Oligonucleotides were annealed and ligated into the Lenti CRISPR V2 plasmid that was previously digested with BsmBI. The cloning protocol associated with the plasmid was followed: http://genome-engineering.org/gecko/wp-content/uploads/2013/12/lentiCRISPRv2-and-lentiGuide-oligo-cloning-protocol.pdf (26). Ligated plasmids were propagated and verified by restriction enzyme digest and by sequencing. Lentiviruses were generated using HEK293T cells by transfecting with packaging plasmid (psPAX2, Addgene 12260), envelope plasmid (pMD2.G, Addgene 12259) and lenti-CRISPR V2 plasmid (Addgene 52961). Cells were transfected using Lipofectamine 2000. Viral supernatant was collected at 48 and 72 h, filtered with 0.45 μm membrane and incubated on MDA-MB-231 cells for 7 h in the presence of polybrene (6 μg/ml, Millipore). Following transduction, we performed antibiotic selection for 3 days using puromycin (1 μg/ml), followed by single cell cloning using serial dilution in 96-well plates. After selection, cells were maintained in normal media supplemented with 0.25 μg/ml puromycin.

To activate the $TET1^{ALT}$ and $TET1^{FL}$ promoter, we used plasmids previously generated following published protocols (27). Plasmids used included pLKO.1-puro U6 sgRNA CAG (Addgene 50927), pLKO.1-puro U6 sgRNA BfuAI stuffer (Addgene 50920) and pHAGE EF1α dCas9–VP64 (Addgene 50918). sgRNA's were designed as described above, but target regions were limited to the promoter regions of $TET1^{ALT}$ and $TET1^{FL}$. Promoters were identified using UCSC genome browser by locating the TSS, RNA polymerase II binding and $H3K4me^3$ enrichment. Two gRNAs were simultaneously used to target the $TET1^{FL}$ promoter and three gRNAs to target the $TET1^{ALT}$ promoter. gRNAs used can be found in Supplementary Table S4. An sgRNA targeting the CAG (CMV-IE, chicken actin, rabbit beta-globin) promoter was used as an off-target control. First, stable cell lines were generated that overexpress the dCas9–VP64 fusion protein. Lentiviruses were made in HEK293T cells, and then MCF10A cells were transduced and puromycin selected for 1 week. Next, lentiviruses were generated in HEK293T cells using the pLKO.1 gRNA plasmid mentioned above. After viral collection, MCF10A-dCas9–VP64 expressing cell lines were transduced with the viral containing gRNAs. Cells were selected for 3 days using puromycin (1 μg/ml).

Bisulfite pyrosequencing

Genomic DNA from HEK293T cells expressing either empty vector, $TET1^{CD}$, $TET1^{ALT}$ or $TET1^{FL}$ was bisulfite converted using the EpiTect bisulfite kit (Qiagen) following

the manufacturer's protocol. For bisulfite pyrosequencing of LINE1, a polymerase chain reaction (PCR) for amplification was used as previously described (28). Pyro Q-CpG Software (Qiagen) was used to analyze the data. PCR and pyrosequencing primers can be found in Supplementary Table S5.

PCR and clonal sequencing

Genomic DNA from MDA-MB-231 TET1 CRISPR knockout cells was amplified using primers surrounding the gRNA target sequencing and Phusion High-Fidelity DNA polymerase (Neb). Amplified PCR products were cloned into the Zero Blunt TOPO pCR4 system and OneShot Top10 chemically competent *Escherichia coli* were transformed following the manufacturer's protocol. (ThermoFisher). Twelve clones were sequenced and analyzed using Serial Cloner. PCR and sequencing primers can be found in Supplementary Table S6.

Bioinformatics and statistics

DNA methylation data (450K array) for normal breast tissue ($N = 41$) were accessed from The Cancer Genome Atlas public data portal in 2014. DNA methylation data (450K array) from 63 human cell lines were downloaded from the UCSC genome browser track HAIB Methyl450 (wgEncodeHaibMethyl450) as part of the Encode project (29). The methylation values for each CpG site associated with TET1 ($N = 30$) were averaged for all normal breast samples and for all cell lines and plotted versus the distance to the TET1 TSS.

RNA-sequencing BAM files were downloaded from the Genomic Data Commons Portal for breast, ovarian and uterine cancer, acute myeloid leukemia (AML) and glioblastoma patient samples. Access was granted by the NCI Center for Biomedical Informatics and Information Technology to obtain The Cancer Genome Atlas (TCGA) controlled access datasets. Reads mapping to the canonical or alternate TET1 exons were extracted and normalized to all reads on chromosome 10.

To illustrate the exon usage of TET1 in selected cancers, we plotted median length normalized exon quantification based on RNA-seq level 3 TCGA data in breast and uterine cancer, glioblastoma and AML. To exclude subjects that do not express TET1 at sufficient levels, we filtered out the samples with average usages of exon 3–12 below a 0.5 cut off.

Survival curves were generated using GraphPad Prism 4.0. Survival data were downloaded from CBioPortal using the following studies: Breast cancer (METABRIC, Nature 2012 & Nature Communications 2016) (30,31), Glioblastoma (TCGA, Provisional), Uterine cancer (TCGA, Nature 2013) (32), Ovarian Serous Cystadenocarcinoma (TCGA, Provisional), Acute Myeloid Leukemia (TCGA, NEJM 2013) (33). TET1 high was considered >1 standard deviation (SD) above the mean and all other patients were classified as TET1 low for uterine cancer (RNA-seq), AML (RNA-seq), glioblastoma (Microarray) and ovarian cancer (RNA-seq). TET1 high was considered SD > 2 above the mean for breast cancer (Microarray U133).

Calculations were done in GraphPad. The Student's *t*-test was used to calculate significant *P* values unless otherwise stated. All *P*-values are two-sided; $*P < 0.05$, $**P < 0.01$, $***P < 0.001$ denotes significance. Mann–Whitney test was used to test significance of $TET1^{ALT}$ exon reads *in vivo*. Error bars are standard error of the mean (SEM). Significance of survival curves were calculated using the Gehan–Breslow–Wilcoxon test. Hierarchical clustering and principal component analysis (PCA) were performed using R software.

RESULTS

Identification of a novel TET1 isoform

NCBI lists one TET1 gene (NM_030625.2), with no alternate isoforms. It has been reported that the canonical TET1 promoter can be hypermethylated in cancer (15). In examining TET1 methylation using publicly available data from TCGA (http://cancergenome.nih.gov) and ENCODE, we found that the TET1 TSS, which is in a CGI, is occasionally hypermethylated in cancer. However, we noted a CpG site (which is not in a CGI) in intron 2 of TET1 that was unmethylated in 62/63 cell lines and in all 41 normal breast tissue samples surveyed (Figure 1A). Given that most CpG sites outside of CGIs are highly methylated unless they are in a regulatory region, we examined this genomic area in greater detail.

This region containing the unmethylated CpG site aligned with the start site of two expressed sequence tags (ESTs), which we called $TET1^{ALT}$ exon 1a and 1b and which spliced into TET1 exon 3 (Figure 1B). This suggested the presence of an alternate TSS. Indeed, the 1 kb region upstream of TET1 exon 3 is highly conserved in primates and placental mammals but not among more distant vertebrates such as chicken and zebrafish (Figure 1B). Of note, other TET1 exons are conserved in these species (Supplementary Figure S1) suggesting more recent evolution for this region. We next queried Poly A (+) CAGE data (which identify TSSs) (34) and found peaks in multiple cell lines, including MCF-7, GM12878 and IMR90 in the $TET1^{ALT}$ putative promoter region (Figure 1B). We then put the predicted $TET1^{ALT}$ sequence into open reading frame finder (35) and found two possible start codons in exon 4, which would encode proteins with molecular weights of 147 and 162 kilodaltons (kDa). Both alternate start codons (ATG/Methionine) have moderate to strong Kozak sequences that are amenable to translation initiation (Supplementary Figure S2A). Aligning the predicted amino acid sequences of the alternate proteins to the canonical protein, we discovered that the alternate isoforms are in frame with TET1 and contain the catalytic domain, but lack the CXXC DNA binding domain (Figure 1C).

To confirm that $TET1^{ALT}$ is indeed expressed, we designed forward PCR primers to target either $TET1^{ALT}$ exon 1a or 1b (the non-coding exons of $TET1^{ALT}$) and placed the reverse primer in TET1 exon 3. In MDA-MB-231 cells, the primer set for exon1a amplified a PCR product of 104 bp in the cDNA samples, and 315 bp in the DNA control samples, indicating that the splicing machinery splices out 211 bp between $TET1^{ALT}$ exon1a and TET1 exon 3 (Supplementary Figure S2B). To confirm our findings, we gel ex-

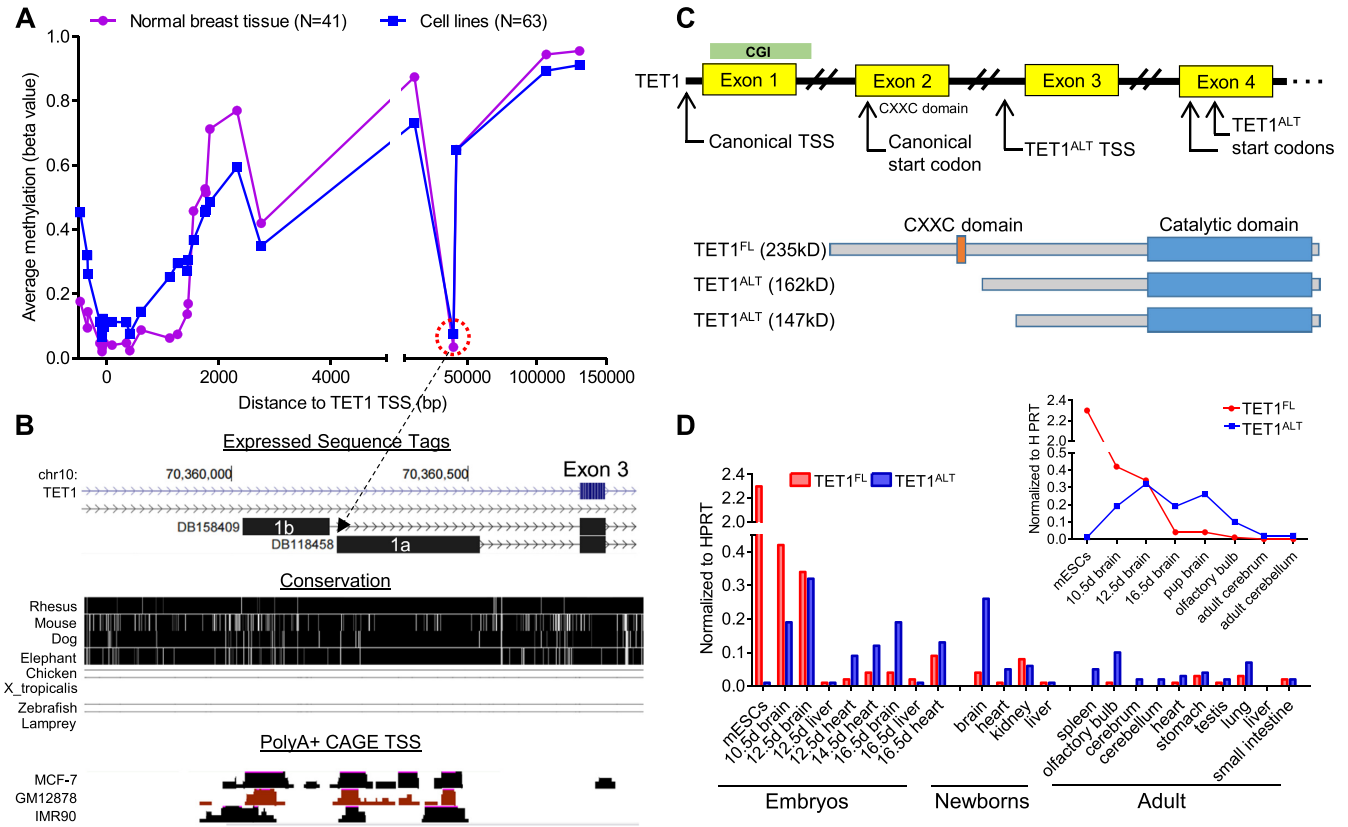

Figure 1. Identification of a novel TET1 isoform. (**A**) Average DNA methylation values from 450K array data of 41 normal breast tissues (TCGA) and 63 human cell lines (ENCODE) across the TET1 gene. An unmethylated CpG site is indicated by a red circle, located ~40 000 bp downstream from the TET1 TSS. (**B**) Mapping of the CpG site to intron 2 of TET1 using the UCSC genome browser. Characteristics of this region include the start sites of two expressed sequence tags (ESTs), high conservation and PolyA+ CAGE plus TSS peaks (ENCODE). (**C**) Schematic illustrating the gene models for TET1 and $TET1^{ALT}$. (**D**) qRT-PCR for TET1 and $TET1^{ALT}$ in mouse embryonic and adult tissues, normalized to HPRT (2^{-dCt}). Inset depicts expression of the isoforms in the developing brain ($N = 1$, two independent experiments performed).

tracted and sequenced the PCR products (boxed in red) and aligned them to the genome using BLAT (Supplementary Figure S2C). The spliced PCR product was also found in additional cell lines including MCF10A, HCC1599, BT549, HCC2218 and HMLE (data not shown). Furthermore, PCR of $TET1^{ALT}$ exon 1b in MCF10A cDNA also led to the amplification of the spliced product of 107 bp, while no bands were detected at the unspliced size of 637 bp (Supplementary Figure S2D).

To measure relative abundance of the TET1 isoforms, we used isoform specific qRT-PCR. We found that $TET1^{FL}$ is highly expressed in mESCs, whereas $TET1^{ALT}$ is repressed (Figure 1D). During embryonic development, we observed an isoform switch, where the $TET1^{FL}$ isoform slowly decreases in expression, as $TET1^{ALT}$ appears and progressively increases. This is most evident in brain development (see Figure 1D inset), where $TET1^{ALT}$ becomes the dominant isoform expressed during development and in the adult olfactory bulb.

$TET1^{ALT}$ promoter activity

To identify the chromatin architecture surrounding the $TET1^{ALT}$ promoter, we analyzed publicly available ChIP-seq datasets for $H3K4me^3$, H3K27Ac and $H3K27me^3$ (36). $H3K4me^3$ is a histone mark permissive for transcription and generally marks active or poised promoters, while $H2K27me^3$ is generally a repressive mark. The 17/19 cell lines were marked by $H3K4me^3$ and 0/19 had $H3K27me^3$, indicating that $TET1^{ALT}$ is active or permissive for transcription in these cell lines. In Figure 2A, histone marks from three representative cell lines H1-hESCs, GM12878 and HeLa cells are displayed for both the canonical and $TET1^{ALT}$ promoter. GM12878, a lymphoblastoid cell line, is enriched for $H3K4me^3$ at the $TET1^{ALT}$ promoter but not at the $TET1^{FL}$ promoter. In contrast, $H3K4me^3$ is enriched at the $TET1^{FL}$ promoter but not at the $TET1^{ALT}$ promoter. HeLa cells are used as a negative control, where no enrichment for either promoter is found. Additional cell lines are shown in Supplementary Figure S3. We also analyzed publicly available ChIP-seq datasets for transcription factors and found that the $TET1^{ALT}$ promoter is bound by a multitude of factors that are likely regulating its activity, including the transcription factors MYC (MCF-7, NB4), YY1 (GM12878, GM12892, K562) and NF-kappa-B (GM12891) (Supplementary Figure S4).

We next analyzed ChIP-seq datasets for mouse tissues, including both embryonic and adult (Figure 2B) (36). In agreeance with the human data, mouse ESCs have active

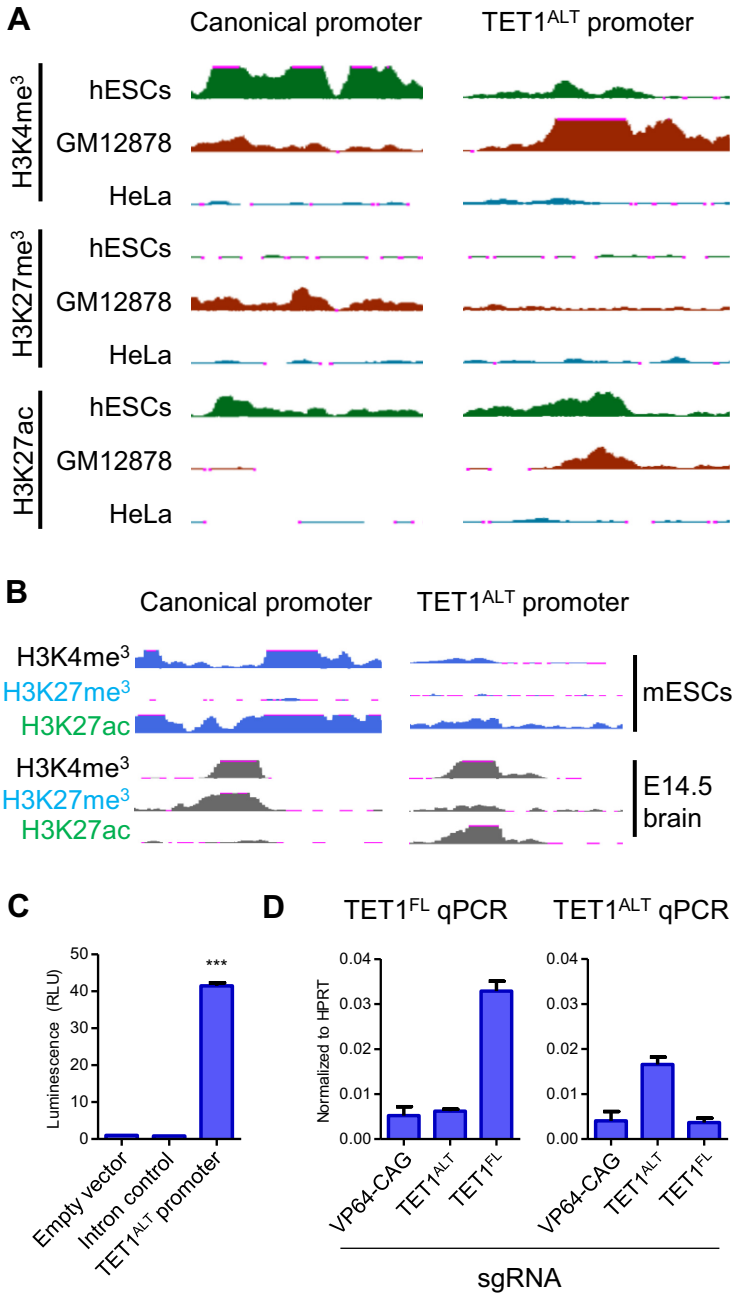

Figure 2. TET1^{ALT} promoter activity. (**A**) ChIP-seq histone marks (ENCODE) for H3K4me^3, H3K27me^3 and H3K27Ac in H1-hESCs, GM12878 and HeLa cells for the TET1 canonical and TET1^{ALT} promoters. (**B**) ChIP-seq histone marks H3K4me^3, H3K27me^3 and H3K27Ac in mESCs and embryonic day 14.5 brain (ENCODE). (**C**) DNA fragments from a control intron region and the TET1^{ALT} promoter region were each cloned into pGL4 luciferase reporter construct and transfected into HEK293T cells, along with empty vector control. Luminescence (RLU) is normalized to Renilla and empty vector control (technical and biological triplicates). (**D**) MCF10A cells were infected with lentiviruses expressing dCas9–VP64 with scrambled gRNAs (VP64-CAG) or gRNAs targeting the TET1^{ALT} promoter or the TET1^{FL} promoter. TET1^{FL} and TET1^{ALT} transcript expression was assayed by qRT-PCR (data shown as the average of biological duplicates, technical triplicates).

marks for TET1^{FL} but not for TET1^{ALT}, indicating that TET1^{ALT} is likely inactive in both mouse and human ESCs. However, during embryonic development, we see that active promoter marks are gained at the TET1^{ALT} promoter in several tissues with an isoform switch in many tissues. For example, in embryonic day 14.5 brain, the canonical promoter becomes poised (marked by H3K27me^3 and H3K4me^3) and the TET1^{ALT} promoter becomes active (H3K4me^3 and H3K27Ac). Histone marks in additional tissues can be found in Supplementary Figure S5. Importantly, the histone marks at the TET1^{ALT} promoter agree with our qRT-PCR RNA expression data. Active histone marks are found at the TET1^{ALT} promoter in tissues where the alternate isoform is expressed (embryonic brain), but are absent in tissues where the isoform is not expressed (mESCs).

To confirm the promoter activity of TET1^{ALT} *in vitro*, we cloned the promoter into a luciferase reporter assay and transfected it into HEK293T cells. A 996 bp region corresponding to the alternate promoter sequence drove 40-fold higher levels of luciferase compared to empty vector and intron control (Figure 2C). To confirm that the TET1^{ALT} promoter is driving expression of the TET1^{ALT} transcripts, we used CRISPR dCas9 fused to the VP64 activator domain to activate transcription of each isoform independently. We designed gRNAs to target either the canonical or the alternate TET1 promoter and transfected them into an immortalized mammary cell line (MCF10A) that overexpressed the dCas9–VP64 fusion protein. By qRT-PCR, we found that only gRNAs tethered to the alternate promoter lead to an increase in TET1^{ALT}, while gRNAs tethered to the canonical promoter did not affect TET1^{ALT} transcription (Figure 2D). Taken together, our data show that a conserved intragenic alternate promoter is used to activate transcription of an alternate isoform of TET1 that potentially retains the catalytic domain but lacks the CXXC DNA binding domain.

TET1^{ALT} produces a truncated protein

To test if TET1^{ALT} produces a detectable protein, we performed a western blot analysis using protein from a variety of cell lines and found that in addition to the TET1^{FL} band at 235 kDa, there is a strong band at ∼162 kDa and a smaller more variable band at 150 kDa (Figure 3A). The smaller band (marked by an asterisk) may be non-specific as it persists after CRISPR knockout (see below). This was observed with multiple antibodies and is consistent with western blots from published studies of the TET1 protein (37). We evaluated expression levels in human ESCs since they express TET1^{FL}, but not TET1^{ALT}. As expected, a band is observed for TET1^{FL} in the hESCs. Additionally, the smaller band at ∼150 kDa is also observed, which is further evidence that this band is indeed non-specific. A closer look at the TET1^{ALT} band at ∼162 kDa reveals overexpression in several breast cancer cell lines (HCC2218, HCC1599, MCF7, MDA-MB-231) compared to the untransformed, immortalized breast lines (HMLE, MCF10A). To verify that the bands are indeed TET1^{ALT}, we used a pIRES expression construct to overexpress either empty vector, TET1^{ALT} or TET1^{FL} in HEK293T cells. Western blot analysis of these lysates using a TET1 antibody (Figure 3B) and a FLAG tagged antibody (Figure 3C) revealed that TET1^{ALT} overexpression resulted in one major band around 162 kDa, which is not observed upon overexpression of the TET1^{FL} isoform. Protein produced by the TET1^{FL} isoform is found at ∼235 kDa, as expected. We quantified the non-saturated bands (normalized to β-actin) and found TET1^{ALT} is overexpressed 9.4-fold and TET1^{FL} is overexpressed 5.4-fold compared to empty vector control. To complement the overexpression experiments and to verify that the band is specific to TET1, we designed CRISPR gRNAs to target a common exon shared between TET1 and TET1^{ALT} (targeting the non-coding TET1^{ALT} exons 1a or 1b alone was not successful). We generated knockouts in MDA-MB-231 cells, as this cell line expresses high levels of TET1^{ALT} and low levels of TET1^{FL}. Upon knockout, we see a loss of the TET1^{ALT} band at 162 kDa, and see minimal to no change in the lower band, further suggesting the lower band is likely non-specific (Figure 3D). Cloning and sequencing of the knockout cells confirmed that CRISPR induced frameshift mutations resulting in an early stop codon for both TET1 alleles (Supplementary Figure S6). These data confirm that TET1^{ALT} expression results in a detectable protein that is truncated compared to TET1^{FL}.

TET1^{ALT} is functional and distinct from TET1^{FL}

To test TET1^{ALT} catalytic activity, we blotted DNA from the overexpressing HEK293T cells onto a membrane and probed with 5hmC (Figure 3E). As predicted, both TET1^{ALT} and TET1^{FL} display increased 5hmC, indicating TET1^{ALT} can hydroxymethylate DNA. To estimate the extent to which 5hmC is dependent on TET1^{FL} or TET1^{ALT}, we performed a 5hmC slot blot in our TET1 knockout MDA-MB-231 cells. Upon loss of TET1^{ALT}, we see a 30% reduction compared to empty vector control (Supplementary Figure S7). This is consistent with previous reports suggesting that TET2 and/or TET3 can compensate for 5hmC production in the absence of TET1 (13). TET1 has previously been reported to affect gene transcription, as loss of TET1 leads to both the upregulation and downregulation of genes (18,19). To determine if TET1^{ALT} also affects gene expression, we overexpressed empty vector, TET1^{ALT} and TET1^{FL} in HEK293T cells and performed RNA-seq. An unsupervised hierarchal cluster analysis and a PCA of the data showed that TET1^{FL} clusters separated from empty vector and TET1^{ALT} (Figure 4A, B). Although both isoforms have only modest effects on transcription (changes in <1.5% of the transcriptome), TET1^{FL} overexpression induced substantially more gene expression changes than TET1^{ALT}: TET1^{FL} led to the upregulation of 7.5-fold more genes than TET1^{ALT} and repressed ∼3-fold more genes. There was only a moderate overlap between the gene expression targets of TET1^{FL} and TET1^{ALT} with a large number of unique targets (Figure 4C).

In our previous work, we found that the TET1 catalytic domain (TET1^{CD}, which lacks the CXXC domain) induced widespread hypomethylation, but TET1^{FL} produced small changes in methylation at sites with low basal methylation (19). Since TET1^{ALT} lacks the CXXC domain, we wondered if it would induce widespread or targeted DNA demethylation. We first examined LINE1 methylation using bisulfite pyrosequencing. As expected, the expression of TET1^{CD} led to demethylation of the LINE1 repetitive elements. However, neither TET1^{FL} nor TET1^{ALT} had major effects on LINE1 methylation (change of methylation <1%), indicating preserved target specificity (Figure 4D). We next analyzed genome-wide DNA methylation data using the DREAM method (23). Hierarchal cluster analysis of HEK293T cells overexpressing the TET1 isoforms showed that TET1^{ALT} clusters more closely with empty vector than TET1^{FL} (Supplementary Figure S8A), indicating weaker effects on DNA methylation. To visualize the sites that changed methylation, we used volcano plots and found that both TET1^{FL} and TET1^{ALT} expression led to changes in DNA methylation compared to empty vector

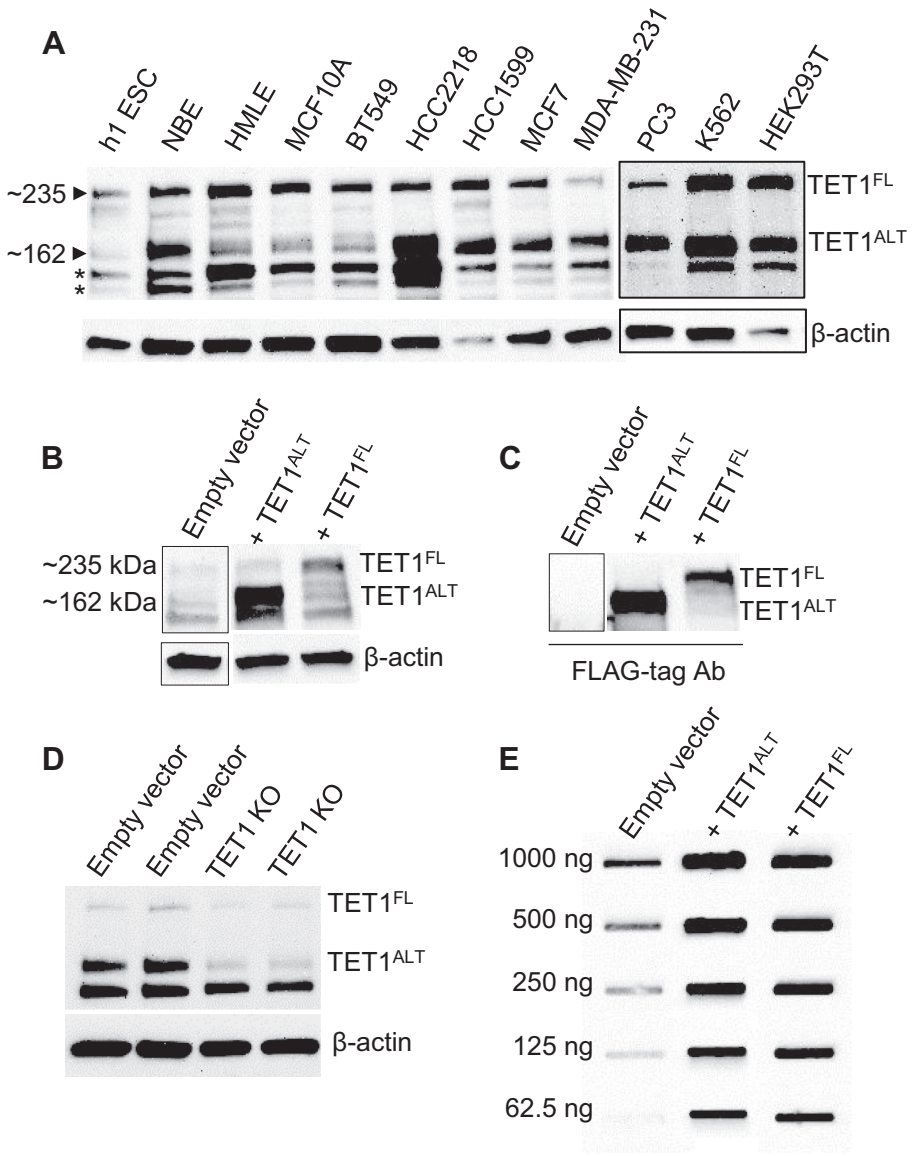

Figure 3. TET1^{ALT} produces a truncated protein. (**A**) Western blot of the TET1 isoforms in human cell lines (two independent experiments performed, asterisk denotes non-specific bands). (**B**) Western blot analysis of empty vector, TET1^{ALT} or TET1^{FL} overexpression in HEK293T cells with anti-TET1 antibody (two independent experiments performed). (**C**) Western blot analysis of empty vector, TET1^{ALT} or TET1^{FL} overexpression in HEK293T cells with anti-FLAG antibody. (**D**) Western blot of CRISPR TET1 knockout in MDA-MB-231 cells using an anti-TET1 antibody (technical duplicates, three independent experiments performed). (**E**) 5-hydroxymethylcytosine (5hmC) DNA slot blot of HEK293T cells expressing either empty vector, TET1^{ALT} or TET1^{FL}.

control (Figure 4E). TET1^{FL} demethylated 3-fold more target genes than TET1^{ALT} (770 CpG sites), but TET1^{ALT} still decreased the methylation of 225 CpG sites by at least 5%. Furthermore, TET1^{FL} and TET1^{ALT} mostly have their own individual target CpG sites (Figure 4F). A small number of genes gained methylation with overexpression of either isoform, which may represent background/false positives. In this short-term experiment, both TET1^{ALT} and TET1^{FL} affected predominantly non-CGIs (Supplementary Figure S8B). Taken together, our data show that TET1^{ALT} is functional but relatively weak when overexpressed alone, likely because it needs to be recruited to DNA by specific co-factors, as previously shown (37,38). Our study does not address whether TET1^{ALT} is physiologically different from TET1^{FL}, but instead illustrates that the two proteins have different gene expression and methylation targets and are thus functionally distinct.

TET1^{ALT} is overexpressed in cancer

The role of TET1 in cancer remains under debate and previous reports of 'loss' of TET1 may relate to the downregulation of TET1^{FL} post-development. We, therefore, sought to examine isoform specific TET1 expression in cancer. We first examined publicly available TSS-seq datasets generated by a method that combines oligo capping with Illumina GA technology to map the exact positions of transcriptional start sites (39). We found very few reads map to

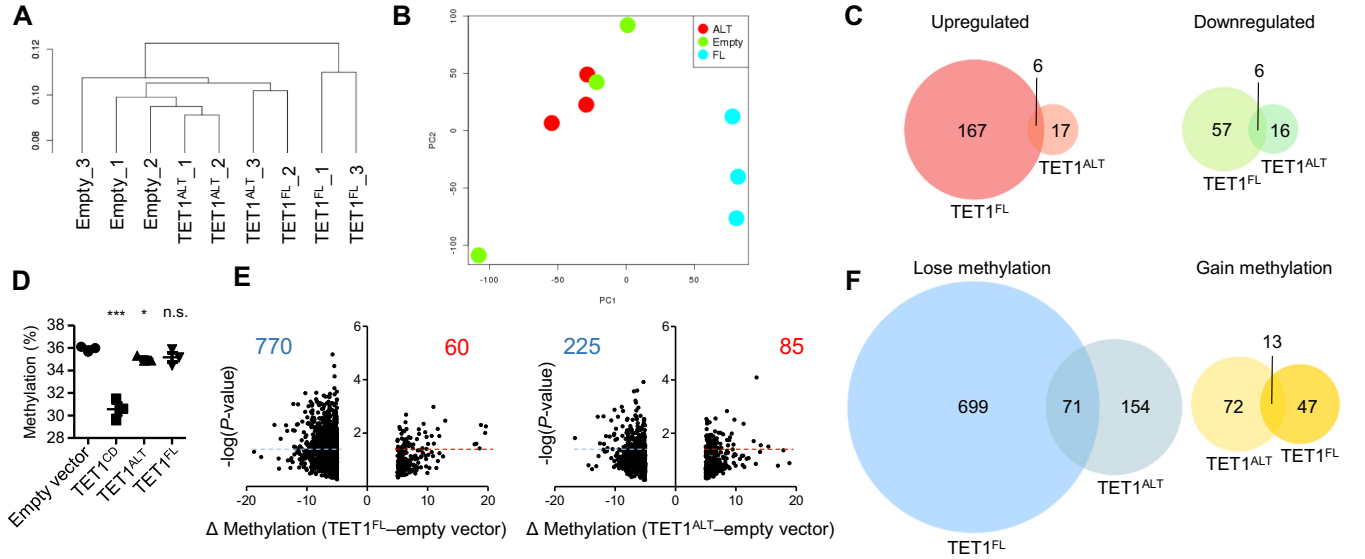

Figure 4. TET1^{ALT} is functional and distinct from TET1^{FL}. (**A**) Dendrogram of RNA-seq hierarchical clustering of HEK293T cells overexpressing empty vector, TET1^{ALT} or TET1^{FL} ($N = 3$, biological triplicates). (**B**) Principal component analysis of RNA-seq data. (**C**) Overlap of genes upregulated and downregulated in TET1^{ALT} versus TET1^{FL} HEK293T cells, FC > 1.5 or FC < 0.66, $P < 0.05$. (**D**) LINE1 pyrosequencing assay of methylation in HEK293T cells overexpressing empty vector, TET1^{CD}, TET1^{ALT} or TET1^{FL} (technical and biological triplicates). (**E**) Volcano plot analysis from genome-wide DNA methylation data (digital restriction enzyme analysis of methylation, DREAM) comparing empty vector to TET1^{ALT} or TET1^{FL}. CpG sites were filtered for minimum sequencing depth of 50 reads and methylation changes >5%. Each point is the average methylation value for biological triplicates. (**F**) Overlap of genes that lose and gain methylation in HEK293T cells that overexpress either TET1^{ALT} or TET1^{FL} compared to empty vector control.

the TET1^{FL} or TET1^{ALT} TSS in normal tissues, which corroborates RNA-seq data showing that TET1 is expressed at low levels in normal adult tissues (GTEx Portal) (Supplementary Figures S9 and 10) (39,40). However, compared to normal breast, lung and lymphocytes, breast and lung cancer as well as Burkitt's lymphoma have substantially more TET1^{ALT} TSS reads (Figure 5A). Interestingly, there is little to no increased activity at the canonical TET1^{FL} promoter in these cell lines (Supplementary Figure S11) indicating that TET1^{ALT} is specifically activated in cancer according to TSS data. For example, 17/22 lung cancer cell lines have multiple reads mapping to the TET1^{ALT} alternate exons (Supplementary Figure S12). We used isoform specific qRT-PCR to confirm these data in breast cells. Figure 5B shows TET1^{ALT} expression in two untransformed, immortalized breast cell lines (MCF10A and HMLE) and four breast cancer cell lines (MCF7, HCC2218, HCC1599 and BT549). TET1^{ALT} was overexpressed in several breast cancer cell lines, with the highest expression being in HCC1599.

To determine if TET1^{ALT} is overexpressed in primary human samples, we used RNA-sequencing files from TCGA to quantify TET1^{ALT} by counting the reads aligning to the TET1^{ALT} exons (exon 1a and exon 1b). TET1^{ALT} is expressed at low levels in normal breast ($N = 107$) but is substantially overexpressed in breast cancer patient samples ($N = 807$), Mann–Whitney, $P = 0.03$ (Figure 5C). Compared to the average TET1^{ALT} expression in normal breast, most breast cancers overexpress TET1^{ALT}; however, a subset of cases shows dramatic overexpression. Next, we analyzed additional cancer types, including uterine, glioblastoma, AML and ovarian cancer. Compared to their normal tissue counterparts, uterine cancer and glioblastoma significantly overexpress the TET1^{ALT} isoform (Figure 5D). Normal tissue was not available for AML and ovarian cancer, so we simply compared TET1^{ALT} expression levels to other cancer types in Supplementary Figure S13A. Reads aligning to TET1^{ALT} exon 1a showed splicing to exon 3 in breast cancer patient samples (Figure 5E), similar to our *in vitro* splicing analyses. Next, we investigated differential TET1 exon usage in cancer. Since TET1^{FL} and TET1^{ALT} both use exons 3–12, and TET1^{ALT} does not use exons 1–2, a discrepancy in usage of the first two exons indicates differential isoform expression. In breast cancer, glioblastoma, uterine cancer and AML exons toward the 3′ end of the gene are used much more frequently than exons 1–2, evidence of an alternate truncated isoform (Figure 5F and Supplementary Figure S13B). Finally, since these datasets indicated that TET1^{ALT} is the predominant isoform overexpressed in cancer, we next looked at TET1 expression across multiple cancers using TCGA data. We find TET1 has a wide range of expression in several cancer types including AML, ovarian, breast and lung, but has very low and tight expression in colorectal, renal, pancreatic and prostate cancer (Supplementary Figure S13C). Interestingly, TET1 appears to be frequently amplified in some cancers, another indicator that TET1^{ALT} may play an oncogenic role in cancer (Supplementary Figure S13D) (30,41–45). Finally, we determined whether TET1 expression associates with overall survival outcomes in patient samples using data downloaded from TCGA and METABRIC (31–33). Patients were considered TET1 high if expression levels were >1 SD above the mean for all cancers except breast (breast SD > 1, $P = 0.05$). For breast cancer, a more stringent cut off of SD > 2 was used as there were many patients that fit this criteria ($N = 83$), whereas in glioblastoma only one patient has SD > 2. Interestingly, TET1 expression is associated with a worse overall

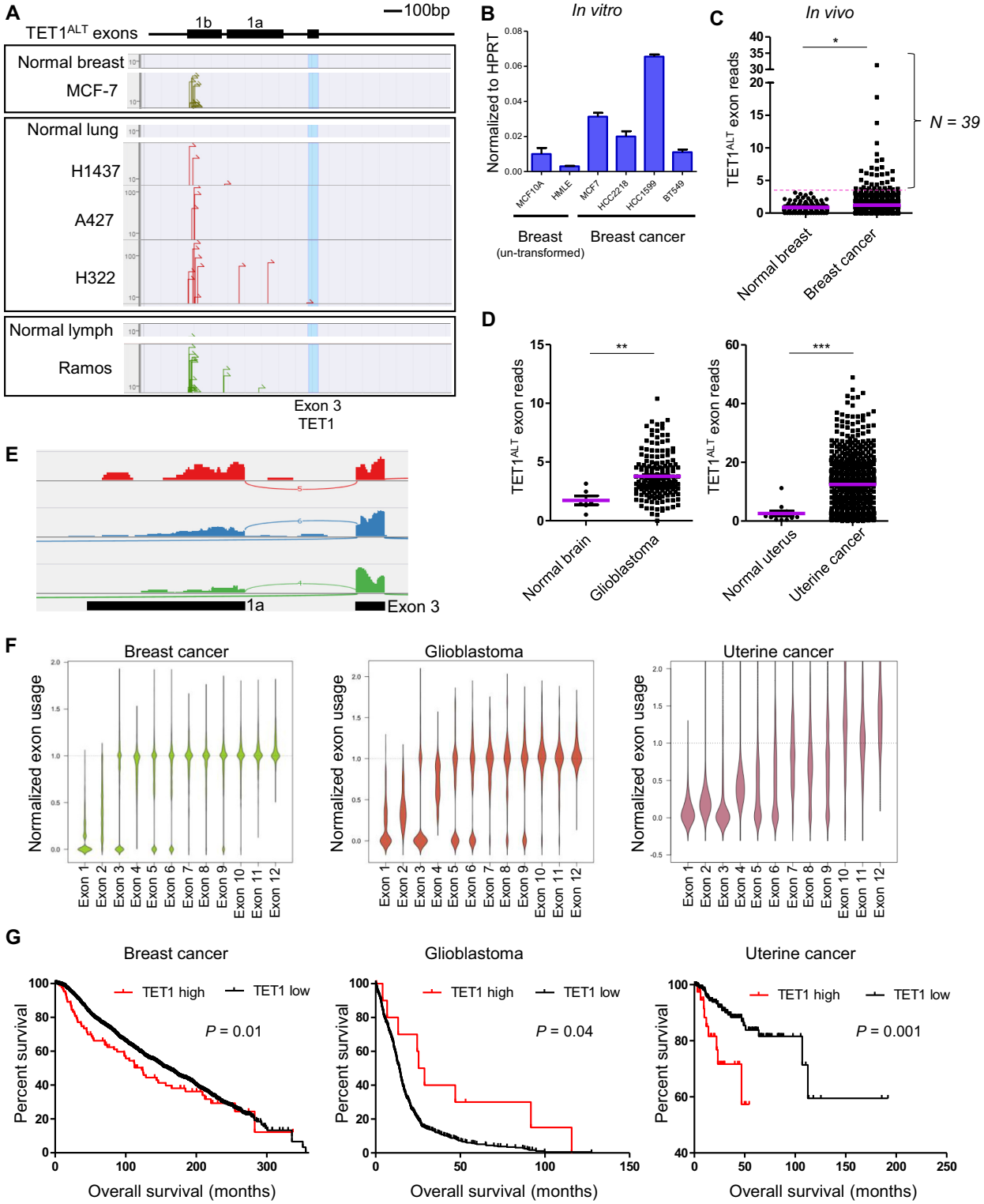

Figure 5. TET1^{ALT} is overexpressed in cancer. (**A**) TSS-seq datasets for normal breast versus a breast cancer cell line, normal lung versus three lung cancer cell lines and normal lymph versus lymphoma cell line showing TSS peaks at the TET1^{ALT} promoter (39). (**B**) *In vitro* analysis of TET1^{ALT} RNA expression across a variety of human cell lines by qRT-PCR (technical and biological triplicates). (**C**) *In vivo* analysis of TCGA data of TET1^{ALT} RNA-sequencing reads from normal breast tissue ($N = 107$) and tissue from breast cancer patients ($N = 807$). Y-axis is TET1^{ALT} reads (TET1^{ALT} exon 1a + exon 1b)/chromosome 10 reads $\times$ 1e^6. Mann–Whitney test, $P = 0.03$. (**D**) *In vivo* analysis of TCGA data of TET1^{ALT} RNA-sequencing reads from normal brain ($N = 6$) and glioblastoma tumor samples ($N = 167$, $P = 0.004$) (left) and normal uterine ($N = 12$) and uterine tumor samples ($N = 574$, $P < 0.0001$) (right). (**E**) Sashimi plot of RNA-sequencing reads from three representative breast cancer patient samples aligned to the TET1^{ALT} exons. (**F**) Normalized exon usage for TET1 in breast cancer, glioblastoma and uterine cancer using TCGA datasets. (**G**) Survival curves based on TET1 expression in breast cancer (TET1 high is SD > 2 above the mean, SD > 1, $P = 0.05$), glioblastoma and uterine cancer (SD > 1 above the mean).

survival in cancers that are predominantly found in women (uterine, $P = 0.001$; breast, $P = 0.01$; ovarian, $P = 0.007$), is associated with a better survival in glioblastoma ($P = 0.04$) and is not associated with survival in AML (Figure 5G and Supplementary Figure S14). Taken together, these data show that the TET1^{ALT} TSS is aberrantly activated in multiple cancers, suggesting that TET1^{ALT} may be a cancer-specific alternate isoform involved in tumorigenesis.

DISCUSSION

In this paper, we identified a novel isoform of TET1 (TET1^{ALT}) that lacks the CXXC DNA binding domain, but retains its catalytic activity. The canonical TET1 protein (TET1^{FL}) is the only isoform expressed in ESCs while TET1^{ALT} is expressed in most adult and cancer cells, suggesting a totally different function for TET1 in ESCs versus adult cells. Previous reports found TET1 to bind to CGIs and protect them from gains of methylation (19). This is an important mechanism in ESCs to protect CGIs during waves of *de novo* methylation (via TET1^{FL}). However, TET1^{FL} and TET1^{ALT} are dramatically different proteins, as the 671 amino acids truncation of TET1^{ALT} causes the loss of important regulatory domains including its DNA binding domain. Our findings suggest that there is no role for TET1^{ALT} in ESCs, and that instead TET1^{ALT} serves as a dynamic regulator in adult cells where it is likely recruited to DNA by specific co-factors as previously shown for TET1 (37,38). This would allow for the precise control of methylation in a tissue-specific manner. It is important to note that much of the literature on TET1 has focused on ESCs, which may not be as relevant to how TET1 functions in adult cells.

The TET1^{ALT} promoter is highly enriched for the active promoter mark H3K4me^3 and is bound by a multitude of transcription factors in multiple cell types. We show that tethering an activator domain to the TET1^{ALT} promoter increases transcription specifically for the TET1^{ALT} isoform. Overexpression of TET1^{ALT} yields a truncated protein at ~162 kDa and results in production of 5hmC, suggesting the protein is catalytically active. Other published work found the TET1^{ALT} promoter to be an enhancer in human ESCs (46). The investigators show the TET1^{ALT} promoter region to be marked by H3K4me1 and H3K27Ac and bound by the transcription factors OCT3/4, MYC and NANOG. Since TET1^{ALT} is not expressed in ESCs, it is possible that the TET1^{ALT} promoter serves as an enhancer in ESCs but switches to a promoter during differentiation. This phenomena has been observed in the literature where intragenic enhancers act as alternative tissue-specific promoters, allowing for transcription to occur in a developmental and cell type-specific manner (47).

TET1^{FL} has been shown by us and others to both activate and repress gene transcription both dependent and independent of its demethylase activity (18,19). We find that TET1^{ALT} has fewer effects on gene expression than TET1^{FL}. One possible explanation is that the co-regulatory proteins targeting TET1^{ALT} to DNA are not expressed in HEK293T cells or that the co-regulatory proteins must be co-expressed with TET1^{ALT} to see robust changes in gene expression. Another possibility is that TET1^{ALT} has few target genes in HEK293T cells. In the future, it will be important to identify the co-regulatory proteins recruiting TET1^{ALT} to DNA. There are several examples of TET1 being targeted to DNA, such as via HIF1A where it affects methylation at hypoxic response genes (38). In addition, TET1 can be targeted to DNA via FOXA1, where these proteins work together to modify the epigenetic signature at linage-specific enhancers (37). A closer look at the published data suggest that TET1^{ALT}, not TET1^{FL}, is the interacting factor because when the investigators pulled down endogenous FOXA1 and probed with a TET1 antibody, the band was visible at ~150 kDa (the approximate size of TET1^{ALT}) (37).

We and others have shown that TET1^{FL} is unable to induce global demethylation because its CXXC domain limits its ability to bind outside of CGIs (19). Our data suggest that TET1^{ALT}, even though it lacks the CXXC domain, is still unable to induce wide-spread hypomethylation. We believe this is likely due to TET1^{ALT} being restricted to specific DNA sites by co-regulatory proteins. We observed that TET1^{ALT} demethylates ~6-fold more CpG sites at non-CGIs, compared to CGIs. We assume this preference for non-CGIs is due to its lack of the CXXC domain. This also provides a rationale as to why there is minimal overlap between TET1^{ALT} and TET1^{FL} gene expression and methylation targets. TET1^{FL} is targeted to CGIs by its CXXC domain and could be targeted to non-CGIs by proteins that interact with its CXXC domain. However, since TET1^{ALT} lacks this domain, it is likely regulated by another set of proteins that bind elsewhere in the TET1^{ALT} protein.

It is important to note that our data do not comment on the physiological differences between TET1^{FL} and TET1^{ALT}, but simply demonstrate that the isoforms have different methylation and gene expression targets and are thus distinct. While this manuscript was under review, another group reported that a truncated isoform of TET1 fails to erase imprints in primordial germ cells, suggesting physiological differences between the two isoforms (48). In their study and our study, overexpression experiments were used to investigate the isoforms function. This is because TET1^{FL} and TET1^{ALT} share the same coding exons and it remains a challenge to cleanly knockout each isoform independently. For now, we believe overexpression is the cleanest way to study the proteins; however, isoform specific knockouts should be the focus of future experimentation.

Although we are in the early stages of understanding the molecular functions of TET1^{ALT}, our results indicate that TET1^{ALT} is activated in cancer. This includes activation at the DNA level (gains of TSS peaks at the TET1^{ALT} promoter), and at the RNA and protein level. TET1^{ALT} appears to be the dominant isoform overexpressed in cancer, as few activation marks are found at the TET1^{FL} promoter and TET1^{FL} protein expression is unchanged between the normal and breast cancer cell lines. Since TET1 is frequently amplified in cancer and is associated with a worse overall survival in breast, uterine and ovarian cancers, we believe that TET1, specifically TET1^{ALT}, could play a more oncogenic role in cancers found in women. Several studies have implicated TET1 as a tumor suppressor (14,49,50); however, TET1^{FL} may have a different function than TET1^{ALT} in cancer. A similar phenomenon has been described for a truncated isoform of ALK (ALK^{ATI}), which is specifically

activated in cancer and is tumorigenic in mouse models (51). Another possibility is that the previous reports of 'loss' of TET1 may relate to the downregulation of $TET1^{FL}$ post-development. Our work emphasizes the importance of distinguishing between $TET1^{FL}$ and $TET1^{ALT}$ both functionally and with respect to their varying roles in cancer.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr Kenneth Zaret at the University of Pennsylvania for providing hESC lysates. We would also like to thank Dr Sendurai A. Mani at The University of Texas, MD Anderson Cancer Center for generously providing HMLE cells and Dr. Xiaowei Chen at Fox Chase Cancer Center for providing the NBE cells.
Author contributions: C.R.G. and J-P.J.I. conceived the project and designed the experiments. C.R.G. and J-P.J.I. generated the figures and wrote the manuscript. C.R.G. performed all experiments, except for DREAM which was performed by B.P. C.R.G., J.M. and J.J. performed bioinformatic analyses. N.E. provided RNA from mESCs and mouse embryonic tissues. S.M. provided adult mouse tissue. All authors discussed the results, commented on the manuscript and declared no conflicts of interest.

FUNDING

National Institutes of Health [CA158112 and CA100632 to J-P.J.I. and CA210424 to C.R.G.]; Ellison Medical Foundation and F. M. Kirby Foundation (to J-P.J.I.). Funding for open access charge: Fels Institute; Temple University.
Conflict of interest statement. None declared.

REFERENCES

1. Robertson,K. (2005) DNA methylation and human disease. *Nat. Rev. Genet.*, **6**, 597–610.
2. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
3. Ehrlich,M. (2010) DNA hypomethylation in cancer cells. *Epigenomics*, **1**, 239–259.
4. Hansen,K.D., Timp,W., Bravo,H.C., Sabunciyan,S., Langmead,B., McDonald,O.G., Wen,B., Wu,H., Liu,Y., Diep,D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.*, **43**, 768–775.
5. Fang,F., Turcan,S., Rimner,A., Kaufman,A., Giri,D., Morris,L.G.T., Shen,R., Seshan,V., Mo,Q., Heguy,A. *et al.* (2011) Breast cancer methylomes establish an epigenomic foundation for metastasis. *Sci. Transl. Med.*, **3**, 75ra25.
6. Baylin,S.B. and Jones,P.A. (2011) A decade of exploring the cancer epigenome – biological and translational implications. *Nat. Rev. Cancer*, **11**, 726–734.
7. Hermann,A., Goyal,R. and Jeltsch,A. (2004) The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. *J. Biol. Chem.*, **279**, 48350–48359.
8. Okano,M., Bell,D.W., Haber,D.A. and Li,E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.
9. Ito,S., Shen,L., Dai,Q., Wu,S.C., Collins,L.B., Swenberg,J.A., He,C. and Zhang,Y. (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* , **333**, 1300–1303.
10. Guo,J.U., Su,Y., Zhong,C., Ming,G.L. and Song,H. (2011) Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell*, **145**, 423–434.
11. Wu,H. and Zhang,Y. (2011) Mechanisms and functions of Tet protein- mediated 5-methylcytosine oxidation. *Genes Dev.*, **25**, 2436–2452.
12. Tahiliani,M., Koh,K.P., Shen,Y., Pastor,W.A., Bandukwala,H., Brudno,Y., Agarwal,S., Iyer,L.M., Liu,D.R., Aravind,L. *et al.* (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930–935.
13. Dawlaty,M.M., Ganz,K., Powell,B.E., Hu,Y.C., Markoulaki,S., Cheng,A.W., Gao,Q., Kim,J., Choi,S.W., Page,D.C. *et al.* (2011) Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. *Cell Stem Cell*, **9**, 166–175.
14. Hsu,C.-H., Peng,K.-L., Kang,M.-L., Chen,Y.-R., Yang,Y.-C., Tsai,C.-H., Chu,C.-S., Jeng,Y.-M., Chen,Y.-T., Lin,F.-M. *et al.* (2012) TET1 suppresses cancer invasion by activating the tissue inhibitors of metalloproteinases. *Cell Rep.*, **2**, 568–579.
15. Li,L., Li,C., Mao,H., Du,Z., Chan,W.Y., Murray,P., Luo,B., Chan,A.T., Mok,T.S., Chan,F.K. *et al.* (2016) Epigenetic inactivation of the CpG demethylase TET1 as a DNA methylation feedback loop in human cancers. *Sci. Rep.*, **6**, 26591–26603.
16. Ito,S., D'Alessio,A.C., Taranova,O.V., Hong,K., Sowers,L.C. and Zhang,Y. (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, **466**, 1129–1133.
17. Williams,K., Christensen,J., Pedersen,M.T., Johansen,J. V, Cloos,P.a.C., Rappsilber,J. and Helin,K. (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*, **473**, 343–348.
18. Wu,H., D'Alessio,A.C., Ito,S., Xia,K., Wang,Z., Cui,K., Zhao,K., Sun,Y.E. and Zhang,Y. (2011) Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature*, **473**, 389–393.
19. Jin,C., Lu,Y., Jelinek,J., Liang,S., Estecio,M.R.H., Barton,M.C. and Issa,J.-P. (2014) TET1 is a maintenance DNA demethylase that prevents methylation spreading in differentiated cells. *Nucleic Acids Res.*, **42**, 6956–6971.
20. Iyer,L.M., Tahiliani,M., Rao,A. and Aravind,L. (2009) Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell Cycle*, **8**, 1698–1710.
21. Teif,V.B., Beshnova,D.a, Vainshtein,Y., Marth,C., Mallm,J.-P., Höfer,T. and Rippe,K. (2014) Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Res.*, **24**, 1285–1295.
22. Elenbaas,B., Spirio,L., Koerner,F., Fleming,M.D., Zimonjic,D.B., Donaher,J.L., Popescu,N.C., Hahn,W.C. and Weinberg,R.A. (2001) Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. *Genes Dev.*, **15**, 50–65.
23. Jelinek,J., Liang,S., Lu,Y., He,R., Ramagli,L.S., Shpall,E.J., Estecio,M.R.H. and Issa,J.P.J. (2012) Conserved DNA methylation patterns in healthy blood cells and extensive changes in leukemia measured by a new quantitative technique. *Epigenetics*, **7**, 1368–1378.
24. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
25. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
26. Sanjana,N.E., Shalem,O. and Zhang,F. (2014) Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods*, **11**, 783–784.
27. Kearns,N.A., Genga,R.M.J., Enuameh,M.S., Garber,M., Wolfe,S.A. and Maehr,R. (2014) Cas9 effector-mediated regulation of transcription and differentiation in human pluripotent stem cells. *Development*, **141**, 219–223.
28. Yang,A.S., Estecio,M.R., Doshi,K., Kondo,Y., Tajara,E.H. and Issa,J.P. (2004) A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements. *Nucleic Acids Res.*, **32**, e38.

29. Dunham,I.T.E.P.C. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
30. Pereira,B., Chin,S.-F., Rueda,O.M., Vollan,H.-K.M., Provenzano,E., Bardwell,H.A., Pugh,M., Jones,L., Russell,R., Sammut,S.-J. *et al.* (2016) The somatic mutation profiles of 2, 433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.*, **7**, 11479–11494.
31. Curtis,C., Shah,S.P., Chin,S.-F., Turashvili,G., Rueda,O.M., Dunning,M.J., Speed,D., Lynch,A.G., Samarajiwa,S., Yuan,Y. *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.
32. Getz,G., Gabriel,S.B., Cibulskis,K., Lander,E., Sivachenko,A., Sougnez,C., Lawrence,M., Kandoth,C., Dooling,D., Fulton,R. *et al.* (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**, 67–73.
33. Voigt,P. and Reinberg,D. (2013) Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia The Cancer Genome Atlas Research Network. *N. Engl. J. Med.*, **368**, 2059–2074.
34. Djebali,S., Merkel,A., Lassmann,T., Tanzer,A., Lin,W., Schlesinger,F., Xue,C., Marinov,G.K., Khatun,J., Williams,B.A. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
35. Stothard,P. (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques*, **28**, 1102–1104.
36. Rosenbloom,K.R., Sloan,C.A., Malladi,V.S., Dreszer,T.R., Learned,K., Kirkup,V.M., Wong,M.C., Maddren,M., Fang,R., Heitner,S.G. *et al.* (2013) ENCODE Data in the UCSC Genome Browser: Year 5 update. *Nucleic Acids Res.*, **41**, 56–63.
37. Yang,Y.A., Zhao,J.C., Fong,K., Kim,J., Li,S., Song,C., Song,B., Zheng,B., He,C. and Yu,J. (2016) FOXA1 potentiates lineage-specific enhancer activation through modulating TET1 expression and function. *Nucleic Acids Res.*, **44**, 8153–8164.
38. Tsai,Y.-P., Chen,H.-F., Chen,S.-Y., Cheng,W.-C., Wang,H.-W., Shen,Z.-J., Song,C., Teng,S.-C., He,C. and Wu,K.-J. (2014) TET1 regulates hypoxia-induced epithelial-mesenchymal transition by acting as a co-activator. *Genome Biol.*, **15**, 513–526.
39. Suzuki,A., Wakaguri,H., Yamashita,R., Kawano,S., Tsuchihara,K., Sugano,S., Suzuki,Y. and Nakai,K. (2015) DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. *Nucleic Acids Res.*, **43**, D87–D91.
40. Ardlie,K.G., Deluca,D.S., Segrè,A. V, Sullivan,T.J., Young,T.R., Gelfand,E.T., Trowbridge,C.A., Maller,J.B., Tukiainen,T., Lek,M. *et al.* (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, **348**, 648–660.
41. Eirew,P., Steif,A., Khattra,J., Ha,G., Yap,D., Farahani,H., Gelmon,K., Chia,S., Mar,C., Wan,A. *et al.* (2015) Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, **518**, 422–426.
42. Baca,S.C., Prandi,D., Lawrence,M.S., Mosquera,J.M., Romanel,A., Drier,Y., Park,K., Kitabayashi,N., MacDonald,T.Y., Ghandi,M. *et al.* (2013) Punctuated evolution of prostate cancer genomes. *Cell*, **153**, 666–677.
43. Kumar,A., Coleman,I., Morrissey,C., Zhang,X., True,L.D., Gulati,R., Etzioni,R., Bolouri,H., Montgomery,B., White,T. *et al.* (2016) Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nat. Med.*, **22**, 369–378.
44. Grasso,C.S., Wu,Y.-M., Robinson,D.R., Cao,X., Dhanasekaran,S.M., Khan,A.P., Quist,M.J., Jing,X., Lonigro,R.J., Brenner,J.C. *et al.* (2012) The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, **487**, 239–243.
45. Beltran,H., Prandi,D., Mosquera,J.M., Benelli,M., Puca,L., Cyrta,J., Marotz,C., Giannopoulou,E., Chakravarthi,B.V.S.K., Varambally,S. *et al.* (2016) Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat. Med.*, **22**, 298–305.
46. Neri,F., Incarnato,D., Krepelova,A., Dettori,D., Rapelli,S., Maldotti,M., Parlato,C., Anselmi,F., Galvagni,F. and Oliviero,S. (2015) TET1 is controlled by pluripotency-associated factors in ESCs and downmodulated by PRC2 in differentiated cells and tissues. *Nucleic Acids Res.*, **43**, 6814–6826.
47. Kowalczyk,M.S., Hughes,J.R., Garrick,D., Lynch,M.D., Sharpe,J. a., Sloane-Stanley,J. a., McGowan,S.J., De Gobbi,M., Hosseini,M., Vernimmen,D. *et al.* (2012) Intragenic Enhancers Act as Alternative Promoters. *Mol. Cell*, **45**, 447–458.
48. Zhang,W., Xia,W., Wang,Q., Towers,A.J., Chen,J., Gao,R., Zhang,Y., Yen,C., Lee,A.Y., Li,Y. *et al.* (2016) Isoform switch of TET1 regulates DNA demethylation and mouse development. *Mol. Cell*, **64**, 1062–1073.
49. Neri,F., Dettori,D., Incarnato,D., Krepelova,A., Rapelli,S., Maldotti,M., Parlato,C., Paliogiannis,P. and Oliviero,S. (2015) TET1 is a tumour suppressor that inhibits colon cancer growth by derepressing inhibitors of the WNT pathway. *Oncogene*, **34**, 4168–4176.
50. Song,S.J., Poliseno,L., Song,M.S., Ala,U., Webster,K., Ng,C., Beringer,G., Brikbak,N.J., Yuan,X., Cantley,L.C. *et al.* (2013) MicroRNA-antagonism regulates breast cancer stemness and metastasis via TET-family-dependent chromatin remodeling. *Cell*, **154**, 311–324.
51. Wiesner,T., Lee,W., Obenauf,A.C., Ran,L., Murali,R., Zhang,Q.F., Wong,E.W.P., Hu,W., Scott,S.N., Shah,R.H. *et al.* (2015) Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. *Nature*, **526**, 453–457.