PNAS

Supporting Information for

Super-enhancer drives a tumor-specific splicing variant of MARCO to promote triple-negative breast cancer progression

Yun-Song Yang[1,2,3†], Xi Jin[1,3,†], Qin Li[3,4], Yi-Yu Chen[1,3], Fenfang Chen[1,3], Hena Zhang[3,4], Ying Su[3,4], Yi Xiao[1,3], Gen-Hong Di[1, 3], Yi-Zhou Jiang[1, 3,*], Shenglin Huang[3,4,*], Zhi-Ming Shao[1, 3, 4,*]

*** Correspondence to:**

Yi-Zhou Jiang, MD, Fudan University Shanghai Cancer Center, No. 270 Dong'an Road, Shanghai, 200032, P.R. China; Email: yizhoujiang@fudan.edu.cn

Shenglin Huang, PhD, Fudan University Shanghai Cancer Center, No. 270 Dong'an Road, Shanghai, 200032, P.R. China; Email: slhuang@fudan.edu.cn

Zhi-Ming Shao, MD, Fudan University Shanghai Cancer Center, No. 270 Dong'an Road, Shanghai, 200032, P.R. China; Email: zhimin_shao@yeah.net

This PDF file includes:

Supporting Materials and Methods

Study cohorts and data sources

In this study, we included patients from Fudan University Shanghai Cancer Center Triple Negative Breast Cancer (FUSCCTNBC) project (a total of 465 patients; 360 patients with RNA-seq data, 401 patients with copy number alteration (CNA) data, 279 patients with whole-exome sequencing (WES) data) (1) and The Cancer Genome Atlas (TCGA) cohort (9,581 patients across 33 cancer types with RNA-seq data and 338 breast cancer patients with 450k array methylation data) (2), and all patients from these cohorts provided appropriate informed consent for data and tissue use as previously described.

Raw RNA-seq data (in FASTQ format) of the FUSCCTNBC cohort (360 TNBC tumor tissues and 88 paired adjacent non-tumor tissues) were downloaded from the Sequence Read Archive (SRA: SRP157974). TCGA data (9,581 tumor tissues and 2,411 paired adjacent non-tumor tissues) were downloaded from the Genomic Data Commons (https://gdc.cancer.gov/). Adjacent non-tumor tissues were taken from sites greater than 2 cm from the tumor as previously described (2). In addition, the Cancer Cell Line Encyclopedia (CCLE) portal (https://portals.broadinstitute.org/ccle) and Genotype-Tissue Expression (GTEx) portal were used to obtain data for 933 cell lines and 14,674 healthy samples, respectively (3, 4).

RNA-seq analysis

Total RNA from duplicates was extracted from MDA-MB-231 LM2 cells with control or MARCO-TST overexpression by using TRIzol (Qiagen). Library preparation and sequencing were performed by BGI as paired-end 100-bp reads. Samples were sequenced on NovaSeq 6000 sequencing system. The raw data were subjected to quality control analyses using Seqtk. The RNA-seq data were then mapped to the human reference genome hg38 with TopHat2. The level 3 data of RNA-seq data has been provided in Dataset S3.

Identification and quantification of splice junctions from RNA-seq data

All previously downloaded RNA-seq data (FUSCCTNBC [360 tumors and 88 paired non-tumor samples], TCGA [9,581 tumors and 2,411 paired non-tumor samples], GTEx [14.674 normal samples], CCLE [933 cell lines]) were individually aligned using a two-pass method with the STAR algorithm. StringTie (version 1.2.3) was used to assemble the BAM files for all samples (5). Next, splice junctions were identified, quantified, and normalized using the Assembling Splice Junctions Analysis (ASJA) software package (6). ASJA package identifies and characterizes three types of splice junctions (linear, back-splice, and fusion junctions). Linear junctions were extracted from assembled transcripts representing for canonical splicing events, while noncanonical splicing processes, resulting in circular RNAs (back-splice junctions) and fusion gene (fusion junctions), can be identified from chimeric transcripts. The expression level of junctions was calculated according to the following formula: $CPT(AB) = \min(\sum COV(A), \sum COV(B)) \times 1e10^7/TC$. where

COV(A) and COV(B) represent the coverage of both ends of the exon and TC is the total coverage of annotated junctions.

Identification and annotation of tumor-specific transcripts

We defined tumor-specific junctions (TSJs) as follows: (1) max_PT = 0 & max_normal = 0 and (2) median_tumor/max_PT > 10 in tumor and paired adjacent non-tumor samples. Max_PT is the maximal expression value of paired non-tumor samples in the FUSCCTNBC (n = 88) and TCGA cohorts (n = 2411). Max_normal is the maximal expression value of normal samples in the GTEx cohort (n = 14674). Median_tumor is the median expression value of tumor samples. We proceeded with the above criteria to analysis both the FUSCCTNBC and TCGA cohorts (Dataset S1). TSJs were used to extract transcripts from the StringTie results, and the transcripts were annotated with their gene positions (GENCODE version 22) using bedtools intersect (Dataset S2). The expression of TSTs was calculated by the mean expression value of junctions annotated to this transcript. The expression frequency of TSTs was defined as the frequency of patients with positive expression of TST. TST with an expression level > 0.8 CPT was defined as positive expression.

Joint analysis of DNA and RNA alterations

We mainly compared the alteration frequency between nonsynonymous mutations, CNAs and alternative splicing within a single gene (7). Nonsynonymous mutations can occur several times within a single gene body, so these events were denoted as 1 if they occurred at least once within a

gene–sample pair. For copy number, a value of 1 was given to copy-number alterations defined by GISTIC as +2 or -2. Similar to mutations, alternative splicing events were denoted as 1 if they occurred once within a gene–sample pair.

Gene set enrichment analysis (GSEA)

GSEA (version 2.2.0)(8) was performed using the hallmark, GO, and chemical and genetic perturbation gene sets and a preranked differential expression gene list. Differential gene analysis between the TST frequency high and low groups was conducted using the DEseq2 package, and the genes were ranked according to the log2-fold change of the high/low group to indicate the most to least enriched genes.

ChIP-seq data analysis and super-enhancer calling

The H3K4me1 ChIP-seq data of HCC1599 cells were obtain from the Gene Expression Omnibus (GSE116871). The H3K27ac ChIP-seq data of MCF7, T47D, MDA-MB-231, CAL51 and SUM159 cells were obtain from GSE69112. The ChIP-seq data of macrophages were obtained from GSE109440. Raw sequencing data was mapped to the hg38 build of the human genome with Bowtie2 v2.3.0 with default settings and the parameters –p 4 –k 1. Mapped reads were filtered to remove duplicate reads, and to regions in the ENCODE blacklist. MACS v1.4 was used for peak identification with a p value cutoff of 1e-6 (9).

Super-enhancers were called based on H3K27ac in HCC1599 cells using the ROSE algorithm (Rank Ordering of Super-Enhancers) (stitching distance 12.5kb and transcriptional start site [TSS] exclusion zone size 2.5kb) (10). Other parameters were set to default.

Cell culture

The human breast epithelial cell line MCF10A; human breast cancer cell lines MCF7, T47D, HCC1143, HCC1187, HCC1395, HCC1806, HCC1937, HCC38, HCC70, Hs-578T, MDA-MB-231, MDA-MB-436, MDA-MB-453, MDA-MB-468, BT-549, and BT-20; and human embryonic kidney (HEK) 293T cells were purchased from the American Type Culture Collection and cultured as suggested by ATCC's guidelines. LM2 was kindly provided by Dr. Kang from Princeton University and cultured in Dulbecco's modified Eagle's medium (DMEM) with 10% FBS (11).

Full-length sequence of MARCO-TST

The full-length cDNA of HCC1599 cells was reverse transcribed using the HiScript III 1st Strand cDNA Synthesis Kit (Vazyme). Full-length MARCO-TST was amplified using Phanta master mix (Vazyme) with the primers listed in Table S2 and analyzed by agarose gel electrophoresis. The specific DNA band at 1900 bp was extracted using the Gel DNA Extraction Mini Kit (Vazyme) and cloned into the pCE2 TA/Blunt-Zero Vector (Vazyme) for Sanger sequencing.

Plasmids and cloning procedures

The coding sequences of MARCO-TST and MARCO-TST-WT were cloned from the cDNA of HCC1599 and THP-1 cells, respectively. Cloned sequences flanked by the indicated homologous sequence were inserted into pCDH-CMV or pcDNA3.1 plasmids. The Flag tag was fused to the N-terminus of MARCO-TST and inserted into the pCDH-CMV and pcDNA3.1 plasmids. For human wild-type PLOD2, cloned sequences were flanked by the indicated homologous sequence and fused with a Flag or HA tag and inserted into the pcDNA3.1 backbone. Human wild-type and mutant HIF1A plasmids were gifts from Dr Jia (12). For the MARCO-TST P-Luc, the -1630 ~ +628 bp region of the MARCO-TST promoter was cloned into the pGL3-basic reporter vector (Promega). To generate the MARCO-TST P-E reporter, five enhancer loci were cloned with the primers listed in Table S2 and flanked with the XbaI homologous sequence. Then, five fragments were cloned into the BRCA1 P-Luc luciferase reporter via XbaI restriction sites.

Lentivirus production and transduction of cell lines

To produce lentiviral particles, HEK293T cells were seeded into one 10-cm dish and incubated overnight to reach approximately 80% confluence before transfection. Transfection was performed using polyethyleneimine linear (PEI, MW 25,000, POLYSCIENCES) according to the recommended protocol. Then, 4.8 mg of the pCDH, pCDH-Flag-TST and pCDH-Flag-WT plasmids; 3.6 mg of psPAX2 (Addgene, #12260); and 1.44 mg of pVSV-G (Addgene, #8454) were used for each 10-cm dish. After transfection for 8-12 h, the medium was

changed to fresh DMEM containing 10% FBS, and the cells were incubated for another 48 h. Culture medium containing the lentiviral particles was collected and filtered through a 0.45-mm filter to remove any remaining cells and debris. The target cells were infected for 24 h with lentiviral particles in the presence of 10 μg/ml polybrene and screened with 1 μg/ml puromycin to establish stable cells.

siRNA and plasmid transfection

For siRNA transfection, the cells were transfected with 35 nM siRNA for MARCO-TST, PLOD2, or PLOD3 or a control siRNA using Lipofectamine RNAiMAX (Thermo Fisher) according to the manufacturer's instructions. Total RNA was isolated 48-72 h later for qRT-PCR analysis. The siRNA sequences are listed in Table S2. Total RNA was isolated 48-72 h later for qRT-PCR analysis. For plasmid transfection, the cells were transfected with plasmid using Lipofectamine 2000 (Thermo Fisher) or PEI as suggested.

RNA isolation and quantitative real-time PCR (qRT-PCR) assay

Total RNA was isolated from cultured cells using TRIzol (Invitrogen) according to the manufacturer's instructions. cDNA was synthesized from 1 μg of total RNA using HiScript III RT SuperMix for qPCR (Vazyme). Real-time PCR was performed with SYBR qPCR Master Mix (Vazyme) or AceQ qPCR Probe Master Mix (Vazyme) for TaqMan probes. cDNA amplification and quantification were performed on a QuantStudio 6 Flex Real-Time PCR System (Applied Biosystems). Relative gene expression was determined by

ΔΔCt normalization to GAPDH. The primers and probes used are listed in Table S2.

Tissue isolation and fluorescence-activated cell sorting (FACS)

Tissues were isolated with 1 mg/ml collagen I and 1 mg/ml collagen III (Sigma) in DMEM for 1-2 h in a 37°C incubator. The cell suspension was filtered through a 40 μm filter, and red blood cells were lysed with TAC buffer (0.16 M NH_4CL, 0.17 M Tris-HCL) for 5 min at RT. The single cell suspension was stained with EpCAM (BioLegend) and CD45 (BioLegend) directly conjugated antibodies in FACS buffer for 25 min on ice in the dark, followed by two washes with FACS buffer. Then, the single-cell suspension was sorted with a Moflo XDP sorter (Beckman Coulter).

Co-Immunoprecipitation (Co-IP)

Co-IP was performed as follows: briefly, cells were collected and lysed in IP lysis buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, PMSF freshly added to a final concentration of 1 mM) supplemented with 1x protease inhibitor cocktail at 4°C for 30 min. Then, the lysates were cleared by centrifugation at 12,000 g for 10 min at 4°C and quantified using a BCA protein assay kit (Thermo Fisher). The supernatants were incubated with anti-FLAG antibody-conjugated magnetic beads (Sigma) or anti-HA antibody-conjugated magnetic beads (Sigma) at 4°C overnight. Then, the beads were washed at least 5 times with lysis buffer and boiled for 10 min with 1x SDS loading buffer to elute the binding proteins.

Western blotting

To prepare whole-cell lysates, the cells were lysed with 1% SDS lysis buffer (50 mM Tris pH 8.1, 1 mM EDTA, 1% SDS, 1 mM fresh dithiothreitol, sodium fluoride, and leupeptin) supplemented with protease and phosphatase inhibitor cocktail (Roche), resolved by SDS-PAGE under denaturing conditions and transferred onto 0.45-μm PVDF membranes (Millipore). The membranes were blocked with 10% nonfat milk in 1x TBST (0.9% NaCl, 10 mM Tris-HCl, pH 7.5, containing 0.05% Tween 20) at RT for 1 h and incubated with primary antibody overnight at 4°C followed by incubation with horseradish peroxidase-conjugated secondary antibodies (Jackson ImmunoResearch) for 1 hr at RT. Specific bands were visualized with enhanced chemiluminescence substrate (Millipore) and exposed onto an Amersham Imager 600 (GE Healthcare). Antibody used in western blotting were as follows: Flag (Sigma), HA, V5, PLOD1 (Proteintech, 12475-1-AP), PLOD2 (Proteintech, 66342-1-Ig), PLOD3 (Proteintech, 11027-1-AP), HIF-1α (Abcam, ab51608), Collagen IV (Abcam, ab6586), Hydroxy-HIF-1α (P564) (Cell Signaling Technology, 3434), , GAPDH.

Stable-isotope labelling by amino acids in cell culture (SILAC)

Light (Lys^0Arg^0) and heavy (Lys^8Arg^{10}) DMEM were prepared as previously described (13). HEK293T cells cultured in light or heavy medium for 5 passages before heavy and light isotope-labeled cells were transfected with pcDNA3.1-Flag-MAROC-TST and pcDNA3.1-Flag plasmids, respectively.

After 48 h of transfection, both cell lines were lysed on ice in IP lysis buffer with protease inhibitor, followed by centrifugation. The lysates were incubated with Flag antibody-conjugated beads (Sigma). The binding protein was eluted with Flag peptide (Sigma). Equal concentrations of proteins in the eluted material were pooled together and concentrated with Amicon Ultra-0.5 10K (Millipore). Then, the concentrated binding protein was resolved on an SDS-PAGE gel prior to band excision and mass spectrometry analysis.

LC-MS/MS

Then, mass spectrometry was performed on an EASY-nLC 1000 system (Thermo Fisher Scientific) connected to an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific) equipped with an online nanoelectrospray ion source. The MS raw file was processed using MaxQuant software (version 1.5.2.8, http://www.maxquant.org/) for protein identification and quantification. The MS data were searched with the Andromeda search engine against the human UniProt database. Candidate bound proteins were filtered as follows: (1) at least two unique peptides were detected, and (2) the false discovery rate was less than 1% at both the peptide and protein levels.

Immunofluorescence

Cells were fixed with 4% paraformaldehyde for 15 min at RT and then blocked with 5% BSA in PBS with 0.3% Triton X-100 (Sigma) at RT for 1 h. Primary antibodies were incubated at 4°C overnight. Antibody dilutions were as follows: Flag (Sigma, 1:200) and PLOD2 (Proteintech, 1:200). Then, coverslips were

mounted in ProLong Gold Antifade Mountant with DAPI (Life Technologies). Images were acquired using a Lecia SP5 laser-scanning confocal microscope and LAS AF software (Leica).

Secreted protein isolation from conditioned medium samples

Cells were seeded in 10-cm plates until they reached 80-90% confluence, and then serum-free medium was added. Conditioned medium samples were collected 24 h later, filtered through a 0.45-μm filter, and concentrated consecutively using Amicon Ultra-15 50K centrifugal filters (Dayer and Stamenkovic, 2015).

Chromatin immunoprecipitation (ChIP)

The ChIP assay was performed as previously described (Shu et al., 2016). Briefly, 1 x 10^7 cells were cross-linked with a final concentration of 1% formaldehyde in growth medium for 10 min at RT, and cross-linking was quenched by the addition of glycine to a final concentration of 125 mM and incubation for 5 min at RT. The cells were rinsed twice with cold PBS, harvested in 1 ml of ChIP LB1 (50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40 and 0.25% Triton X-100) supplemented with 1x protease inhibitor cocktail and lysed on ice for 10 min. The cells were pelleted, resuspended in 1 ml of LB2 buffer (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, and 0.5 mM EGTA), and mixed for 5 min. Cells were pelleted and resuspended in 1 ml of LB3 buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.1% SDS) and sonicated at a Bioruptor ultrasonicator (Diagenode) to

shear the chromatin to yield DNA fragments. The lysate was cleared by centrifugation at 12,000 g for 10 min at 4°C. Then, 100 µl of 10% Triton X-100 and 30 µl of 5 M NaCl were added. The sample was then incubated with primary antibodies overnight at 4°C (BRD4, Cell Signaling Technology, 13440S; H3K4me3, Cell Signaling Technology, 9751S; H3K27ac, active motif, 39133; H3K4me1, Cell Signaling Technology, 5326S; H3K4me1, Millipore, 05-623). The cross-linked complexes were incubated with Dynabeads Protein G for 2 h at 4°C. The beads were washed 3 times with cold ChIP LB1 and 3 times with ChIP wash buffer. DNA was eluted in elution buffer (100 mM sodium bicarbonate and 1% SDS). Cross-links were reversed overnight at 65°C RNA and protein were digested with 0.2 mg/ml RNase A for 30 min at 37°C followed by 0.2 mg/ml Proteinase K for 1 h at 65°C. DNA was purified with the ChIP DNA Clean Kit (ZYMO). ChIP-qPCR was performed using SYBR qPCR Master Mix (Vazyme). The primers are listed in Table S2. ChIP-seq libraries were prepared using the KAPA Hyper DNA Kit (KAPA Biosystems, KK8503) according to the manufacturer's instruction and sequenced on an NovaSeq 6000 platform. The ChIP-seq data of HCC1599 cells have been deposited in the GSE214133.

***In vitro* proliferation, migration and invasion**

For the proliferation assay, cells were seeded at 1,000-3,000 cells/well in 96-well culture plates, and Cell Counting Kit-8 (CCK-8) assays (Vazyme) were performed every 2 days following the manufacturer's guidelines. The

absorbance value was measured at 450 nm using a SpectraMax M5 (Molecular Devices). For the migration assay, 5×10^4 cells were suspended in serum-free culture medium and seeded into 24-well Transwell inserts (pore size, 8 μm; BD Biosciences). Medium with 10% serum was added to the wells. After 8 h, the migrated cells were fixed and stained with crystal violet (0.05%, Solarbio) and then imaged using a microscope. For the invasion assay, 1×10^5 cells were suspended in serum-free culture medium and seeded into 24-well Matrigel-coated Transwell inserts (pore size, 8 μm; BD Biosciences). Medium with 10% serum was added to the wells. After 16h, the invaded cells were fixed and stained with crystal violet (0.05%, Solarbio) and then imaged using a microscope.

Animal experiments

Four-week-old female NOD/SCID mice were purchased from Beijing Vital River Laboratory Animal Technology Co., Ltd. and housed under SPF conditions at the animal care facility of the Experimental Animal Center of Fudan University Shanghai Cancer Center. For xenograft models, 1×10^6 LM2 cells were orthotopically injected directly into the inguinal mammary fat pads of mice in 50 μl of sterile PBS (n = 8 in each group). Tumor were measured by caliper twice a week. The xenografts were fixed in polyformaldehyde for paraffin embedding or frozen for RNA. To investigate tumor metastasis, 8 weeks after injection, each mouse was intraperitoneally injected with 200 mg/kg D-Luciferin. Bioluminescence was analyzed with an IVIS system. All

experiments were performed in accordance with the relevant institutional and national guidelines and regulations of Shanghai Medical Experimental Animal Care Commission.

OTX015 treatment *in vivo*

For MARCO-TST positive xenograft model, $5x10^6$ HCC1599 cells were orthotopically injected directly into the inguinal mammary fat pads of NOD/SCID mice (n = 16). Tumor bearing mice were randomized into vehicle (n=8) and OTX015 group (n=8), and started dosing when average tumor volume reached 50–100 mm^3. OTX015 (50-70 mg/kg, 1%DMSO + 30% PEG300 + 1%Tween80 + ddH2O) was given orally daily for the number of days indicated. Body weight was measured twice or three times per week. Mice were euthanized and tumor tissues collected 6 hr after the last dosing and prepared for immunoblotting or immunohistochemistry staining.

Drug dose-response curves

The cells were seeded in 96-well plates. Twenty-four hours after seeding, the cells were treated with a 9-point 3-fold dilution series of OTX015 (Selleck, from 50 μM to 7.62 nM). Cells were grown in the presence of drug for 3 days. Cell viability was assessed by the CCK-8 assay following the manufacturer's recommendations. The half-maximal inhibitory concentration (IC50) was estimated, and the IC50 values from replicate experiments were compared by Student's t test.

Patient derived organoids (PDOs) culture and drug response test

Fresh breast cancer tumor tissues were delivered to the laboratory in DMEM/F12 with penicillin/streptomycin on ice in 2 hours. Tissues were washed in cold PBS for several times and then minced into tiny fragments. Then tissue fragments were digested in 10ml digestion buffer (DMEM/F12 with 5% BSA, insulin and hydrocortisone) supplemented with collagenase I (1mg/ml, Sigma), collagenase III (1mg/ml, Sigma) and hyaluronidase (0.1mg/ml, Sigma) at 37℃ for 30-60 minutes with shaking. Tumor cells were filtered through a 100μm filter and filtrate was centiguated at 300-500g for 5 minutes. Tumor cells were resuspended into basement membrane extract (Cultrex) and seeded in a pre-warmed 24-well cell culture plate. After completed gelation, 500 ul of breast cancer organoid medium was added to each well and PDOs were culture in 37℃ 5% CO2 incubators(14). For drug response test, PDOs were harvested and diluted to 50 organoids/μl in breast cancer organoid medium containing 10% BME and 25 μl organoid suspension was added into each well of clear bottom 384-well suspension culture plates (Greiner). Then, 6 concentrations of OTX015 and DMSO control were added in triplicate. After 5 days, the viability of PDOs was measured using CellTiter-Glo 3D Reagent (Promega) according to the manufacture's instructions.

Statistical analysis

RNA-seq and 450K methylation array data analyses were conducted in R, and statistical analyses were performed as described above. All other data plotting and statistical analyses were performed with GraphPad Prism version 8

(GraphPad Software, Inc.). Data are presented as the mean ± SD of replicate experiments, as indicated, and presented as individual values, scatter plots, heatmaps, box plots, and bar graphs. Significance was determined using unpaired/paired Student's t test, the Mann-Whitney test, the χ^2 test or the log-rank test (Kaplan-Meier curves), where appropriate.

Supplementary Figures

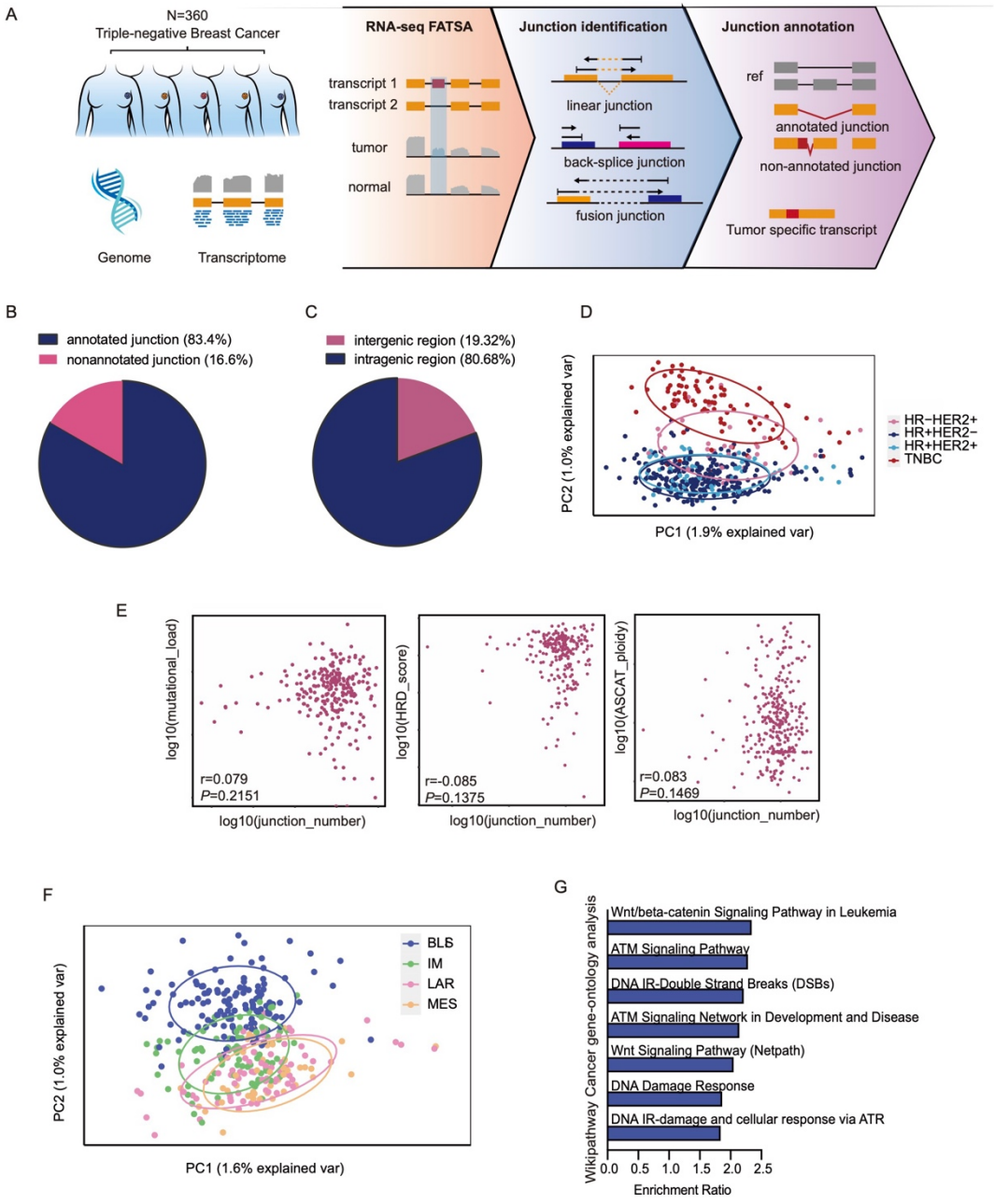

Figure S1. Detection of splicing junctions and association with genome alteration, related to Figure 1. (A) Study cohort and flow diagram of splicing junction identification in this study. **(B and C)** The pie chart shows the proportion of annotated and unannotated junctions (B) and the distribution of

genome locations (C). **(D)** Principal component analysis (PCA) of splicing junctions of breast cancer tumors in TCGA cohort (n=548) using the PSI value. **(E)** Correlation of mutational load, homologous recombination deficiency (HRD) score and ploidy with the number of splicing junctions. Pearson's correlation coefficient was used to determine the correlation. **(F)** PCA of splicing junctions of TNBC tumors in the FUSCCTNBC cohort (n=360) using the PSI value. **(G)** WikiPathway cancer analysis of genes within the top 20% positive and negative loadings of principal component one from (F).

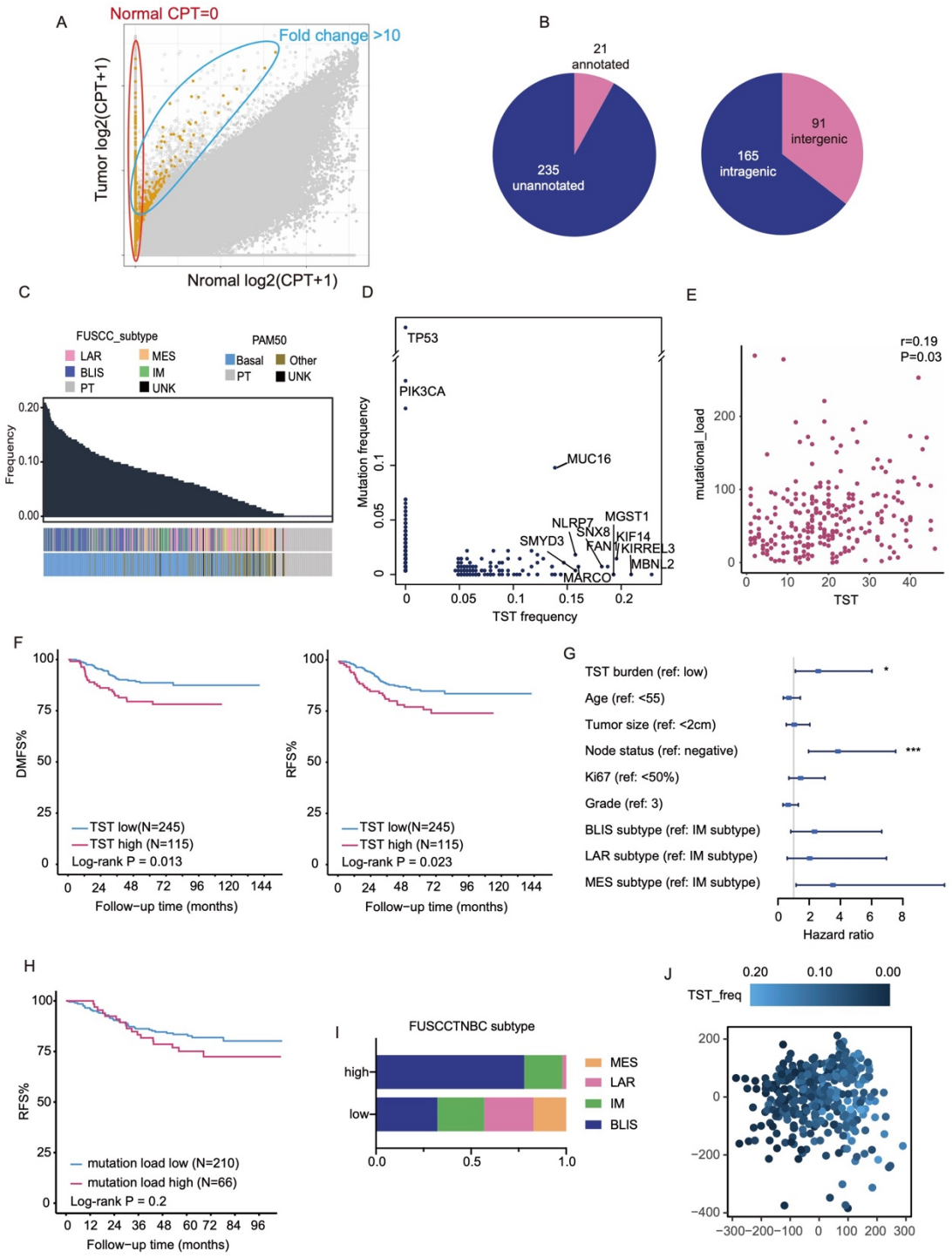

Figure S2 Clinical implications of tumor-specific transcripts (TSTs) in TNBC, related to Figure 1. (A) Comparison of the expression of splicing

junctions between tumor and paired adjacent non-tumor tissues in the FUSCCTNBC cohort. Solid points represent junction expression frequencies ≥5%, and circles represent junction expression frequencies < 5%. Yellow points represent tumor-specific junctions (CPT in paired adjacent non-tumor tissues = 0 or a fold change > 10 between tumor and paired adjacent non-tumor tissues). **(B)** The pie chart shows the proportion of annotated and unannotated tumor-specific junctions and their genome location. **(C)** Bar plot shows the TST expression frequency of each sample ordered from high to low. **(D)** The frequency of mutations and TSTs in genes. Genes of interest are highlighted. **(E)** Correlation of mutational load with the number of TSTs. Pearson's correlation coefficient was used to determine the correlation. **(F)** Kaplan–Meier plots for disease-metastasis free survival (DMFS) and relapse-free survival (RFS) of patients with high (>22 TST) and low (≤22 TST) TST abundance. P values were calculated using the log-rank test. **(G)** Forest plot of multivariate Cox regression analysis for DMFS adjusting for age, tumor size, lymph node status, Ki-67, grade, and TNBC mRNA subtypes. Data are presented as hazard ratios with 95% confidence intervals. **p < 0.01; ***p < 0.001. P values were calculated using Cox regression analysis. **(H)** Kaplan–Meier plots for RFS of patients with high and low nonsynonymous mutation frequencies. P values were calculated using the log-rank test. **(I)** Distribution of TNBC mRNA subtypes in the TST-high and TST-low groups. **(J)**

PCA of splicing junctions in TNBCs using the junction expression value. The TST frequency of each patient is shown on the graph with the indicated color.

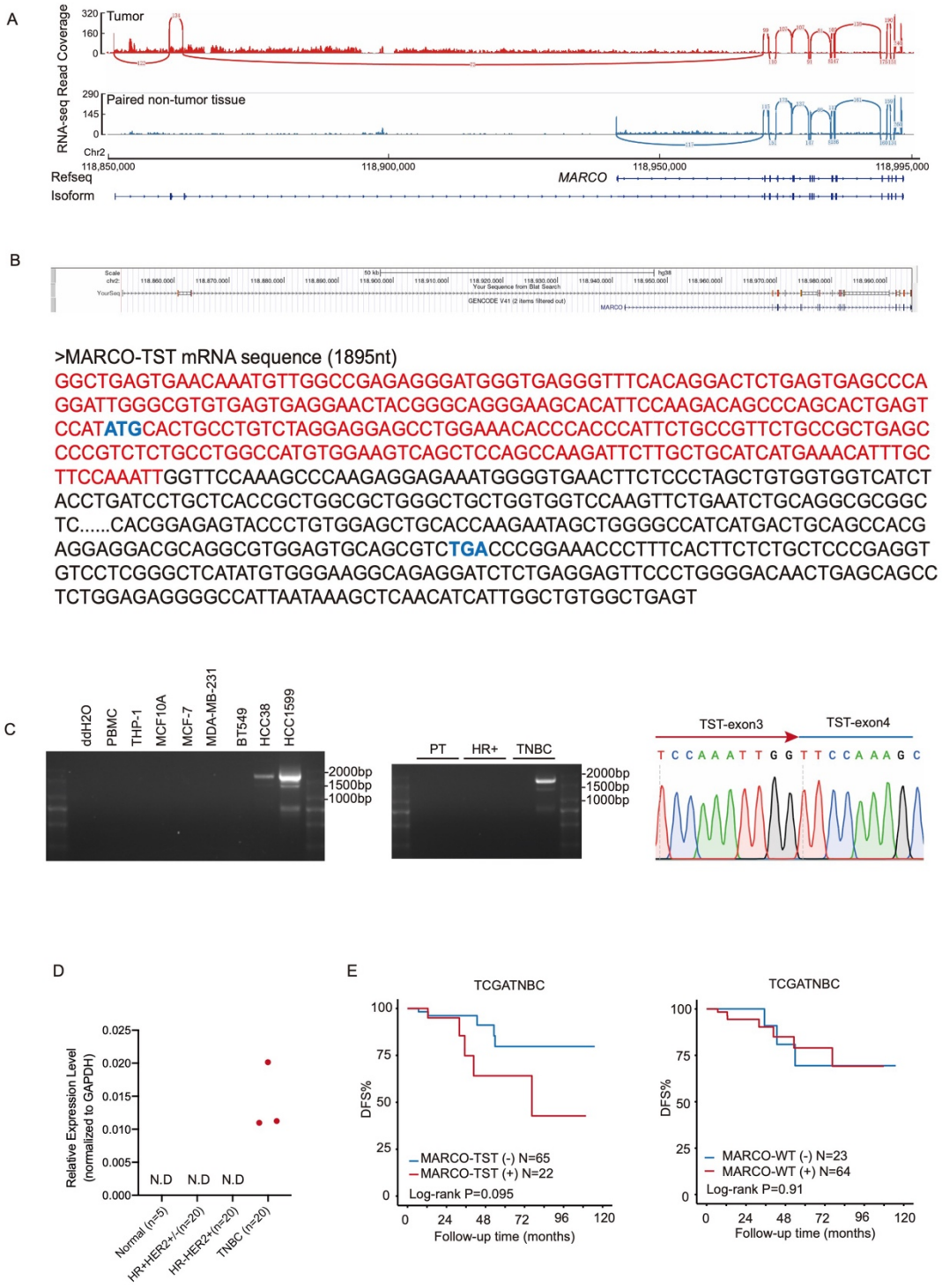

Figure S3. Identification of the MARCO-TST transcript, related to Figure 2.

(A) Sashimi plot showing the structure of MARCO variants. Tumor samples

expressing MARCO-TST are shown in red, and paired non-tumor tissues expressing MARCO-WT are shown in blue. The coverage of junctions is labeled on the graph. **(B)** The BLAST result and sequence of MARCO-TST sequence extracted from long-read RNA-seq of HCC1599 cells. The MARCO-TST transcript consists of a region (marked in red) different from the wild-type MARCO transcript. **(C)** Agarose gel electrophoresis and Sanger sequencing results of cloned full-length MARCO-TST in different cell lines and tumor tissues processed by polymerase chain reaction. PT, paired non-tumor tissues. **(D)** qRT-PCR analysis of MARCO-TST expression in tissues. Samples with cycle threshold values over 40 were defined as not detected (N.D.) **(E)** Survival analysis of MARCO-TST or -WT expression of TNBCs in TCGA cohorts. P values were calculated using the log-rank test. DFS, disease-free survival.

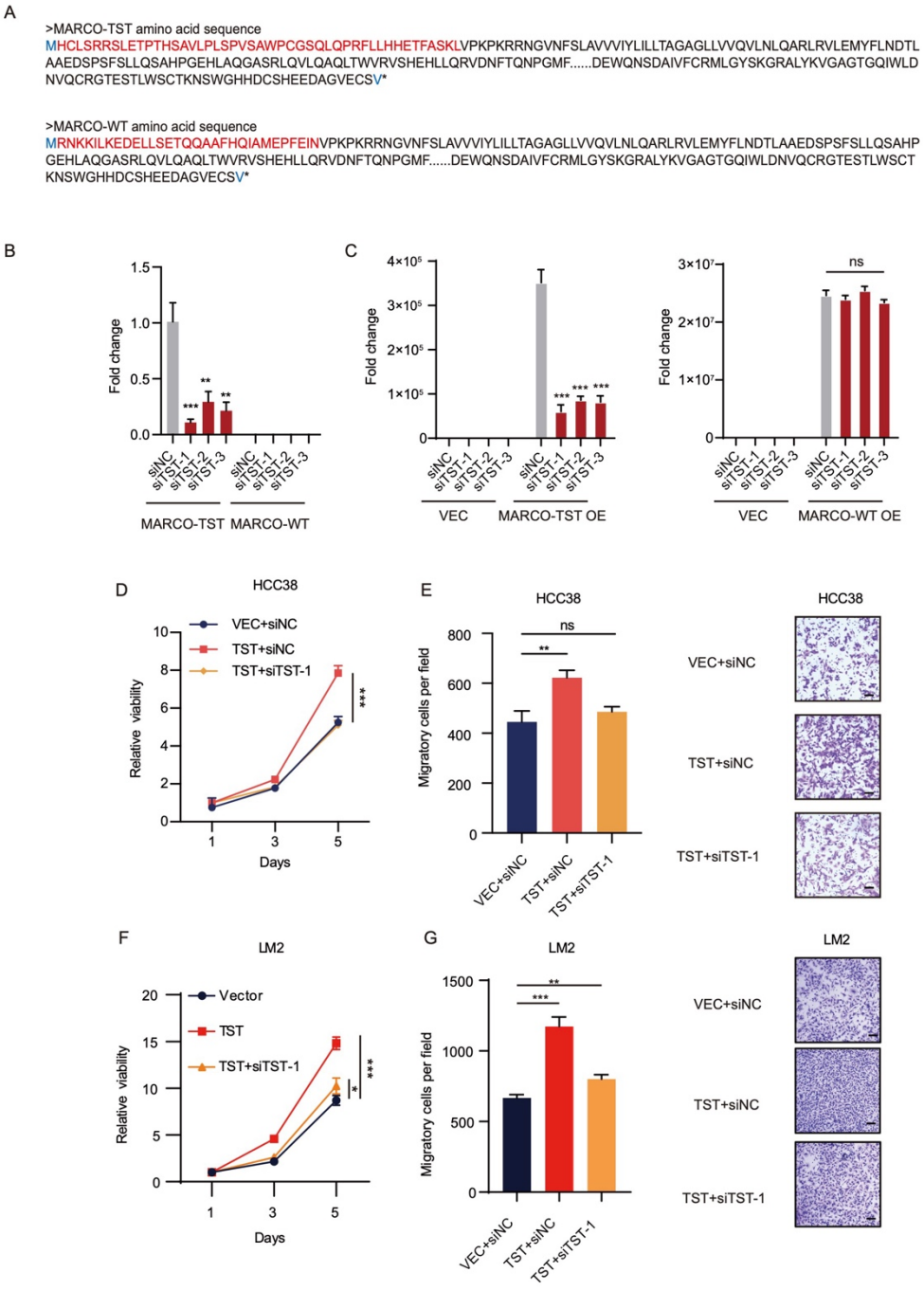

Figure S4. The amino acid sequence and function of the MARCO-TST protein, related to Figure 3. (A) The amino acid sequences of MARCO-TST and -WT. Different amino acids are marked in red. **(B)** qRT-PCR analysis of the knockdown efficacy of siRNAs targeting MARCO-TST in HCC38 cells. **(C)**

qRT–PCR analysis of LM2 cells overexpressing MARCO-TST or MARCO-WT transfected with siRNAs targeting MARCO-TST. Data are presented as the mean ± SD. *p < 0.05; **p < 0.01; ***p < 0.001. P values were calculated using two-tailed Student's t test. **(D)** The growth rate of HCC38 cells stably expressing MARCO-TST transfected with MARCO-TST siRNAs. **(E)** Quantification and representative images of migration assays of HCC38 cells stably expressing MARCO-TST transfected with MARCO-TST siRNAs. Scale bar, 100 μm. **(F)** The growth rate of LM2 cells stably expressing MARCO-TST transfected with MARCO-TST siRNAs. **(G)** Quantification and representative images of migration assays of LM2 cells stably expressing MARCO-TST transfected with MARCO-TST siRNAs. Scale bar, 100 μm. Data are represented as the mean ± SD. *p < 0.05; **p < 0.01; ***p < 0.001. P values were calculated using two-tailed Student's t test.

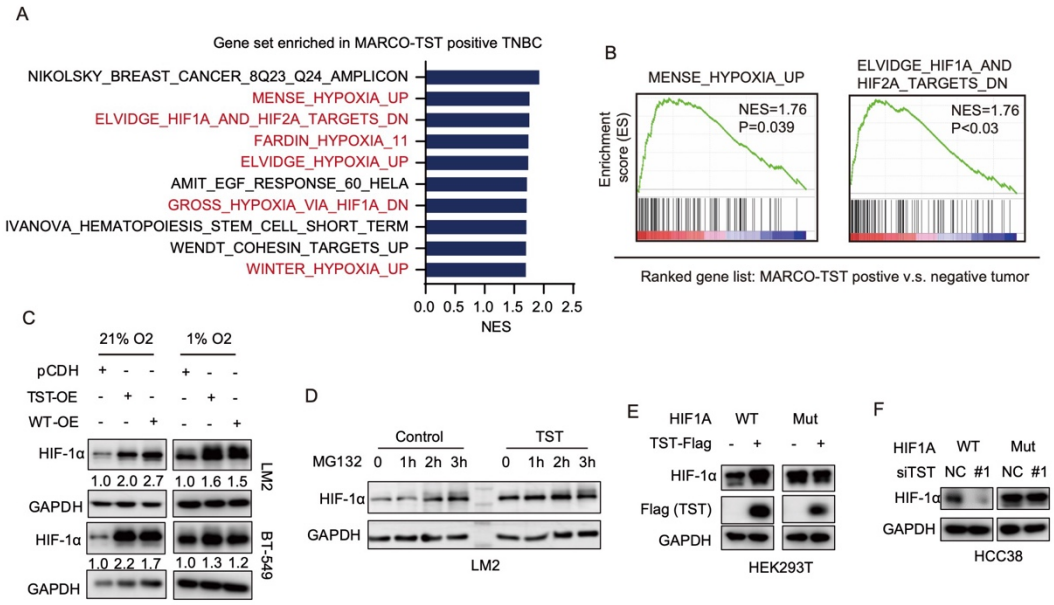

Figure S5. MARCO-TST regulates HIF-1α protein stability, related to Figure 4. (A) Gene set enrichment analysis (GSEA) of pathways significantly enriched (*P*<0.05) in MARCO-TST-positive tumors in the FUSCCTNBC cohort, showing the top ten enriched pathways. **(B)** Enrichment of hypoxia pathway gene sets in MARCO-TST-positive tumors in the FUSCCTNBC cohort. **(C)** Immunoblot analysis of HIF-1α protein levels in MARCO-TST- and MARCO-WT-expressing tumor cells under normoxic and hypoxic conditions. **(D)** Immunoblot analysis of HIF-1α levels in control and MARCO-TST-overexpressing cell lines treated with or without proteasome inhibitor (MG132) for different time periods. **(E and F)** HIF-1α protein levels were detected in MARCO-TST-expressing HEK293T cells (E) or

MARCO-TST-silenced HCC38 cells (F) transfected with wild-type HIF1A (WT) or proline hydroxylation-deficient HIF1A vector (Mut).

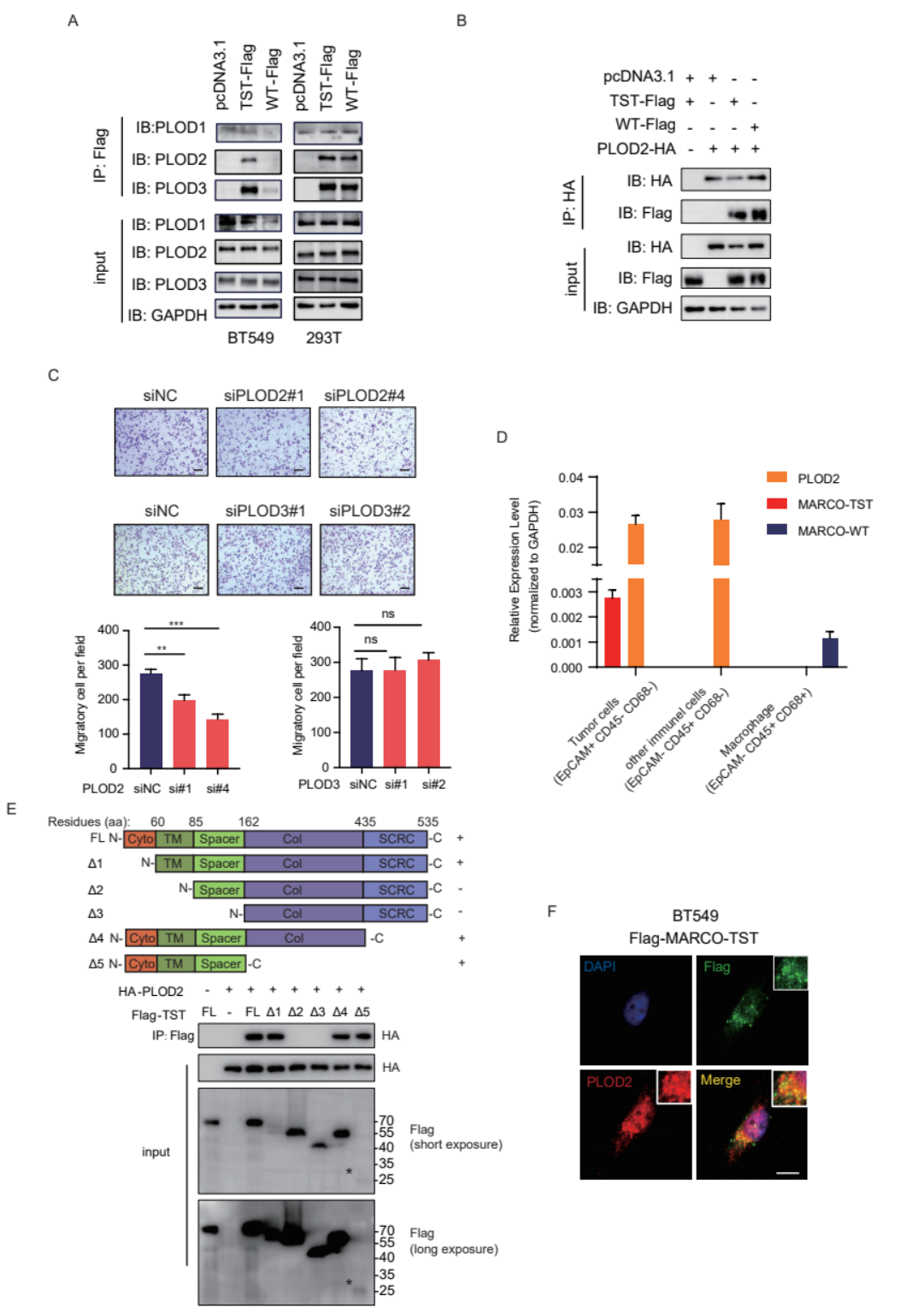

Figure S6. MARCO-TST interacts with PLOD2 via the transmembrane domain, related to Figure 5. (A) HEK293T and BT549 cells expressing Flag-tagged MARCO-TST or MARCO-WT were collected and

immunoprecipitated with anti-Flag beads. IP samples were subjected to immunoblotting with PLOD1, PLOD2 and PLOD3 antibodies. PLOD, procollagen lysyl hydroxylase. **(B)** HEK293T cells expressing Flag-tagged MARCO-TST or MARCO-WT with HA-tagged PLOD2 were harvested for IP with anti-HA beads, and proteins were assessed by western blotting as indicated. **(C)** Representative images and quantification of the migration activity of BT-549 cells transfected with control or PLOD2- or PLOD3-targeting siRNAs. Scale bar, 100 μm. Data are represented as the mean ± SD; **p < 0.01; ***p < 0.001. The P values were calculated using two-tailed Student's t test. **(D)** PLOD2 and MARCO variants expression among different cell types sorted from tumor tissues. **(E)** Schematic of MARCO-TST truncation mutants (top) and PLOD2 binding capability is shown (+). HEK293T cells were transfected with different MARCO-TST truncation mutants and immunoprecipitated with anti-Flag or anti-HA beads. IP samples were assessed by immunoblotting with the indicated antibodies (bottom). FL, full length. **(F)** Confocal fluorescence photomicrographs of BT549 cells stably expressing Flag-tagged MARCO-TST and immunostained for Flag (green) or endogenous PLOD2 protein (red). Scale bars, 10 μm.

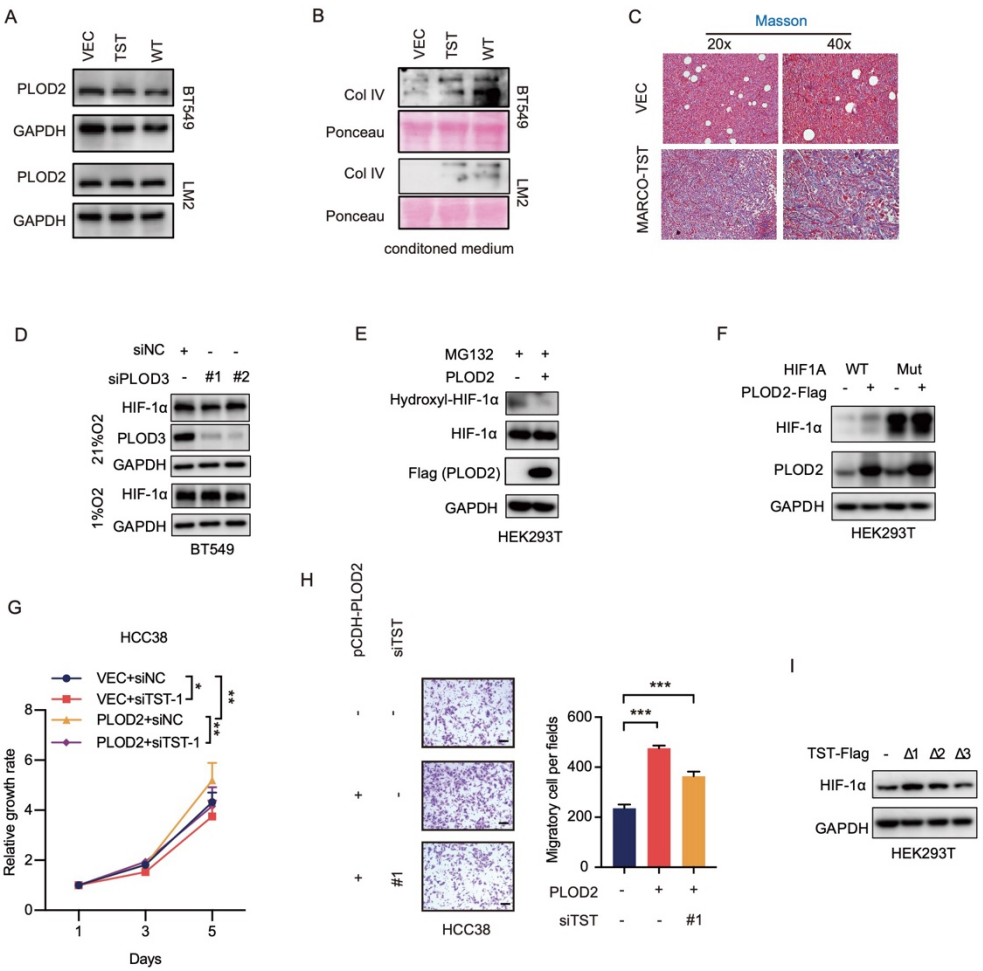

Figure S7. PLOD2 is required for the function of MARCO-TST, related to Figure 5. (A) Immunoblots of PLOD2 in BT549 and LM2 cells with MARCO-TST or -WT expression. **(B)** Immunoblots of collagen IV (Col IV) using conditioned medium from LM2 and BT549 cells with MARCO-TST or -WT expression. **(C)** Masson staining of tumors stably expressing control vector or MARCO-TST. **(D)** HIF-1α levels in BT549 and LM2 cells depleted with PLOD3. **(E)** Hydroxylation of HIF-1α in HEK293T cells expressing PLOD2. **(F)** Immunoblots of HIF-1α hydroxylation in control and PLOD2-expressing HEK293T cells transfected with HIF1A wild-type (WT) or

hydroxylation-deficient mutants (Mut). **(G and H)** Growth rate (G) and migration assay (H) of control and PLOD2-overexpressing LM2 cells transfected with MARCO-TST siRNAs. Data are represented as the mean ± SD. *p < 0.05; **p < 0.01; ***p < 0.001. P values were calculated using two-sided two-tailed Student's t test. **(I)** HIF-1α protein levels in HEK293T cells expressing different truncated proteins of MARCO-TST.

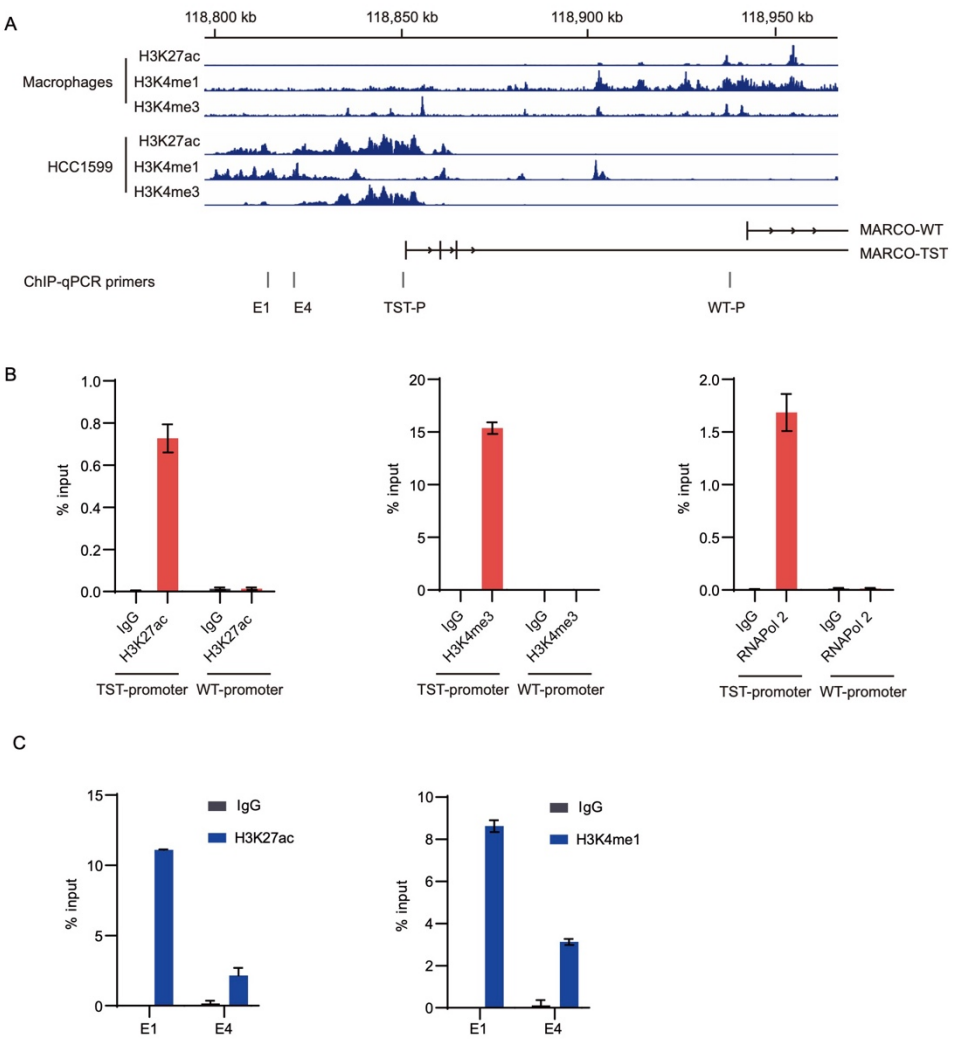

Figure S8. The proximal promoter and superenhancer of MARCO-TST, related to Figure 6. (A) Schematic of MARCO variants promoter and enhancer region showing primers used for ChIP-qPCR assay (B and C). **(B)** ChIP-qPCR of H3K27ac, H3K4me3 and RNA Polymerase II enrichment at the promoter region of MARCO-TST and -WT in HCC38 cells. Data are represented as the mean ± SD. **(C)** ChIP-qPCR of H3K27ac and H3K4me1

enrichment at the enhancer region of MARCO-TST in HCC38 cells. Data are represented as the mean ± SD.

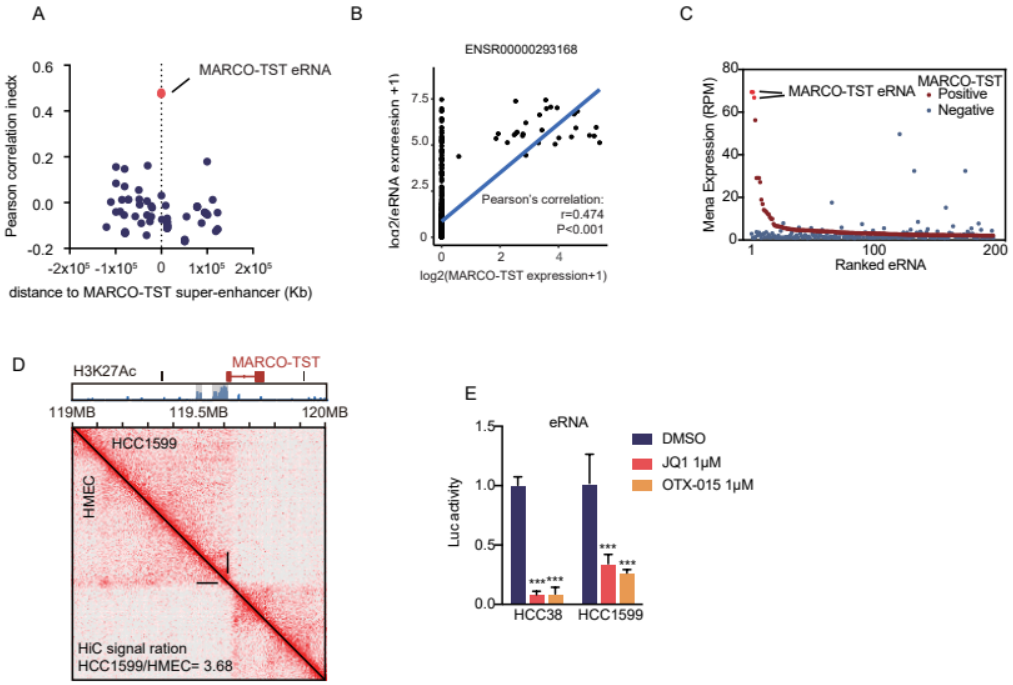

Figure S9. The MARCO-TST associated eRNAs, related to Figure 6. (A) Correlation analysis of eRNAs at chromosome 2 and MARCO-TST mRNA in breast cancer samples in the TCGA cohort. eRNAs associated with MARCO-TST were labeled (ENSR00000122296, chr2:118763939; ENSR00000293168, chr2: 118763900; and ENSR00000122295, chr2:118762005). The expression profile of eRNAs were extracted from eRic database (https://hanlab.uth.edu/) (Dataset S4). **(B)** Correlation analysis of MARCO-TST eRNA and MARCO-TST expression in breast cancer samples in the TCGA database. Pearson's correlation coefficient was used to determine the correlation. **(C)** Rank of MARCO-TST eRNAs expression levels in MARCO-TST-positive and -negative tumors in the TCGA cohort. **(D)** Hi-C contact matrices for the MARCO-TST locus in HCC1599 cells (upper right,

GSE116872) versus HMECs (lower left, GSE63525). MARCO-TST promoter-enhancer contacts are underlined in the matrices, and the ratio of Hi-C signals between HCC1599 cells and HMECs (3.68) is indicated. H3K27ac ChIP-seq tracks of HCC1599 cells are shown, and the enhancer region is shadowed. **(E)** qRT–PCR analysis of MARCO-TST eRNA levels in HCC38 or HCC1599 cells treated with 1 μM JQ1 or 1 μM OTX015 for 24 h. Data are represented as the mean ± SD. ***p < 0.001. P values were calculated using two-sided two-tailed Student's t test.

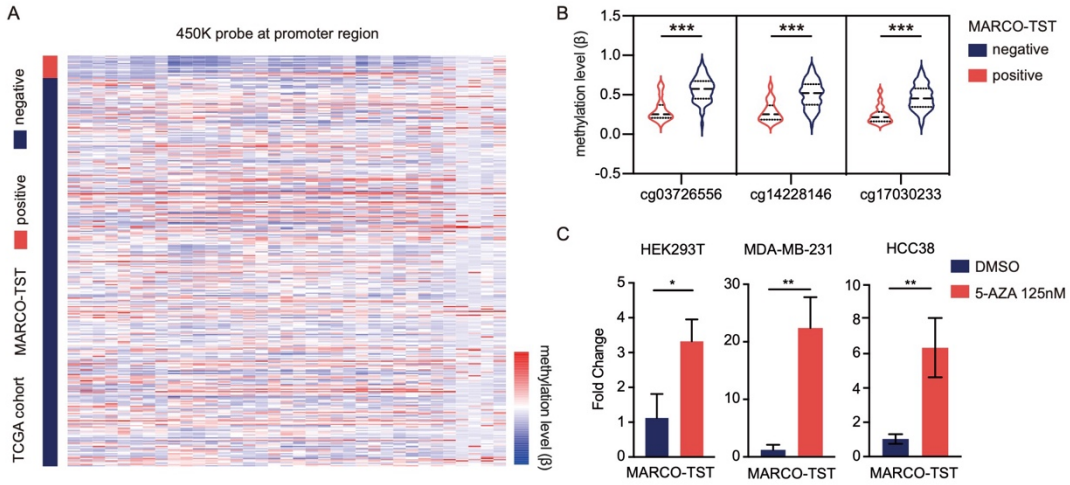

Figure S10. Demethylation of the MARCO-TST promoter is associated with its expression, related to Figure 6. (A) Heatmap with β-values of DNA methylation obtained from HM450K probes at MARCO-TST promoter loci in breast tumors from TCGA database. **(B)** Violin plot of DNA methylation β-values of HM450K probes at the promoter of MARCO-TST. **(C)** qRT-PCR analysis of MARCO-TST mRNA levels in MARCO-TST-negative (HEK293T, MDA-MB-231) and MARCO-TST-positive (HCC38) cell lines treated with or without 5-AZA (125 nM). *p < 0.05; **p < 0.01; ***p < 0.001. P values were calculated using two-sided two-tailed Student's t test.

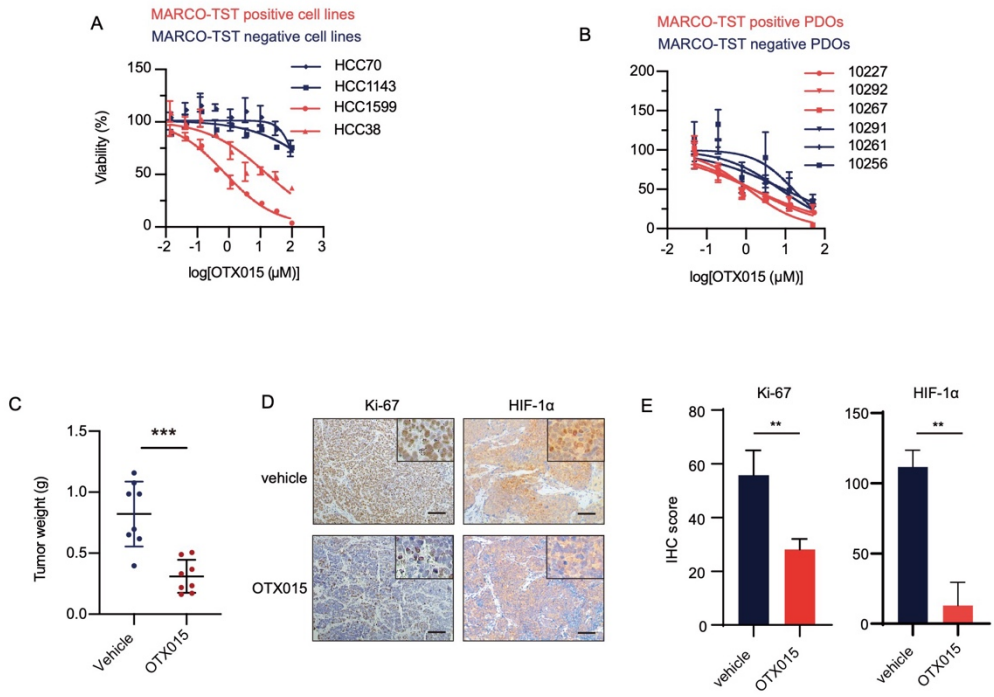

Figure S11. BET inhibitor activity in MARCO-TST expression TNBC tumors, related to Figure 6. (A and B) Dose–response curves of TNBC cell lines (A) and patient-derived organoid models (B) treated with OTX015. PDO, patient-derived organoid. **(C)** Tumor weight of HCC1599 cell-implanted NOD/SCID mice (n=8) treated with vehicle (n=8) or OTX015 (n=8). **(D)** Immunohistochemistry staining images (Ki-67 and HIF-1α) of tumor sections from mice treated with vehicle or OTX015. Scale bar, 20 μm. **(E)** Quantification of Ki-67 and HIF-1α levels in tumor sections from mice treated with vehicle or OTX015. Data are represented as the mean ± SD. **p < 0.01. P values were calculated using two-sided two-tailed Student's t test.

Table S1. Clinicopathological correlation of MARCO-TST in TNBC

	MARCO-TST		P value
	Negative (N=301) N (%)	Positive (N=59) N (%)	
Age			
≦55	167 (55.5)	41 (69.5)	0.065
>55	134 (44.5)	18 (30.5)	
Tumor size			
≦2cm	111 (36.9)	20 (33.9)	0.818
>2cm	189 (62.8)	39 (66.1)	
unknown	1 (0.3)	0 (0.0)	
Node status			
negative	178 (59.1)	37 (62.7)	0.357
positive	122 (40.5)	21 (35.6)	
unknown	1 (0.3)	1 (1.7)	
Ki-67			
≦50%	148 (49.2)	19 (32.2)	0.011
>50%	143 (47.5)	40 (67.8)	
unknown	10 (3.3)	0 (0.0)	
Grade			
2 to 3	87 (28.2)	8 (13.6)	0.012
3	184 (61.1)	48 (81.4)	
unknown	32 (10.6)	3 (5.1)	
Liver_metastasis			
YES	291 (96.7)	57 (96.6)	1.000
NO	10 (3.3)	2 (3.4)	
Bone_metastasis			
YES	285 (94.7)	54 (91.5)	0.520
NO	16 (5.3)	5 (8.5)	
Lung_metastasis			
YES	292 (97.0)	52 (88.1)	0.007
NO	9 (3.0)	7 (11.9)	
Brain_metastasis			
YES	298 (99.0)	56 (94.9)	0.092
NO	3 (1.0)	3 (5.1)	
mRNA_Subtype			
BLIS	96 (31.9)	43 (72.9)	<0.001
IM	73 (24.3)	14 (23.7)	
LAR	81 (26.9)	0 (0.0)	
MES	51 (16.9)	2 (3.4)	

Table S2, Sequences of DNA oligos, primers, and siRNAs.

Sequences for siRNAs	**5'-3'**
siNC	GUGCGUUGUUAGUACUAAU
siMARCO-TST-1	GGAAACACCCACCCAUUCU
siMARCO-TST-2	GCUGCAUCAUGAAACAUUU
siMARCO-TST-3	CCAAGAUUCUUGCUGCAUC
siPLOD2-1	GGUUGUCAUGUUUACUGAA
siPLOD2-4	CCGUAUAUCUGGUGGUUAU
siPLOD3-1	CUUCCUCAAUUCUGGUGGAUU
siPLOD3-2	UGGAUAGCUACGACGUGAUUC
Primers for RT-PCR	
GAPDH-RT-F	GTCTCCTCTGACTTCAACAGCG
GAPDH-RT-R	ACCACCCTGTTGCTGTAGCCAA
PLOD2-RT-F	GACAGCGTTCTCTTCGTCCTCA
PLOD2-RT-R	CTCCAGCCTTTTCGTGGTGACT
HIF1A-RT-F	TATGAGCCAGAAGAACTTTTAGGC
HIF1A-RT-R	CACCTCTTTTGGCAAGCATCCTG
LOX-RT-F	GATACGGCACTGGCTACTTCCA
LOX-RT-R	GCCAGACAGTTTTCCTCCGCC
CA9-RT-F	GTGCCTATGAGCAGTTGCTGTC
CA9-RT-R	AAGTAGCGGCTGAAGTCAGAGG
LDHA-RT-F	GGATCTCCAACATGGCAGCCTT
LDHA-RT-R	AGACGGCTTTCTCCCTCTTGCT
VEGFA-RT-F	TTGCCTTGCTGCTCTACCTCCA
VEGFA-RT-R	GATGGCAGTAGCTGCGCTGATA
MARCO-TST-eRNA-R T-F	AGGACATTGGCAAGGACCAA
MARCO-TST-eRNA-R T-R	CCATGATGGAAACCGTGTGC
Primers for ChIP-qPCR validation	
MARCO-TST-E1-F	CCTCTTCAGTAGCACACCCA
MARCO-TST-E1-R	GCTTGCTTGCGGGATAAAGA
MARCO-TST-E4-F	ACAATGGGCAGATGGAAGGA
MARCO-TST-E4-R	CTACATAAGGGTTTGGCGGC
MARCO-TST-P1-F	AGTGGAGCCCGGATTTGTTA
MARCO-TST-P1-R	TTTGCAGGAGGGAGAAGAGG
Sequence of taqman probe	
MARCO-TST (FAM)	ACAGCCCAGCACTGAGTCCA
MARCO-TST-Taq-F	GTGAGTGAGGAACTACGGGC
MARCO-TST-Taq-R	TCCAGGCTCCTCCTAGACAG
MARCO-WT (FAM)	CTGGCTCCAGGACTTTGGCC

MARCO-WT-Taq-F	AGGAGGACGAGCTCTTGAGT
MARCO-WT-Taq-R	TGACTGCAGCAAGGAGAAGG
GAPDH (VIC)	CCTGGTCACCAGGGCTGCTTTTAA
GAPDH-Taq-F	ACGGATTTGGTCGTATTGGGC
GAPDH-Taq-R	TTGACGGTGCCATGGAATTTG

Primers for expression plasmids

MARCO-TST-Full-F	AGACAGCCCAGCACTGAGTC
MARCO-TST-Full-R	CGGGAGCAGAGAAGTGAAAG
pcDNA3.1-MARCO-TST-F	TTGGTACCGAGCTCGGATCCAATGCACTGCCTGTCTAGGAGG
pcDNA3.1-MARCO-TST-R	CCACTGTGCTGGATATCTGCAGAATTCGGCCGCTACTTGTCATCGTCATCC
pcDNA3.1-MARCO-WT-F	TTGGTACCGAGCTCGGATCCAATGAGAAATAAGAAAATTCTCAAGGAGGA
pCDH-MARCO-TST-F	ATTCTAGAGCTAGCGAATTCATGCACTGCCTGTCTAGGAGG
pCDH-MARCO-WT-F	ATTCTAGAGCTAGCGAATTCATGAGAAATAAGAAAATTCTCAAGGAGGA
pCDH-MARCO-TST-R	TGGTCTTTGTAGTCGGATCCGACGCTGCACTCCACGCCT
pcDNA3.1-MARCO-TST-short1-F	TTGGTACCGAGCTCGGATCCATCCCTAGCTGTGGTGGTCATCTA
pcDNA3.1-MARCO-TST-short2-F	TTGGTACCGAGCTCGGATCCAAATCTGCAGGCGCGGCTCC
pcDNA3.1-MARCO-TST-short3-F	TTGGTACCGAGCTCGGATCCAGGTGAACAAGGCGCCCCAGG
pcDNA3.1-MARCO-TST-short4-R	CGTCATGGTCTTTGTAGTCGAATTCTCAGTTTTCACCTCTTTCACCTTTTTC
pcDNA3.1-MARCO-TST-short5-R	CGTCATGGTCTTTGTAGTCGAATTCTCAGCCTTGTTCACCTTTGATTCT
pcDNA3.1-PLOD2-F	TTGGTACCGAGCTCGGATCCAATGGGGGGATGCACGGT
pcDNA3.1-PLOD2-R	CGTCATGGTCTTTGTAGTCGAATTCGGGATCTATAAATGACACTGCAATGTATCT
pCDH-PLOD2-F	ATTCTAGAGCTAGCGAATTCATGGGGGGATGCACGGTGA
pCDH-PLOD2-R	TGGTCTTTGTAGTCGGATCCGGGATCTATAAATGACACTGCAATGTATCTT

Primers for luciferase expression plasmids

pGL3-P-F	CCAGAACATTTCTCTATCGATAGGTACCTACCCAAACCCCAGAA
pGL3-P-R	ACTTAGATCGCAGATCTCGAAGGGGACTGGAAGCG

pGL3-P-E1-F	GTGGTAAAATCGATAAGGATCCGGGGTGGACACTGTTGTTTAGC
pGL3-P-E1-R	TCTCAAGGGCATCGGTCGACGAGTCCCTTGTGAGATTAATCACCC
pGL3-P-E2-F	GTGGTAAAATCGATAAGGATCCCCATTTGTTTGCTACAAGTTTCAGC
pGL3-P-E2-R	TCTCAAGGGCATCGGTCGACGCTGTTTTTAGTCGCTACCCCCT
pGL3-P-E3-F	GTGGTAAAATCGATAAGGATCCGGCCTCAGAAGGGGAGCT
pGL3-P-E3-R	TCTCAAGGGCATCGGTCGACTGGGCTCTCCTGTTGACCAG

SI References

1. Y.-Z. Jiang, *et al.*, Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies. *Cancer Cell* **35**, 428-440.e5 (2019).

2. Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

3. GTEx Consortium, *et al.*, Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

4. M. Ghandi, *et al.*, Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).

5. M. Pertea, *et al.*, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).

6. J. Zhao, *et al.*, ASJA: A Program for Assembling Splice Junctions Analysis. *Comput Struct Biotechnol J* **17**, 1143–1150 (2019).

7. C. Calabrese, *et al.*, Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).

8. A. Subramanian, *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–15550 (2005).

9. X. Zhang, *et al.*, Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat Genet* **48**, 176–182 (2016).

10. W. A. Whyte, *et al.*, Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).

11. M. Shen, *et al.*, Tinagl1 Suppresses Triple-Negative Breast Cancer Progression and Metastasis by Simultaneously Inhibiting Integrin/FAK and EGFR Signaling. *Cancer Cell* **35**, 64-80.e7 (2019).

12. X. Jia, *et al.*, Basal and therapy-driven hypoxia-inducible factor-1α confers resistance to endocrine therapy in estrogen receptor-positive breast cancer. *Oncotarget* **6**, 8648–8662 (2015).

13. X. Jin, *et al.*, The endogenous retrovirus-derived long noncoding RNA TROJAN promotes triple-negative breast cancer progression via ZMYND8 degradation. *Sci. Adv.* **5**, eaat9820 (2019).

14. N. Sachs, *et al.*, A Living Biobank of Breast Cancer Organoids Captures Disease Heterogeneity. *Cell* **172**, 373-386.e10 (2018).