ARTICLE OPEN

Check for updates

X-linked cancer-associated polypeptide (XCP) from *lncRNA1456* modulates PHF8 histone demethylase activity to regulate the epigenome, gene expression, and cellular pathways in breast cancer

Shrikanth S. Gadad[1,2,3,6,7,9], Cristel V. Camacho[1,2,9], Xuan Gong[1,8], Micah Thornton[1,4], Venkat S. Malladi[1,4], Anusha Nagari[1,4], Aishwarya Sundaresan[1,4], Tulip Nandu[1,4], Sneh Koul[1,4], Yan Peng[5] and W. Lee Kraus[1,2]✉

Recent studies have demonstrated that a subset of long "noncoding" RNAs (lncRNAs) produce functional polypeptides and proteins. In this study, we discovered a 132 amino acid protein in human breast cancer cells named XCP (X-linked Cancer-associated Polypeptide), which is encoded by *lncRNA1456* (a.k.a. *RHOXF1P3*), a transcript previously thought to be noncoding. *lncRNA1456* is a pancreas- and testis-specific RNA whose gene is located on chromosome X. We found that the expression of *lncRNA1456* and XCP is highly upregulated in the luminal A, luminal B, and HER2 molecular subtypes of breast cancer. XCP modulates both estrogen-dependent and estrogen-independent growth of breast cancer cells by regulating cancer pathways, as shown in cell and xenograft models. XCP shares some homology with homeodomain-containing proteins and interacts with the histone demethylase plant homeodomain finger protein 8 (PHF8), which is also encoded by an X-linked gene. Mechanistically, XCP is required for the binding of PHF8 to chromatin. Moreover, XCP stimulates the histone demethylase activity of PHF8 to regulate gene expression in breast cancer cells. These findings identify XCP as a coregulator of PHF8 in the chromatin-dependent regulation of gene expression and emphasize the need to interrogate the potential functional roles of open reading frames originating from noncoding RNAs.

Oncogene; https://doi.org/10.1038/s41388-026-03740-w

INTRODUCTION

Recent advancements in genomics and deep sequencing technologies have uncovered that a significant fraction of the noncoding mammalian genome is transcribed, leading to the discovery of many long noncoding RNAs (lncRNAs) [1]. While lncRNAs are typically classified as noncoding based on computational approaches (e.g., codon-substitution frequency; CSF) [2], this may not be ideal for some highly evolved RNAs that serve complex biological functions in humans and non-human primates. Additionally, many peptides have gone unnoticed due to technical limitations and assumptions made during the annotation of genome databases [3]. As a result, research has recently focused on identifying functional open reading frames (ORFs) that might be translated into functional novel peptides and proteins derived from transcripts classified as noncoding, particularly lncRNAs.

The remarkable number of expressed lncRNAs, some of which conceivably also have protein-coding potential, have led to a significant gap in our knowledge and understanding of the biological roles of this distinct group of peptides. One of the most interesting features of lncRNAs is their tightly regulated developmental and tissue-specific expression, which is often dysregulated in diseased states, such as cancer [4]. The discovery of lncRNA-derived peptides has the potential to expand our understanding of the function of lncRNAs and their roles in cancer, ultimately uncovering exciting new targets for therapeutic intervention.

Emerging evidence suggests that peptides derived from lncRNAs play a crucial role in physiological processes, as well as in the manifestation of various diseases, including cancer [5–17]. Studies of lncRNAs, such as *long noncoding RNA 152* (*lncRNA152*), which we previously annotated and characterized [18], suggest

[1]Laboratory of Signaling and Gene Regulation, Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA. [2]Division of Basic Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [3]Center of Emphasis in Cancer, Department of Molecular and Translational Medicine, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, TX, USA. [4]Computational Core Facility, Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA. [5]Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [6]Present address: Division of Cancer Immunology and Microbiology, Medicine and Oncology Integrated Service Unit, The University of Texas Rio Grande Valley School of Medicine, McAllen, TX, USA. [7]Present address: South Texas Center of Excellence in Cancer Research, The University of Texas Rio Grande Valley School of Medicine, McAllen, TX, USA. [8]Present address: Department of Bone Marrow Transplantation and Cellular Therapy, St. Jude Children's Research Hospital, Memphis, TN, USA. [9]These authors contributed equally: Shrikanth S. Gadad, Cristel V. Camacho. ✉email: LEE.KRAUS@utsouthwestern.edu

Received: 9 May 2025 Revised: 14 February 2026 Accepted: 12 March 2026
Published online: 01 April 2026

that they are critical regulators of cancer biology [4]. However, in the studies described herein, we found the presence of potential ORFs in many annotated lncRNAs identified in breast cancer. We focused our analyses on XCP (X-linked Cancer-associated Polypeptide), which is encoded by a pancreas- and testis-specific RNA, *lncRNA1456* (a.k.a. *RHOXF1P3*) [19]. XCP interacts with plant homeodomain finger protein 8 (PHF8) to modulate its demethylase activity and regulate gene expression outcomes in breast cancer. Overall, this study underscores the importance of the novel polypeptide XCP in epigenetic regulation of gene expression and tumor biology.

MATERIALS AND METHODS

Cell culture and treatments

MCF-7 (ATCC; RRD:CVCL_0031) cells were kindly provided by Benita S. Katzenellenbogen (University of Illinois, Urbana-Champaign, Champaign, IL) and MDA-MB-231 cells were purchased from the American Type Culture Collection (ATCC; RRID:CVCL_0062). MCF-7 cells were maintained in Minimum Essential Medium Eagle (Sigma-Aldrich, M1018) supplemented with 5% calf serum (Sigma-Aldrich). MDA-MB-231 cells were maintained in RPMI-1640 Medium (Sigma-Aldrich, R8758) supplemented with 10% fetal bovine serum (Sigma-Aldrich). 293 T (ATCC; RRID:CVCL_0063) cells were maintained in high glucose Dulbecco's Modified Eagle's Medium (Sigma-Aldrich, D5796) supplemented with 10% fetal bovine serum. For doxycycline (Dox; Sigma-Aldrich, D9891) induction, cells were treated with 250 ng/mL of Dox for 24–48 h. All cell lines were authenticated for cell type identity using the GenePrint 24 system (Promega, B1870), and confirmed as *Mycoplasma*-free every 6 months using the Universal Mycoplasma Detection Kit (ATCC, 30-1012 K).

Antibodies

The custom rabbit polyclonal antiserum against XCP was generated by Pocono Rabbit Farm and Laboratory by using purified recombinant XCP. Other antibodies used were as follows: PHF8 (Bethyl, A301-772A; RRID:AB_1211497), GFP (Abcam, ab13970; RRID:AB_300798), FLAG M2 (Sigma, F3165; RRID:AB_259529), DYKDDDDK Tag (Invitrogen, PA1-984B RRID:AB_347227), β-tubulin (Abcam, ab6046); RRID:AB_2210370, His Tag (H-3) (Santa Cruz, sc-8036; RRID:AB_627727), Histone H3K9me2 (Cell Signaling, 9753; RRID:AB_659848), Histone H3 (Abcam, ab1791; RRID:AB_302613), goat anti-rabbit HRP- conjugated IgG (Pierce, 31460; RRID:AB_228341), goat anti-mouse HRP-conjugated IgG (Pierce, 31430; RRID:AB_228307), goat anti-chicken HRP-conjugated IgG (Abcam, ab6877; RRID:AB_955465), and Alexa Fluor 488 goat anti-rabbit IgG (Thermo Fisher Scientific, R37116; RRID:AB_2556544).

RNA isolation and polyA+RNA-seq

Total RNA was isolated from MCF-7 cells using the RNeasy kit (Qiagen, 74136) according to the manufacturer's instructions. The RNA collected was processed for whole genome polyadenylated RNA sequencing (polyA + RNA-seq).

The total RNA samples were subjected to enrichment of polyA+ RNA as described previously [20]. Briefly, poly(A) + RNA was enriched using Dynabeads oligo(dT)25 (Invitrogen), heat fragmented, and reverse transcribed using random hexamers in the presence of dNTPs. Second strand cDNA synthesis was performed with dNTPs but replacing dTTP with dUTP. After end-repair, dA tailing, ligation to adaptors containing barcode sequences, and size selection using AMPure XP beads (Beckman Coulter, A63881), the synthesized second-strand was digested using uracil DNA glycosylase (Enzymatics, Y9180L). A final PCR reaction was performed using KAPA HiFi HotStart Ready Mix (KAPA Biosystems, KK2612). After the library quality control assessment using a Bioanalyzer (Agilent), the samples were subjected to 50 bp single-end sequencing using an Illumina NextSeq Sequencing System. At least two biological replicates were sequenced for each cell line with a minimum of ~65 M raw reads per cell line.

Rapid amplification of cDNA ends (RACE)

Rapid amplification of 5′ and 3′ ends was performed using 5′/3′ RACE Kit, 2nd Generation (Roche, 03-353-621-001), according to the manufacturer's instructions.

Computational pipeline for identification of potential lncRNA-derived peptides

Database of ORFs. We started with a comprehensive set of annotated lncRNAs from LNCipedia (v4.0) [21] and converted to bed format using UCSC utilities gtfToGenePred and genePredToBed (http://hgdownload.soe.ucsc.edu/admin/exe/). Extraction of sequences for each annotation was achieved by using getfasta in BEDTools (v2.17.0) [22]. All possible ORFs starting with Methionine were determined using the getorf tool in the European Molecular Biology Open Software Suite [23].

Length of ORFs. Histogram representations were used to assess the median length of all ORFs and the longest ORF for each transcript.

Protein digest. Maximum theoretical coverage of predicted ORFs and known proteins from UniProt (RRID:SCR_002380) [24] digested by trypsin and chymotrypsin was determined using the MS Proteomics tools library (https://github.com/msproteomicstools/msproteomicstools). Histograms representations were used to assess the median coverage of digestion by either enzyme.

Peptide database. We created a comprehensive protein database by combining our database of predicted ORFs with known proteins from UniProtKB/Swiss-Prot [25]. Spectra were analyzed using Proteome Discoverer 2.0 (Thermo Fisher Scientific, Waltham, MA; RRID:SCR_014477) against the combined protein database.

Mass spectrometry and analysis

Preparation of cell extracts for mass spectrometric analysis. Cells were collected, washed with ice-cold PBS and resuspended in Whole Cell Lysis Buffer [50 mM Tris-HCl pH 7.5, 0.5 M NaCl, 1 mM EDTA, 1% NP-40, 10% Glycerol, and 1x complete protease inhibitor cocktail (Roche, 11697498001)] and incubated for 30 min on ice with gentle mixing to lyse the cells and extract the proteins. All lysates were clarified by centrifugation in a microcentrifuge for 5 min at 4 °C at full speed. The extracts were fractionated on 4–12% gradient polyacrylamide-SDS gels, followed by Coomassie staining. Bands between 5 and 20 kDa were excised from the gel and subjected to LC-MS/MS.

Analysis. The Mass-spectrometry spectra were analyzed using Proteome Discoverer 2.0 (Thermo Fisher Scientific, Waltham, MA) against the combined peptide database described above. Software, scripts and other information about the analyses can be obtained by contacting the corresponding author (W.L.K.).

Molecular cloning and purification of recombinant XCP

XCP coding sequence was subcloned between the NcoI and XhoI sites of the pET19b (Novagen, 69677) vector by PCR-based sub cloning from MCF-7 cDNA library. His_6 tagged XCP was purified from E. coli cells using Ni-NTA column as described elsewhere [26].

Preparation of cell extracts and Western blotting

Preparation of whole cell lysates. Cells were collected, washed with ice-cold PBS and resuspended in Whole Cell Lysis Buffer [50 mM Tris-HCl pH 7.5, 0.5 M NaCl, 1 mM EDTA, 1% NP-40, 10% Glycerol, and 1x complete protease inhibitor cocktail (Roche, 11697498001)] and incubated for 30 min on ice with gentle mixing to lyse the cells and extract the proteins. All lysates were clarified by centrifugation in a microcentrifuge for 5 min at 4 °C at 15,000 RPM.

Determination of protein concentrations and Western blotting. Protein concentrations were determined using a BCA protein assay (Pierce, 23225). The cell extracts were aliquoted, flash-frozen in liquid N_2, and stored at −80 °C. Aliquots of the cell extracts were run on polyacrylamide-SDS gels and transferred to nitrocellulose membranes. After blocking with 5% nonfat milk in TBST, the membranes were incubated with the primary antibodies described above in 3% nonfat milk prepared in TBST, followed by anti-rabbit, anti- mouse, or anti- chicken HRP-conjugated IgG. Western blot signals were detected using an ECL detection reagent (Thermo Fisher Scientific, 34077, 34095).

Functional analyses of lncRNAs in MCF-7 cells

siRNA-mediated knockdowns (lncRNA1456 and PHF8) in MCF-7 cells. Transient siRNA-mediated knockdown of *PHF8* or *lncRNA1456* was performed

by transfection of commercially available siRNA oligos targeting human PHF8 (siRNA ID # s61133 and s23107; Invitrogen, 4392420), and custom-designed siRNAs for *lncRNA1456* (designed using SciTools RNAi design software from Integrated DNA Technologies). Commercially available control siRNAs (Sigma-Aldrich, MISSION siRNA universal negative control) were also used. Cells were plated at a density of 2×10^5 cells per well in six-well dishes. Transfections were done at a final concentration of 10 nM using Lipofectamine RNAiMAX reagent (Invitrogen, 13778150) according to the manufacturer's instructions. Forty-eight hours post-transfection, cells were collected for RT-qPCR or RNA-seq.

siRNA sequences. We used the following nucleic acid oligonucleotides for targeted knockdowns:

si1456-1 5′-GUACAAGGUUAAGUGUAAAUA[dT][dT]-3′
si1456-1_as 5′-UAUUUACACUUAACCUUGUAC[dT][dT]-3′
si1456-2 5′-ACCUGCCUCAGUCCUUGAAUA[dT][dT]-3′
si1456-2_as 5′-UAUUCAAGGACUGAGGCAGGU[dT][dT]-3′
si-Luc 5′-GAUUUGUAUUCAGCCCAUA[dT][dT]-3′
si-GFP 5′-ACAACAGCCACAACGUCUA[dT][dT]-3′

Analysis of lncRNA and mRNA expression by RT-qPCR. Total RNA was isolated from MCF-7 or MDA-MB-231 cells using the RNeasy kit (Qiagen, 74136) according to the manufacturer's instructions. For xenograft tumors, tissue was homogenized in RTL Plus Buffer (Qiagen 74136). RNA was subjected to reverse transcription using Oligo-dT and MMLV reverse transcriptase (Promega, M1705). The resulting cDNA pool was treated with three units of RNase H (Enzymatics, Y9220L) for 30 min at 37 °C, and then analyzed by qPCR using a Roche LightCycler 480 system (initial 95 °C for 5 min, 45 cycles of amplification at 95 °C for 10 s, 60 °C for 10 s, 72 °C for 1 s) with SYBR Green detection and gene-specific primers (see list below). Target gene expression was normalized to the expression of *RPL19* mRNA.

qPCR primer sequences. We used the following nucleic acid oligonucleotides for qPCR:

XCP-CDS-F ′-GAAGGCGACAATGCGAAGG-3′
XCP-CDS-R 5′-GCTCGGGTTTGGACCGTATT-3′
XCP-FLAG-R 5′-CTTGTCATCGTCATCCTTATAATC-3′
XCP-5nonCDS-F 5′-CAGGGCAGGAGCCCACT-3′
XCP-5nonCDS-R 5′-CTGCTCTGAGTTCCGGCTC-3′
XCP-Ex3-F 5′-AGCCACTCTCTGAAACCTGC-3′
XCP-Ex3-R 5′-AGGCAACAAAAACAGGCCAA-3′
hPHF8-F1 5′-GACCTGACTATGCTGCCCTC-3′
hPHF8-R1 5′-ACGATGAGGACTGCAGGTTG-3′
GFP-F 5′-ACGACGGCAACTACAAGACC-3′
GFP-R 5′-TTGTACTCCAGCTTGTGCCC-3′
RPL19-F 5′-ACATCCACAAGCTGAAGGCA-3′
RPL19-R 5′-TGCGTGCTTCCTTGGTCTTA-3′

Analysis of TCGA and GTEx data

Data from TCGA and GTEx was downloaded from the recount2 database [27] that was aligned to the human reference genome (GRCh38/hg38) and gene annotations for reference chromosomes from GENCODE (v.25; RRID:SCR_014966).

Box plots. Box plot representations were used to quantitatively assess the expression of genes in normal tissue. Additionally, box plot representations of the mean and standard error were used to assess the expression of genes in normal and cancer tissues. Differential expression of RPKM normalized data was tested by ANOVA, using the aov packaged in R.

Histology and immunostaining

A breast disease spectrum tissue microarray was purchased from TissueArray.com (BR2082b). Immunohistochemical staining was performed on a Dako Autostainer Link 48 system. Briefly, 5 µm paraffin sections were baked for 20 min at 60 °C, then deparaffinized and hydrated before the antigen retrieval step. Heat-induced antigen retrieval was performed at pH 6 (XCP) for 20 min in a Dako PT Link. The tissue was incubated with a peroxidase block and then an antibody incubation (1:100 XCP) for 20 min. The staining was visualized using the EnVision FLEX visualization system. The intensity of staining was scored on a scale of 0–3, where 3 is the highest intensity (expression).

Generation of ectopic cell lines

Molecular cloning of XCP and PHF8. *LncRNA1456* 5′ and 3′ ends were mapped by RACE as described above. The full-length coding sequence was cloned by PCR with a 3′-FLAG epitope tag flanked by NheI and XhoI sites using MCF-7 cDNA. The full length *lncRNA1456* (wild-type and ATG mutant sequences), and PHF8 cDNAs were synthesized by Invitrogen GeneArt Gene Synthesis and sequence verified. Purified plasmid contained NheI and XhoI flanked full length *lncRNA1456* (wild-type or ATG mutant sequences), or PHF8 with 3′-FLAG epitope-tag coding sequence.

Lentiviral expression vector. The cDNAs, or a GFP-FLAG cDNA control, were inserted into pInducer20 lentiviral doxycycline (Dox)- inducible expression vector by ligation using NheI and XhoI sites.

Doxycycline-inducible expression in cell lines. Lentiviruses were generated by transfecting each pInducer20 (Addgene, 44012; RRID:Addgene_44012) vector into 293 T cells, together with expression vectors for the VSV-G envelope protein (pCMV-VSV-G, Addgene plasmid no. 8454; RRID: Addgene_8454), the expression vector for GAG-Pol-Rev (psPAX2, Addgene plasmid no. 12260; RRID: Addgene_12260), and a vector to aid with translation initiation (pAdVAntage, Promega E1711) using Lipofectamine 3000 transfection reagent (Thermo Fisher Scientific, L3000015) according to the manufacturer's instructions. After 24 h, culture medium was replaced with fresh medium, and the cells were maintained for an additional 24 h. The virus-containing supernatants were collected, filtered through a 0.45 µm syringe filter, and concentrated by using Lenti-X Concentrator (Clontech, 631231). The filtered supernatants were used to infect MCF-7 or MDA-MB-231 cells supplemented with 1 µg/mL polybrene (Sigma-Aldrich, H9268) to increase transduction efficiency. The infected cells were placed under selection with 1 mg/mL Geneticin/G418 (Thermo Fisher Scientific, 11811031), and once stable, maintained with 400 µg/mL Geneticin/G418. For doxycycline (Dox; Sigma-Aldrich, D9891) induction, cells were treated with 250 ng/mL of Dox for 24–48 h. Expression of proteins was confirmed by Western blotting.

Xenograft assays

All animal experiments were performed in compliance with the Institutional Animal Care and Use Committee (IACUC) at the UT Southwestern Medical Center. Sample size was based on estimations by power analysis with a level of significance of 0.05 and a power of 0.9. Female NOD *scid* gamma (NSG) mice at 6–8 weeks of age were used, $n = 7$ for MCF-7, $n = 8$ for MDA-MB-231. Cell lines used were verified for cell type identity using the GenePrint 24 system (Promega, B1870). To establish breast cancer xenografts, MDA-MB-231 or MCF-7 cells (5×10^6 in 100 µl), engineered for doxycycline-inducible expression of FLAG-tagged XCP or GFP were injected subcutaneously into the flank of mice in a 1:1 ratio PBS and matrigel (Corning, 354234). For MCF-7 experiments, NSG mice were supplemented with 1.7 mg 17-β-estradiol 60-day release pellets (Innovative Research, SE-121) implanted subcutaneously at the base of the neck under anesthesia. Tumors that did not grow were excluded from experiment, no randomization. Ten days post-tumor cell injection, mice were placed on a doxycycline containing diet (625 mg/kg; Envigo, TD.01306). Mouse weight was monitored once a week and tumor growth measured over time, non-blinded, using electronic calipers approximately every 3-4 days. Tumor volumes were calculated using a modified ellipsoid formula: $\textit{Tumor volume} = ½\ (\textit{length} \times \textit{width}^2)$. Variance between animals within each group was expected to due to natural biological heterogeneity and was quantified and reported as mean ± SEM unless otherwise indicated. Animals were euthanized at 50 days post-injection. Fresh tumor samples were fixed for 24 h in 10% formalin.

Immunofluorescent staining and confocal microscopy

MCF-7 cells were seeded on four-well chambered cover slips (Thermo Fisher, 155411) 1 day prior to fixing. The following day, cells were washed three times with PBS, fixed in 3.7% paraformaldehyde for 15 min at room temperature, and washed with 0.1% PBS-Tween 20 (PBS-T). Cells were permeabilized for 5 min using 0.5% PBS-T, washed three times with 0.1% PBS-T, and incubated for 1 hour at room temperature in Blocking Solution (10% calf serum in PBS containing 0.5% gelatin) and then washed twice with 0.1% PBS-T. Fixed cells were incubated at room temperature with a polyclonal antibody against XCP in Incubation Solution (PBS containing 1% BSA) for 2 h at room temperature and washed four times with 0.1% PBS-T. Then the fixed cells were probed with Alexa Fluor 488 goat anti-rabbit IgG (Thermo Fisher Scientific, R37116; RRID:AB_2556544) in blocking solution

for 1 h, washed four times with 0.1% PBS-T, and incubated at room temperature with TO-PRO™-3 Iodide nuclear dye (Thermo Fisher Scientific, T3605) or 4,6-diamidino-2-phenylindole nuclear dye (Thermo Fisher Scientific, D1306) in PBS for 5 min, and then washed four times with 0.1% PBS-T again. Lastly, cells were treated with VectaShield (Vector Laboratories, H-1000) and imaged using a Zeiss LSM880 confocal microscope, purchased with a shared instrumentation grant from the NIH (1S10OD021684-01 to Katherine Luby-Phelps).

In vitro pull down assays

Ni-NTA pull down assay. One microgram of the His_6-tagged protein XCP was mixed with 2 micrograms of FLAG-PHF8 (Active Motif, 31435) in the interaction buffer (20 mM Tris-HCl [pH 7.9], 20% glycerol, 0.2 mM EDTA [pH 8.0], 0.1% Nonidet P-40, 2 mM phenylmethylsulfonyl fluoride, 150 mM KCl, 30 mM imidazole) along with the Ni-nitrilotriacetic acid (NTA) His Bind resin (Qiagen). The mixture was incubated for 3 h at 4 °C on a rotary shaker. After the beads were extensively washed in the interaction buffer, the proteins were extracted from the beads into the sodium dodecyl sulfate (SDS) sample buffer, separated on an 12% polyacrylamide-SDS gel and visualized by Western blotting with rabbit anti-XCP and mouse anti-FLAG (Sigma, F3165; RRID:AB_259529).

Streptavidin pull down assay. Two micrograms of the His_6-tagged protein XCP was mixed with 2 micrograms of FLAG-PHF8 (Active Motif, 31435), and 1 microgram of Histone H3K9me2 peptide-biotinylated (Active Motif, 81046) in the binding buffer (20 mM Tris-HCl [pH 7.9], 20% glycerol, 0.2 mM EDTA [pH 8.0], 0.1% Nonidet P-40, 2 mM phenylmethylsulfonyl fluoride, 150 mM KCl). The mixture was incubated for 2 h at 4 °C on a rotary shaker, followed by incubation with Dynabeads M-280 Streptavidin (Invitrogen, 11205D) for 1 h at 4 °C on a rotary shaker. After magnetic beads were washed in binding buffer, the proteins were extracted from the beads into the sodium dodecyl sulfate (SDS) sample buffer, separated on an 4–12% gradient polyacrylamide-SDS gel and visualized by Western blotting with rabbit anti-His and mouse anti-FLAG.

Analysis of domain family

We downloaded the list of Homeobox family of proteins from the HGNC database [28]. All 314 Homeobox family proteins and the identified ORF were aligned using Clustal Omega (RRID:SCR_001591) [29].

Analysis of *lncRNA1456*- and PHF8- regulated genes by RNA-seq

PolyA+ RNA-seq libraries were prepared from control and *lncRNA1456* knockdown MCF-7 cells as described above using the dUTP method [20]. Two biological replicates were generated for each sample. The RNA-seq raw reads were mapped to the hg19 human reference genome by TopHat (RRID:SR_013035) [30], using RefSeq (RRID:SCR_003496) gene annotations as the reference for alignment. To determine expressed transcripts, we used custom R scripts to calculate the counts and RPKM using the Bioconductor (RRID:SCR_006442) packages GenomicRanges (RRID:SCR_000025) and edgeR (RRID:SCR_012802) (Lawerence et al., 2013, Robinson et al., 2010). Differentially regulated RefSeq mRNAs were called by Cuffdiff, using a 5% FDR, comparing the control samples to the lncRNA knockdown samples. We derived a high-confidence regulated mRNA set by filtering the Cuffdiff-called regulated mRNA lists with a fold cutoff of either $2^{(0.8)}$ or $2^{(-0.8)}$ for each siRNA-treated condition relative to the control. The resulting mRNAs with their corresponding fold changes were represented in heatmaps using Java Treeview.

Chromatin immunoprecipitation (ChIP)-sequencing and analysis

Chromatin immunoprecipitation. For ChIP assays, cells were cross-linked with 1% formaldehyde in PBS for 10 min at 37 °C, quenched by the addition of 125 mM glycine, and incubated 5 min at 4 °C. The cross-linked cells were collected in ice-cold PBS and pelleted by centrifugation. The cells were then resuspended by gentle mixing by pipetting in ice-cold Farnham Lysis Buffer (5 mM PIPES pH 8.0, 85 mM KCl, 0.5% NP-40) with freshly added 1 mM DTT and 1x protease inhibitor cocktail. The supernatant was removed, and the crude nuclear pellet was collected by centrifugation and resuspended in SDS Lysis Buffer (50 mM Tris-HCl pH 7.9, 1% SDS, 10 mM EDTA) with freshly added 1 mM DTT and 1x protease inhibitor cocktail. After a 10 min incubation on ice, the lysate was sheared by sonication using a Bioruptor (Diagenode) to generate chromatin fragments of approximately 250 bp in length. The sheared chromatin was clarified by centrifugation and the diluted ten-fold in ChIP Dilution Buffer (20 mM Tris-HCl pH 7.9, 0.5% Triton X-100, 2 mM EDTA, 150 mM NaCl) with freshly added 1 mM DTT and 1x protease inhibitor cocktail. The lysate was pre-cleared with equilibrated protein A-agarose beads and subjected to immunoprecipitation reactions with an antibody against PHF8 (Bethyl, A301-772A) at 4 °C overnight.

The immunoprecipitates were collected by incubation with BSA-blocked protein A-agarose beads for 2 h at 4 °C with gentle mixing. After incubation, the beads were washed on ice once each with (1) Low Salt Wash Buffer (20 mM Tris-HCl pH 7.9, 2 mM EDTA, 125 mM NaCl, 0.05% SDS, 1% Triton X-100, 1x protease inhibitor cocktail), (2) High Salt Wash Buffer (20 mM Tris-HCl pH 7.9, 2 mM EDTA, 500 mM NaCl, 0.05% SDS, 1% Triton X-100, 1x protease inhibitor cocktail), (3) LiCl Wash Buffer (10 mM Tris-HCl pH 7.9, 1 mM EDTA, 250 mM LiCl, 1% NP-40, 1% sodium deoxycholate, 1x protease inhibitor cocktail), and (4) 1x Tris-EDTA (TE). The beads were then subjected to a final wash with 1x TE at room temperature. They were then collected by centrifugation, resuspended in ChIP Elution Buffer (100 mM $NaHCO_3$, 1% SDS), and incubated on end-over-end rotator for 15 minutes at room temperature to elute the ChIPed DNA. The ChIPed DNA was de-crosslinked by adding 100 mM NaCl with incubation at 65 °C overnight. The eluted material was cleared of protein and RNA by adding RNase H and proteinase K, and incubating for 2 h at 55 °C. The ChIPed DNA was then extracted with phenol:chloroform:isoamyl alcohol (25:24:1), collected by ethanol precipitation, and dissolved in water. ChIPed DNA was used for library preparation for sequencing.

Library preparation. ChIP libraries were prepared using a modified KAPA LTP Library Preparation kit (KAPA Biosystems, KK8232) for Illumina Platforms. Ten ng of sheared DNA was used to repair the ends of the damaged fragments using a proprietary master mix. The resulted blunted fragments were 3′ A-tailed using a proprietary mixture of enzymes to allow ligation to the specific Illumina adaptors. Each of the steps (i.e., end repair, 3′A tailing, and adaptor ligation) was followed by AMPure XP bead clean up (Beckman Coulter, A63881). After adapter ligation, DNA enrichment was performed using Kapa HiFi Hot Start Ready PCR mix, and a cocktail of primers (1 cycle at 98 °C for 45 s; five cycles at 98 °C for 20 s, 63 °C for 30 s, and 72 °C for 30 s; and 1 cycle at 72 °C for 1 min), and purified with AMPure XP beads. DNA templates were size selected (~200–300 bp) by running on a 2% agarose gel, followed by PCR enrichment (1 cycle at 98 °C for 45 s; 11 cycles at 98 °C for 20 s, 63 °C for 30 s, and 72 °C for 30 s; and 1 cycle at 72 °C for 1 min) and final purification. The quality of the final libraries was assessed using a 2200 TapeStation (Agilent Technologies). The libraries were quantified using Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific, Q32854) and samples pooled at final concentration of 2 nM (GSE288868 for RNA-seq, and GSE287423 for ChIP-seq).

Library sequencing. The libraries were sequenced using a NextSeq sequencer (Illumina; Single-end reads, 75 bp for all samples). At least two biological replicates were sequenced for each cell line for a minimum of roughly 100 million raw reads per cell line.

Quality control. Quality control for the RNA-seq data was performed using the FastQC tool (RRID:SCR_014583) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Read alignment and peak calling. The raw reads were aligned to the human reference genome (GRCh37/hg19) using default parameters in Bowtie (ver. 1.0.0) (RRID:SCR_005476) [31]. The aligned reads were subsequently filtered for quality and uniquely mappable reads using Samtools (ver. 0.1.19) (RRID:SCR_002105) [32] and Picard (ver. 1.127; http://broadinstitute.github.io/picard/) (RRID:SCR_006525). Library complexity was measured using BEDTools (v2.17.0) (RRID:SCR_006646) [22] and met minimum ENCODE (RRID:SCR_006793) data quality standards [33]. Relaxed peaks were called using MACS (v2.1.0 [34] with a p-value $= 1 \times 10^{-2}$ for each replicate, pooled replicates' reads, and pseudoreplicates. Called peaks from the pooled replicate that were observed in both replicates or in both pseudoreplicates were used for subsequent analyses.

Metagene and heatmap analysis. The average read densities of reads were calculated for a 1 or 20 Kb window surrounding peak center using the plotProfile and plotHeatmap functions in deeptools (ver. 2.5.0.1) [35].

Respective boxplots were generated using the boxplot function in R (ver. 4.4.3).

Demethylation assay and dot blots

Purified His_6-XCP, recombinant Histone H3K9me2 (Active Motif, 31280), and recombinant FLAG-PHF8 protein (Active Motif, 31435) were used for in vitro assays. Ten microliters of each demethylation reaction (10 µL) containing 1 µg H3K9me2 peptide and/or 1.2 µg FLAG-PHF8, or 1-2 µg His_6-XCP. PHF8 and/or XCP were incubated in demethylation buffer (50 mM HEPES, 250 mM NaCl, 1 mM $MgCl_2$, 1 mM α-ketoglutarate, 0.04 mM $FeSO_4{\cdot}7H_2O$, and 0.04 mM ascorbic acid) for 15 min at 37 °C. Subsequently, H3K9me2 peptide was added and the mixture was further incubated for 1 h at 37 °C. Each reaction was then spotted onto a nitrocellulose membrane. The membranes were air dried, blocked with 5% nonfat milk in TBST, incubated with the Histone H3K9me2 (Cell Signaling, 9753) or Histone H3 (Abcam, ab1791) primary antibodies described above in 3% non-fat milk prepared in TBS-T, and then incubated with goat anti-rabbit HRP-conjugated IgG (Pierce, 31460). Western blot signals were detected using an ECL detection reagent (Thermo Fisher Scientific, 34077, 34095).

Histone post-translational modification mass spectrometry and analyses

Isolation of histones by acid extraction. Histone analysis was carried out as previously described [36]. After isolation of nuclei, acid extraction was performed to release histone proteins from nuclei, with subsequent purification as described previously [37]. The acid extracted histones were analyzed on 15% polyacrylamide-SDS gels, using ~10 µg of histone proteins, and histone modifications were detected by Western blotting using specific antibodies, as described above.

Propionylation of histones. Chemical derivatization (propionylation) was carried out for mass spectrometry (MS) analysis of histone modifications as previously described [38]. Briefly, the acid extracted histones were dissolved in ammonium bicarbonate, and the pH was adjusted to 8. Propionylation reagent, which was made by combining propionic anhydride and acetonitrile (ACN) in a 1:3 ratio (v/v), was added, followed by adjustment of the pH using NH_4OH. Centrifugation and vacuum drying produced propionylated histones, ideal for proteolytic digestion and HPLC-MS analysis.

Mass spectrometry. The proteins were digested with trypsin and processed for MS as described previously [36, 39]. The resulting peptides were injected onto an Orbitrap Fusion Lumos mass spectrometer (Thermo Electron) coupled to an Ultimate 3000 RSLC-Nano liquid chromatography system (Dionex). The samples were eluted with a gradient from 1 to 28% Buffer B over 90 min. Buffer A contained 2% (v/v) ACN and 0.1% formic acid in water, and Buffer B contained 80% (v/v) ACN, 10% (v/v) trifluoroethanol, and 0.1% formic acid in water. MS scans were acquired at a 120,000 resolution in the Orbitrap and up to 10 MS/MS spectra were obtained in the ion trap for each full spectrum acquired using higher-energy collisional dissociation (HCD) for ions with charges 2-7.

Analysis of mass spectrometry data. Raw MS data files were analyzed using Proteome Discoverer v2.4 SP1 (Thermo Fisher Scientific, Waltham, MA) with peptide identification performed using Sequest HT searching against the *Homo sapiens* database from UniProt. To mine the data for modifications of interest, we first filtered the records in the MS output file to contain only those which were indicated to be contained within either the Histone H3 or H4 complex, of these we then determined which entries covered the sites of interest (H3K9). After determining a listing of all records in the MS data that covered these sites, we calculated the relative abundances of PTM p at that site (k) for each replicate (t) of each condition (c) via:

$$a_{pkc(t)} = \frac{\sum_{i=1}^{N_k} a_{c(t)k(i)} \cdot m_i}{\sum_{i=1}^{N_k} a_{c(t)k(i)}}$$

$$m_i = \begin{cases} 1, & \text{observation i contains PTM p} \\ 0, & \text{observation i doesn't contain PTM p} \end{cases}$$

Where $a_{pkc(t)}$ is the relative abundance of PTM p at site k, of replicate t of condition c, $a_{c(t)k(i)}$ is the ith reported absolute abundance of replicate t of condition c at site k, N_k is the total number of records covering site k, and m_i is an indicator of whether or not a modification p is present in observation i of site k. Software, scripts and other information about the analyses can be obtained by contacting the corresponding author (W.L.K.).

RESULTS

Identification and validation of XCP, a polypeptide encoded by *lncRNA1456*

In this study, we developed a pipeline integrating RNA-sequencing (RNA-seq) and MS data to identify putative functional lncRNA-encoded peptides using a universe of lncRNAs previously discovered and annotated in MCF-7 breast cancer cells [19] (Fig. S1A). Briefly, we curated a peptide database by combining all possible ORFs, starting with methionine from the list of lncRNAs with known proteins from UniProtKB/Swiss-Prot. Further, we performed MS on lysates from MCF-7 cells with peptides under 20 kDa, and spectra were analyzed using Proteome Discoverer 2.0 (Thermo Fisher Scientific, Waltham, MA) against the combined peptide database (Fig. S1A). Using this pipeline, we discovered and functionally characterized a new polypeptide encoded by *lncRNA1456*, which we have named XCP (X-linked Cancer-associated Polypeptide) (Fig. S1B, C). Our previous comprehensive annotation of *lncRNA1456* using RNA-seq coupled with bioinformatics and RACE, identified three isoforms. The longest isoform contains four exons, while the other two isoforms contain three exons each. XCP is translated from an ORF spanning exons 1 and 2 (Fig. S1C). This locus was recently annotated as *RHOXF1P3* coding peptide in GENCODE, which appears to have been identified through computational annotation and proteogenomic approach using PRIDE database and CPTAC Data Portal [40]. To our knowledge, there are no published experimental studies functionally characterizing this locus. Therefore, we have kept the XCP nomenclature, which reflects both the historical precedence of our discovery and the biological function we have uncovered in this study.

RNA-seq analysis revealed that *lncRNA1456* is highly transcribed in MCF-7 cells, which is classified as a luminal breast cancer cell line (Fig. 1A). Its transcription is independent of estrogen treatment, and the transcript can be detected in both cytoplasmic and nuclear fractions (Fig. 1A). Analysis of its evolutionary conservation across genomes using the ECR browser [41] showed that *lncRNA1456* is conserved in humans and primates only, suggesting this is an evolutionarily young polypeptide (Fig. 1B and Fig. S1B). *LncRNA1456* contains an ORF that is translated into a 132-amino acid peptide with a predicted molecular weight of approximately 14 kDa (Fig. 1C and Fig. S1B). This peptide was detected in MCF-7 whole-cell extracts as a single band at approximately 18 kDa using an antibody raised against recombinant XCP (Fig. 1D). siRNA-mediated knockdown of *lncRNA1456* reduced the levels of detectable XCP polypeptide, confirming antibody specificity (Fig. 1E and Fig. S1D).

XCP is highly expressed in luminal subtype breast cancer, promoting tumor cell growth

To rule out the possibility that the XCP peptide is an in vitro artifact, we determined *lncRNA1456* RNA levels across normal tissues (GTEx). We found that *lncRNA1456* is transcribed exclusively in the testes and pancreas, which was further confirmed by immunohistochemistry using a collection of normal tissues (Fig. 2A). In agreement with this, *lncRNA1456* is highly transcribed in primary tumors (T; TCGA) from breast cancer patients but not in normal solid tissue (NT; TCGA) or normal breast tissue (N; GTEx) (Fig. 2B; *left panel*).

Analysis of XCP peptide expression using a commercial breast cancer tissue microarray indicates that XCP is expressed at high levels only in malignant and metastatic (MM) breast cancer tissues but not in non-malignant (NM) tissue (Fig. 2C, D; Fig. S2A–D);

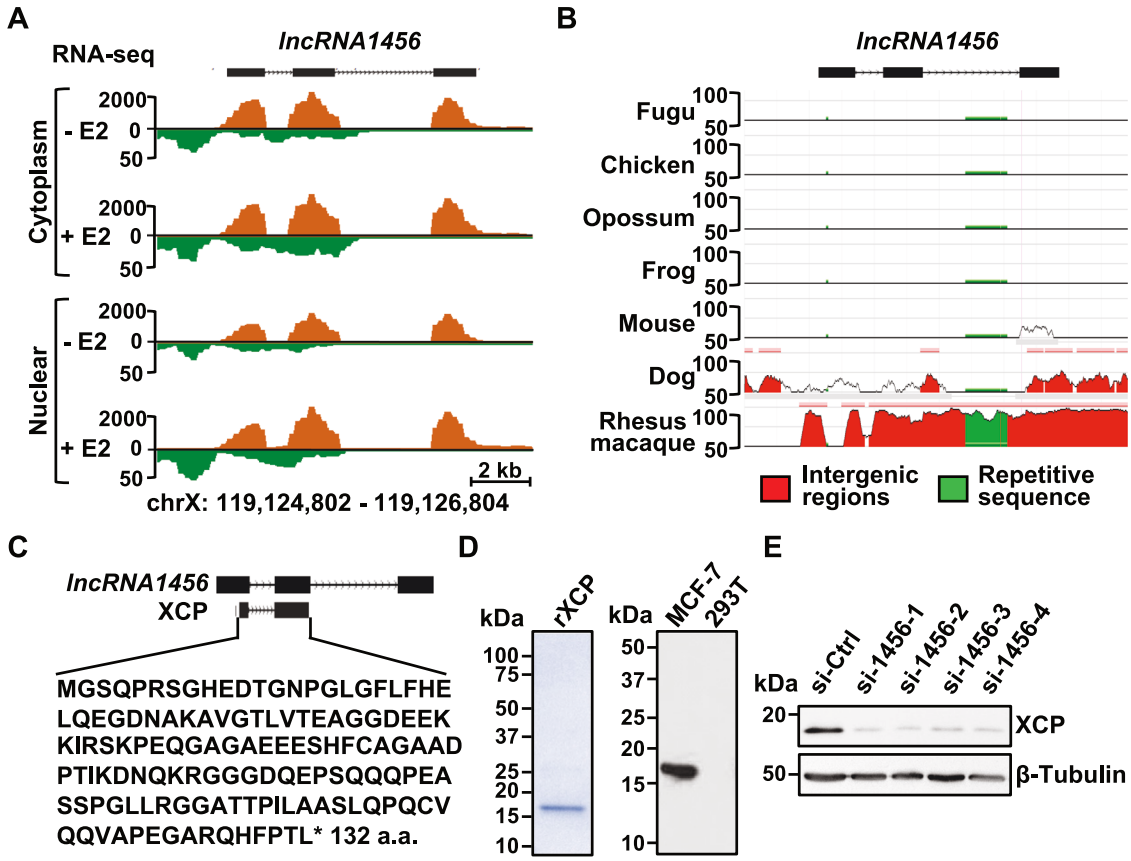

Fig. 1 Characterization of *lncRNA1456* and its encoded peptide, XCP. A Genome browser view for the *lncRNA1456* locus from fractionated RNA-seq data (cytoplasm vs nuclear) in the presence or absence of estrogen (E2) [19]. **B** Evolutionary Conserved Regions (ECR) browser tracks showing conservation of *lncRNA1456* locus across different species. **C** Diagram showing the putative *lncRNA1456*-encoded peptide (XCP) from exons 1 and 2 and amino acid sequence. **D** Coomassie gel showing recombinant XCP peptide (*left*) and its endogenous detection in MCF-7 cells by Western blot using a derived XCP antibody (*right*). **E** Western blot showing reduced levels of endogenous XCP peptide upon siRNA-mediated knockdown of *lncRNA1456* RNA in MCF-7 cells. β-tubulin was used a loading control. [See also Fig. S1].

within malignant and metastatic tissues, XCP was found to be significantly more highly expressed in estrogen receptor-positive (ER +) samples (Fig. 2E). Interestingly, and in agreement with this observation, although the overall frequency of XCP-positive samples was low, *lncRNA1456* was found to be upregulated in HER2, luminal A, and luminal B molecular subtypes of breast cancer (TCGA; Fig. 2B; *right panel*), suggesting a context-specific biological role. To explore this further, we ectopically expressed doxycycline (Dox)-inducible XCP-FLAG in MCF-7 breast cancer cell line, where XCP is naturally expressed and MDA-MB-231, where it is not expressed, representing a luminal and basal subtype of breast cancer, respectively, and examined cell growth in a xenograft model. Ectopic expression of XCP-FLAG in MCF-7 (luminal) cells significantly promoted tumor cell growth (Fig. 3A, C; Fig. S3A). In contrast, an inhibitory effect was observed when XCP-FLAG was ectopically expressed in MDA-MB-231 (basal) cells (Fig. 3B, D; Fig. S3B), indicating that XCP is playing a dual (oncogenic and tumor suppressive) role in a context-dependent manner.

XCP modulates tumor growth and regulates cellular pathways in a context-dependent manner

To better understand the biological consequences of XCP expression, we performed RNA-seq on MCF-7 and MDA-MB-231-derived xenograft tumors. Ectopic expression of XCP-FLAG resulted in differential gene expression in each cell line (Fig. 4A, B). Gene Ontology (GO) analysis showed enriched biological pathways that are regulated in the context of MCF-7 (luminal) or MDA-MB-231 (basal) cells. Interestingly, and consistent with the opposing effects observed in tumor cell growth, the pathways that were negatively enriched in MCF-7 (such as EMT, myogenesis, angiogenesis) were positively enriched in MDA-MB-231 cells. In contrast, pathways that were positively enriched in MCF-7 cells (such as G2/M checkpoint, E2F targets, and Myc targets) were negatively regulated in MDA-MB-231 cells (Fig. 4A, B). We also examined the overlap between XCP-regulated genes in MCF-7 cells and MDA-MB-231 cells and identified only 76 genes common to both datasets (Fig. S3C). Functional annotation suggests that these few overlapping genes are associated with processes such as the estrous cycle, response to steroid hormone, and signal transduction (Fig. S3D). Interestingly, in agreement with the expression of *lncRNA1456* and XCP observed in testes (Fig. 2A), enrichment of the term 'spermatogenesis' (Fig. 4A, B) suggests a biological role for XCP in the normal physiology of the testes.

We determined the expression of the XCP-induced gene set from MCF-7 xenograft tumors in breast tumor samples categorized by molecular subtype (PAM50) (Fig. 4C). Expression of the gene set was significantly upregulated in the non-basal subtypes of breast cancer (HER2 + , Luminal A, Luminal B, etc.), also in tumors positive for ER expression, corroborating the pro-tumorigenic effect of XCP in MCF-7 xenograft tumor growth (Fig. 3A). Intriguingly, the expression of XCP-induced genes from MDA-MB-231 xenograft tumors was also elevated in the luminal subtype of patient breast tumors including ER+ tumors (Fig. 4D), suggesting a role for XCP in driving the gene expression profiles of luminal subtype tumors.

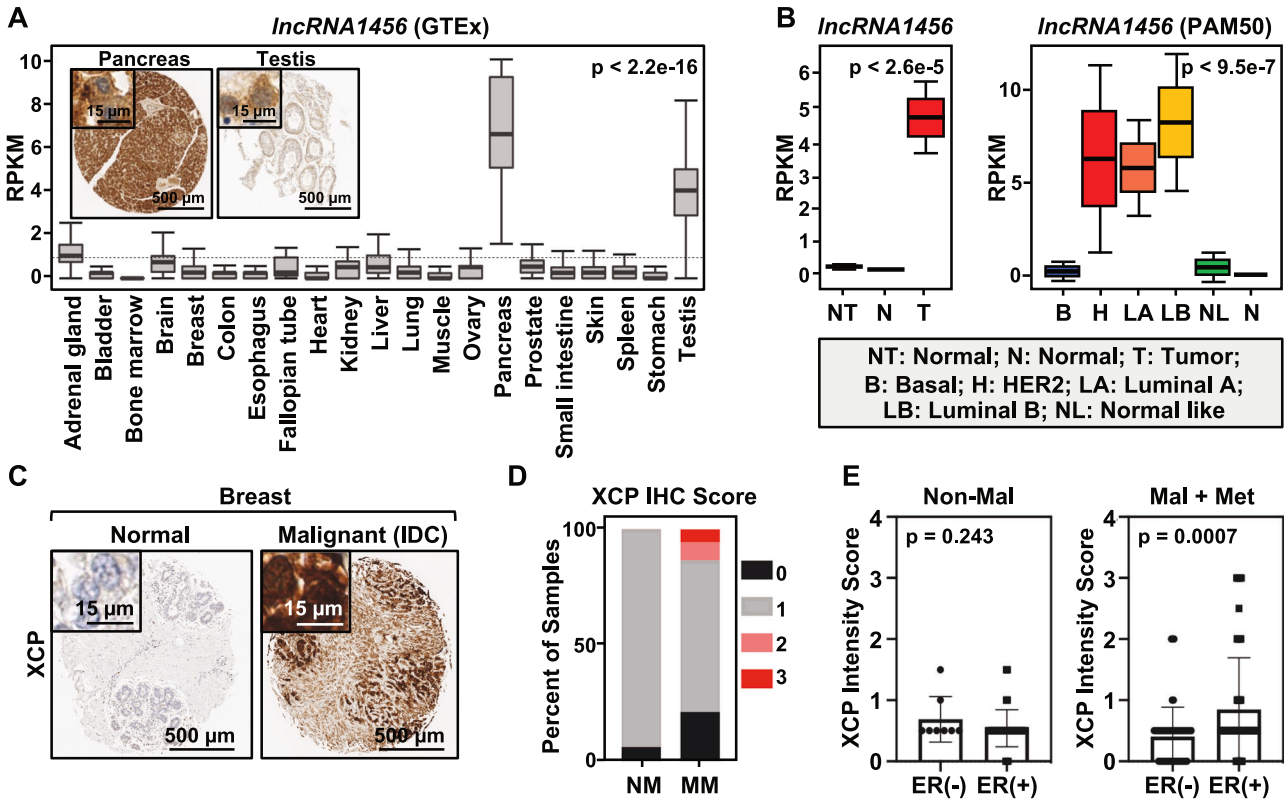

Fig. 2 Expression of *lncRNA1456* RNA and XCP peptide in human normal and breast cancer tissues. A GTEx survey of *lncRNA1456* RNA expression across normal human tissues. Observed differences are significant as determined by an ANOVA comparison of the means (*p*-value < 2.2 e-16). Insets show immunohistochemical staining of pancreas and testis using antibody against XCP. **B** Graph showing *lncRNA1456* RNA expression in normal solid tissue (NT; TCGA), normal breast tissue (N; GTEx) and malignant primary breast tumors (T; TCGA) (*left*) and *lncRNA1456* RNA expression in malignant primary breast tumors (TCGA) stratified into subtypes (B: basal; H: HER2; LA/LB: luminal A/B; NL: normal like) using PAM50 gene set analysis compared to normal breast tissue (N; GTEx) (*right*). Observed differences are significant as determined by an ANOVA comparison of the means. **C** Representative immunohistochemical staining for XCP peptide in normal versus malignant (IDC) breast tissue. **D** Frequency of XCP peptide expression across non-malignant (NM) and malignant plus metastatic (MM) breast tissues based on IHC staining intensity scores. **E** Expression scores based on IHC of XCP in non-malignant (*left*) and malignant plus metastatic (*right*) breast cancer samples, stratified by ER status. Each bar represents the mean ± SEM; For Non-Mal, ER(-) $n = 8$ and ER(+) $n = 36$; For Mal +Met, ER(-) n = 53 and ER(+) $n = 89$. Significance was calculated using unpaired t-test. [See also Fig. S2].

XCP interacts with the demethylase PHF8

Immunofluorescent staining of MCF-7 cells demonstrated XCP expression localized to the nucleus (Fig. S4A). A homology search identified homeobox-containing proteins as distantly related protein sequences (Fig. S4B). However, the regulation of gene expression by XCP and a lack of a DNA-binding domain prompted us to speculate that XCP may be acting on chromatin by interacting with a partner protein. To identify possible XCP binding partners and clues about the functions of XCP, we performed MS analysis of XCP-FLAG complexes pulled down from MCF-7 cell lysates (Fig. 5A). Among a number of candidates, MS analysis revealed a clear interaction with the histone demethylase, PHF8 (Table S1). This was confirmed in an in vitro pull down assay using FLAG-PHF8, in which XCP was also detected (Fig. 5B).

Further rationale in support for the selection of PHF8 as an XCP-interacting candidate to study in more detail was the similarities it shared with *lncRNA1456*: (1) location of the gene to the X chromosome; (2) high expression restricted to the testes in normal tissues (GTEx; Fig. 5C); (3) higher expression levels in normal breast and primary tumor tissues compared to normal solid tissues (Fig. 5D); and (4) higher expression in luminal A, luminal B, and HER2 compared to the basal molecular subtype of breast cancer (Fig. 5E). Although not significant, we also found a positive correlation of *PHF8* mRNA and *lncRNA1456* RNA expression in invasive breast carcinoma and non-seminomatous germ cell tumors (Fig. S4C, D). Collectively, this evidence supports the conclusion that XCP and PHF8 are co-expressed in the same tissues and can interact to regulate biological outcomes.

XCP regulates gene expression by modulating the binding of PHF8 to chromatin

Previous studies have demonstrated that PHF8 is a histone demethylase that acts as a transcriptional coregulator by removing repressive H3K9me2 and H3K9me1 modifications from chromatin [42–46]. To further investigate the function of the interactions between XCP and PHF8, we performed RNA-seq using MCF-7 cells ectopically expressing FLAG-GFP (as a control) or XCP-FLAG, with siRNA-mediated knockdown of either *lncRNA1456* RNA or *PHF8* mRNA, aimed at elevating or reducing XCP protein levels (Fig. S5A, B). Using the overlap of two different siRNAs for each target, we identified 1,449 and 1,070 genes commonly regulated by *lncRNA1456* or *PHF8* knockdown, respectively (1.5 fold change cutoff). Among these, 239 genes were commonly regulated by both *lncRNA1456* and *PHF8* knockdown (p-value of 1.12×10^{-127}; hypergeometric test, assuming a background of 20,000 protein-coding genes). Heat maps using these 239 genes illustrate that a subset of genes differentially regulated upon *lncRNA1456* knockdown can be rescued upon ectopic expression of XCP-FLAG (Fig. 6A–C; lane 2 versus lane 3). These genes are therefore classified as XCP-regulated, whereas genes whose expression is not restored by XCP-FLAG are likely *lncRNA1456*-regulated through RNA-dependent, protein-independent mechanisms. However, rescue of

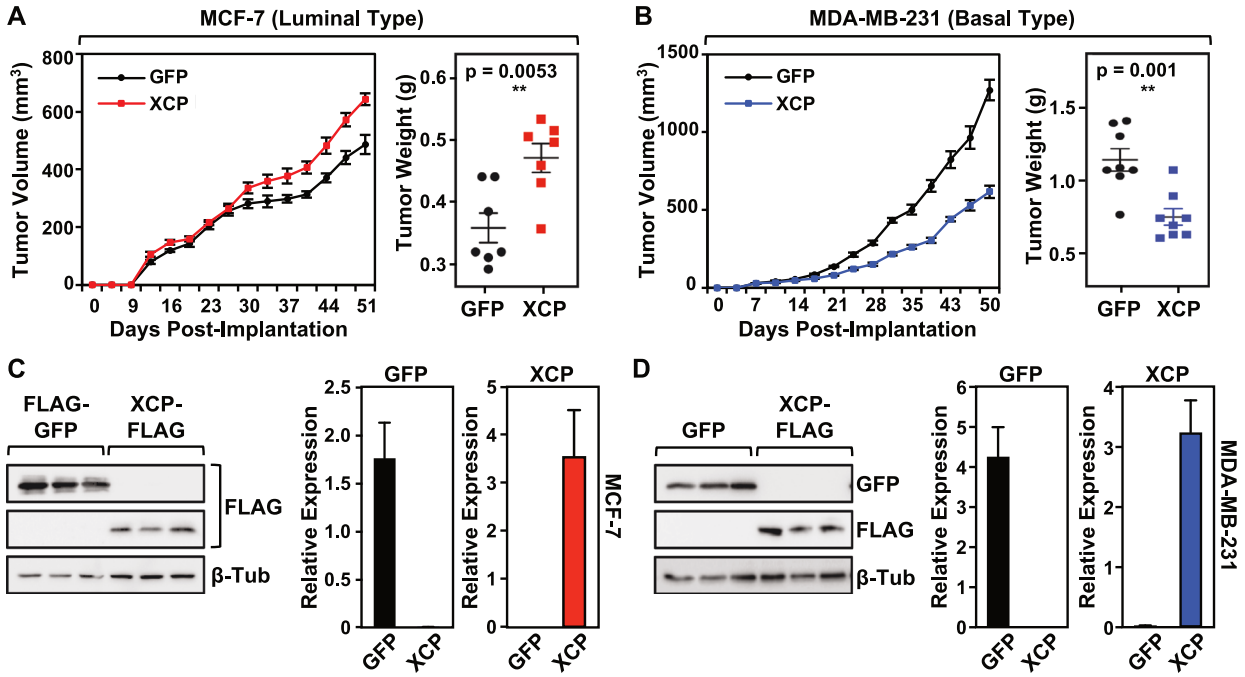

Fig. 3 XCP modulates tumor growth in a context-dependent manner. A,B Xenograft assays showing subcutaneous tumor growth of MCF-7 (**A**) or MDA-MB-231 (**B**) breast cancer cells ectopically expressing dox-inducible FLAG-GFP or XCP-FLAG (*left*). Tumor weight at end of experiment is shown (*right*). Animals injected with MCF-7 cells had E2-pellet implantations at the back of the neck to promote growth of MCF-7 cells in vivo. All animals were fed doxycycline in their chow. Each point represents the mean ± SEM. For MCF-7, GFP $n = 7$ and XCP $n = 7$; For MDA-MB-231, GFP $n = 8$ and XCP $n = 8$. Significance was calculated using unpaired t-test. **C,D** Expression of FLAG-GFP or XCP-FLAG in tumor tissue was validated from MCF-7 (**C**) or MDA-MB-231 (**D**) xenograft tumors by Western blot (*left*) and RT-qPCR (*right*). For Western blot, 3 representative tumors from each group are shown. β-tubulin was used as a loading control. For RT-qPCR, the average of all tumors from each group is shown. RNA levels were quantified by normalizing to housekeeping gene *RPL19* mRNA. Each bar represents the mean + SEM. For MCF-7, GFP $n = 7$ and XCP $n = 7$; For MDA-MB-231, GFP $n = 8$ and XCP $n = 8$. [See also Fig. S3].

the genes differentially regulated upon *PHF8* knockdown was attenuated upon ectopic expression of XCP-FLAG (Fig. 6A–C; lane 5 versus lane 6). GO analysis of the 239 genes indicates enrichment in terms related to response to estradiol, positive regulation of insulin secretion and apoptotic process (Fig. S5C). We observed less significant gene expression changes in the XCP-regulated genes that are not shared with *PHF8* knockdown, and this non-shared gene set is enriched for a distinct set of processes, mostly related to mitosis (Fig. S5D–F). These results suggests that PHF8 is necessary for XCP-mediated gene regulation.

Next, we performed ChIP-seq analysis to assess PHF8 binding across the genome following knockdown of *lncRNA1456* to reduce the levels of XCP. We observe that PHF8 binds primarily near promoters, both globally (Fig. S6A) and near the 239 genes that are regulated by XCP (Fig. 6D), and these promoter regions are highly enriched for H3K4me3, as expected (Fig. S6B). For XCP-regulated genes, PHF8 peaks are enriched within 200 kb of the transcription start sites (TSSs) (Fig. S6C). Interestingly, we detected a dramatic loss of PHF8 binding upon reduction in XCP levels, again both globally and near XCP-regulated genes (Fig. 6E, F; Fig. S6D). For example, the *IGFBP4* and *FOXA1* genes, both found to be modulated by XCP expression, show reduced PHF8 binding near promoters in response to reduced levels of XCP (Fig. 6G, H). Finally, no enrichment in PHF8 binding was observed in XCP-regulated genes that are not shared with *PHF8* knockdown (Fig. S6E, F). These results indicate that XCP directly controls the chromatin binding of PHF8 to regulate gene expression.

XCP modulates the demethylase activity of PHF8

To understand how XCP modulates PHF8 activity, we carried out in vitro demethylase reactions using recombinant FLAG-PHF8, His_6-XCP, and H3K9me2 peptide (Fig. 7A; *left panel*). We observed that PHF8 alone has a limited ability to demethylate the H3K9me2 peptide, as determined by dot blotting (Fig. 7A; *right panel*, lane 2). However, in the presence of XCP, PHF8 was able to robustly demethylate the H3K9me2 peptide (lane 4), and its activity was further enhanced with increasing amounts of XCP (lane 5). XCP alone did not alter the methylation of the H3K9me2 peptide (Fig. 7A; *right panel*, lane 3). We also assessed the binding of XCP and PHF8 to a biotinylated H3K9me2 peptide in a streptavidin pull down assay, and we found that XCP alone cannot bind the peptide, as expected based on sequence, and XCP does not enhance PHF8 binding to the peptide (Fig. S7A).

To examine the effect of XCP on PHF8-mediated demethylation globally in cells, we examined post-translational modification of histones by MS. We isolated histones from MCF-7 cells subjected to siRNA-mediated knockdown of *lncRNA1456* and ectopic co-expression of full-length *lncRNA1456*, which expresses XCP, or an XCP start codon (ATG) mutant that abolishes XCP production (Fig. 7B). We assayed methylation marks that are known to be the preferred targets of PHF8 by MS. We observed a significant reduction in H3K9me2 levels with expression of wild-type *lncRNA1456* (XCP expressed) compared to the ATG mutant (XCP absent) (Fig. 7C; *top panels*). This was validated by Western blotting for H3K9me2 on isolated histones, strengthening our observations (Fig. 7D). Interestingly, acetylation at the neighboring H3K14 residue abrogates the effect on H3K9 dimethylation in wild-type versus mutant *lncRNA1456* expressing cells (Fig. 7C; *bottom panels*). No significant changes were observed over H3K27me2, and H4K20me1 was undetectable in our samples (Fig. S7B). Overall, these results provide evidence for the global modulation of PHF8 demethylase activity by XCP, resulting in the transcriptional regulation of gene expression in breast cancer cells (Fig. 7E).

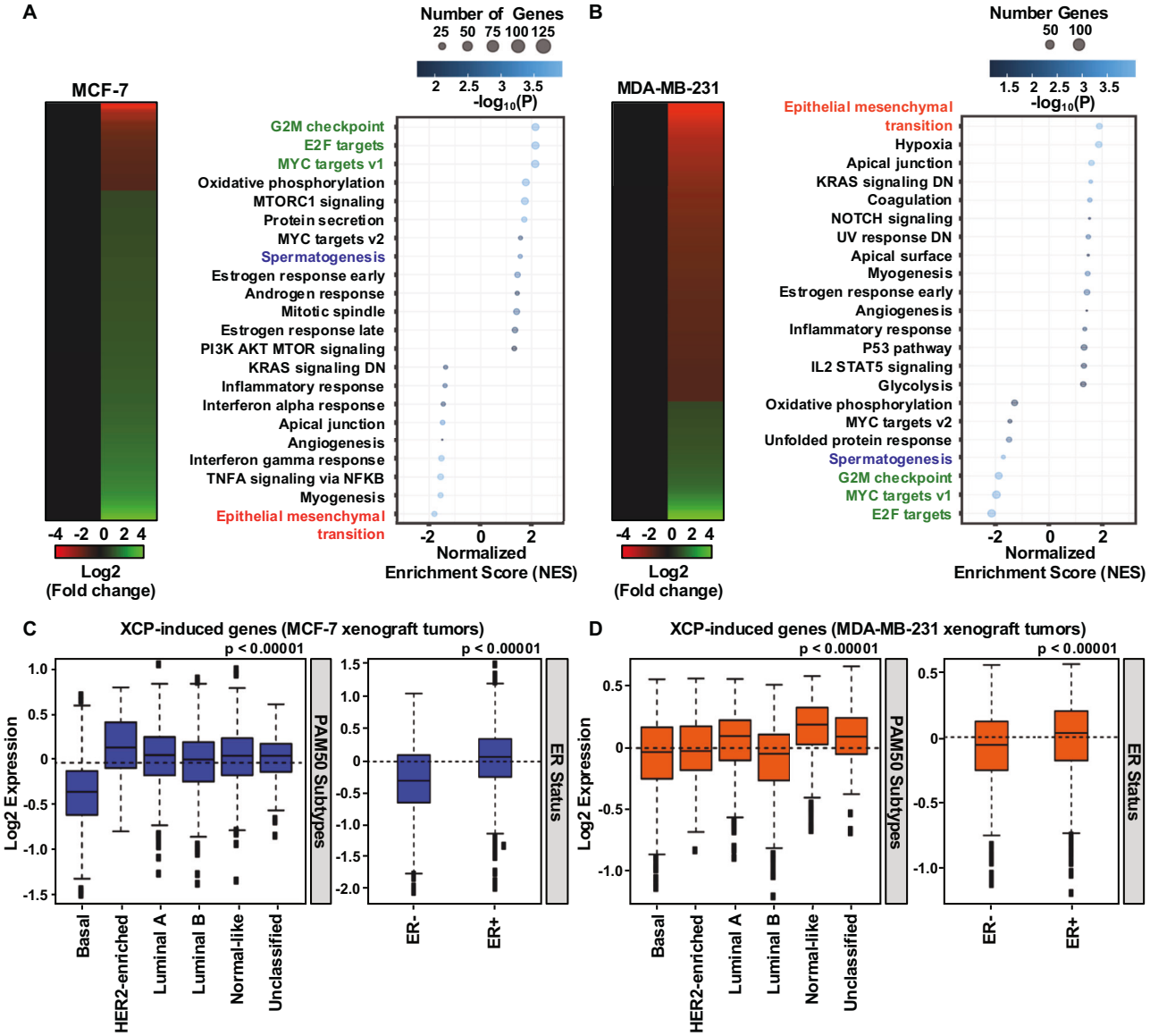

Fig. 4 XCP modulates gene expression and regulates cellular pathways in a context-dependent manner. A,**B** Heatmaps (*left*) and Gene Ontology analysis (*right*) showing XCP-mediated gene expression changes in MCF-7 (**A**) and MDA-MB-231 (**B**) xenograft tumors. **C**,**D** Box plots showing gene set analysis of XCP-induced genes and their expression levels stratified into breast cancer subtypes using PAM50 gene set analysis (*left*) and stratified by ER status (*right*) in MCF-7 (**C**) and MDA-MB-231 (**D**) xenograft tumors. Observed differences are significant as determined by an ANOVA comparison of the means (*P* value < 0.00001). [See also Fig. S3].

DISCUSSION

Expansion of genome-wide transcriptome analysis has revealed that mammalian genomes are pervasively transcribed, and the 'noncoding RNA' genome has garnered significant attention in recent years. LncRNAs are an interesting species of noncoding RNAs exhibiting a high level of cell type and developmental specificity; their expression often becomes dysregulated in disease states. For example, cancers carry a heavy mutational load in the noncoding genome, which could profoundly affect the expression of lncRNAs and, hence, cellular biology [47, 48]. Recently, it has become clear that some lncRNAs, which are sometimes mis-annotated as non-protein-coding, may contain translatable ORFs. Many lncRNAs that have been categorized as non-protein coding are often misclassified as such due to specific cutoffs for ORF sizes set by computational algorithms [49]. As a result, many potential ORFs were often marked as evolutionarily dispensable and, hence, not translatable. These ORFs can vary in size, producing anything on the order of 100 amino acids. In the last five to ten years, there is significant evidence from proteomics and Ribo-seq profiling showing that these ORFs, many encoded within this 'noncoding RNA' genome, are abundant, evolutionarily young, and biologically functional [9, 50–53]. The lack of characterization of specific cellular roles for these small functional peptides remains a gap in our knowledge.

XCP is an X-linked cancer protein

We have identified a lncRNA-encoded peptide, which we have named XCP, and demonstrated its biological relevance in breast cancer cells. The gene encoding XCP, *lncRNA1456*, like many others transcribed in the testis, is located on the X chromosome, and its sequence conservation is primarily observed in humans and non-human primates. Although XCP gene expression is largely restricted to normal tissues such as the testis and pancreas, it can escape regulation in malignant cells from females, likely due to aberrant epigenetic regulation that presumably leads to the reversal of X-inactivation. In fact, the majority of novel transcripts

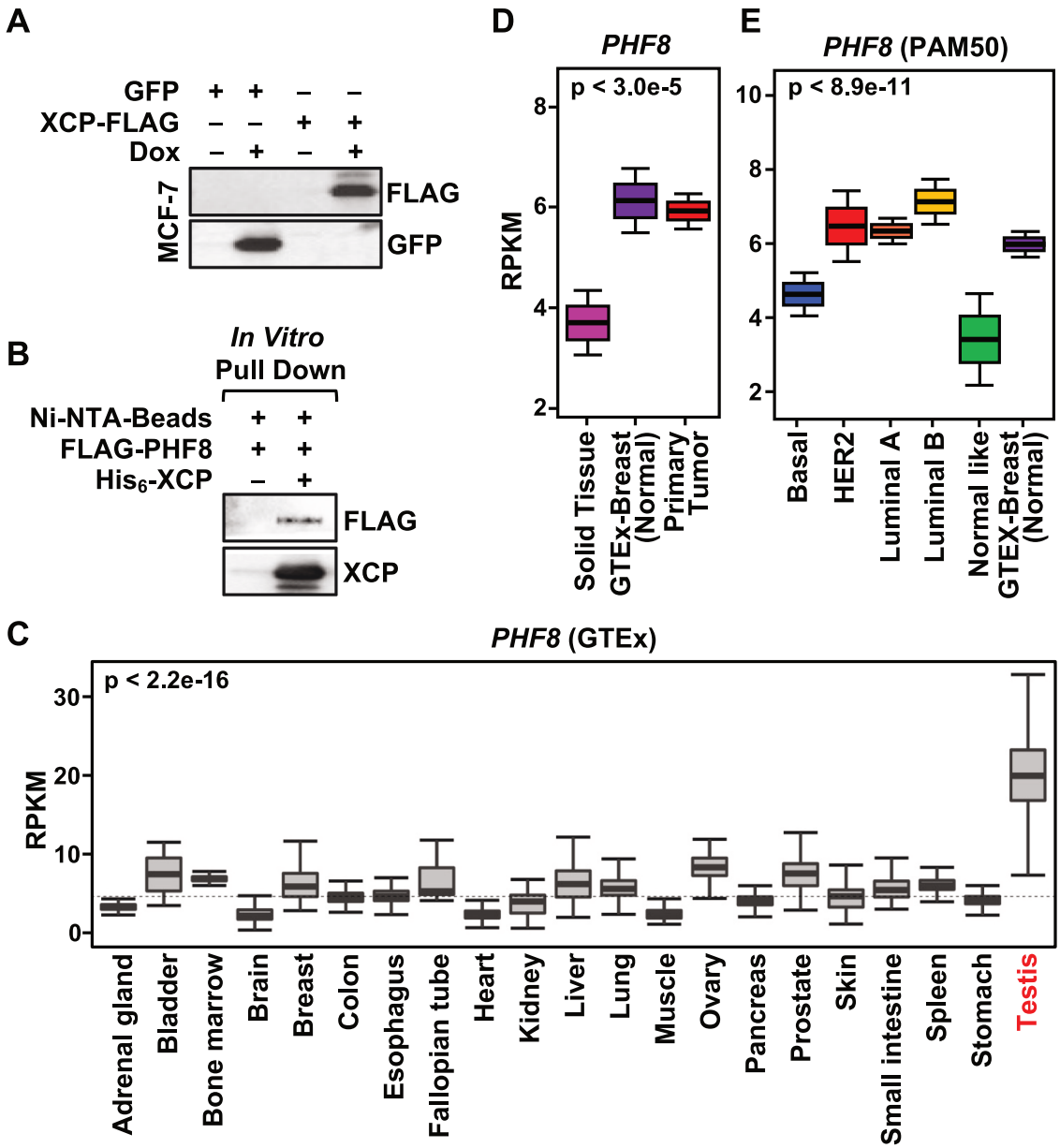

Fig. 5 XCP interacts with the demethylase PHF8. A Western blot showing expression of GFP and FLAG-tagged XCP used for pull down assay from MCF-7 whole cell lysates. **B** Western blot showing interaction of PHF8 and XCP confirmed in an in vitro pull down assay using recombinant FLAG-PHF8 and His_6-XCP. **C** GTEx survey of *PHF8* expression across normal human tissues. Observed differences are significant as determined by an ANOVA comparison of the means (*p*-value < 2.2 e-16). **D** Graph showing *PHF8* RNA expression in normal solid tissue (TCGA), normal breast tissue (GTEx) and malignant primary breast tumors (TCGA). Observed differences are significant as determined by an ANOVA comparison of the means. **E** *PHF8* RNA expression in malignant primary breast tumors (TCGA) stratified into subtypes using PAM50 gene set analysis compared to normal breast tissue (GTEx). Observed differences are significant as determined by an ANOVA comparison of the means. [See also Fig. S4].

(coding and noncoding) that are frequently identified from cancer cells have restricted expression in the testis and originate from the X chromosome [54]. We propose, based on the aberrant expression of XCP in breast cancer cells, and its encoding on the X chromosome, that XCP can be classified as an X-linked cancer antigen.

XCP has dual, context-dependent functions in gene regulation

Our findings indicate that XCP is upregulated not only at the RNA level, but also at the protein level in breast cancers, particularly in non-aggressive breast cancer (luminal subtype), while it is found at reduced or minimal levels in more aggressive cases (basal subtypes). This is in agreement with our observation that ectopic expression of XCP in MCF-7 cells (luminal or ER+ cells) promotes E2-dependent tumor growth, whereas an inhibition of tumor growth is observed in MDA-MB-231 (basal or ER–) cell line.

While XCP expression correlates with ER expression (as in MCF-7 cells, for example), the use of MDA-MB-231 as a heterologous system allowed us to assess whether XCP might exert effects in an ER-negative context. The different effects of XCP in different cellular contexts can be attributed to the specific genes and cellular pathways it regulates in ER+ and ER– cells, as identified through genome-wide expression analyses. Furthermore, the expression patterns of XCP-regulated genes correspond to the breast cancer subtypes. For instance, genes regulated in ER+

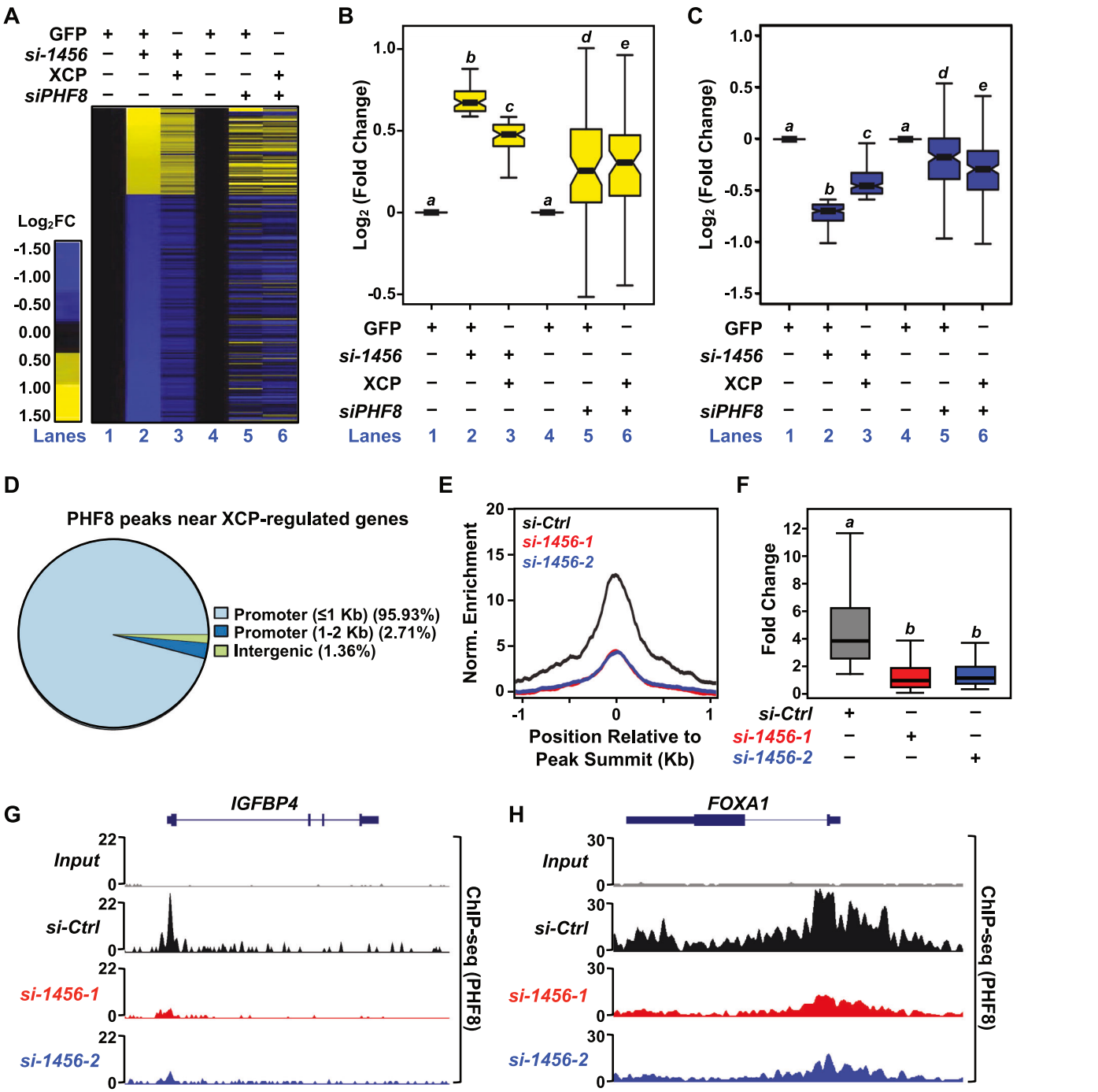

Fig. 6 XCP modulates gene expression through its interaction with the demethylase PHF8. A Heatmap from RNA-seq data showing the regulation of 239 genes upon siRNA-mediated knockdown of *lncRNA1456* RNA or *PHF8* mRNA and co-expression of XCP-FLAG. **B**,**C** Box plots quantifying the aggregate upregulated (**B**) and downregulated (**C**) gene expression changes (Log2 fold change) observed in heatmap of 239 XCP-regulated genes (**A**). Only genes exhibiting at least a 1.5-fold change were included in the analysis. Each bar represents the mean ± SEM; Bars marked with different letters are significantly different from each other, Wilcoxon rank sum test. **D** Pie chart showing the distribution of PHF8 peaks at the 239 XCP-regulated genes from PHF8 ChIP-seq analyses. **E**,**F** Metagene plot (**E**) and box plot (**F**) showing the reduced enrichment of PHF8 binding upon siRNA-mediated knockdown of *lncRNA1456* near 239 XCP-regulated genes from ChIP-seq analyses. Each bar represents the mean ± SEM. Only genes exhibiting at least a 1.5-fold change were included in the analysis. Bars marked with different letters are significantly different from each other, Wilcoxon rank sum test. **G**,**H** ChIP-seq browser tracks showing reduced PHF8 binding near promoter of *IGFBP4* and *FOXA1* genes, 2 representative XCP-regulated genes. [See also Figs. S5 and S6].

breast cancer cells were also expressed in luminal breast cancer subtypes, while genes regulated in triple-negative breast cancer cells were also expressed in basal breast cancer subtypes. This context-dependent regulation may be due to a distinct set of proteins or molecules with which XCP interacts. A similar observation has been made with other proteins specific to epithelial cells, such as *FOXA1* or *GATA3*, which reduce cell growth when overexpressed in basal cell lines [55, 56].

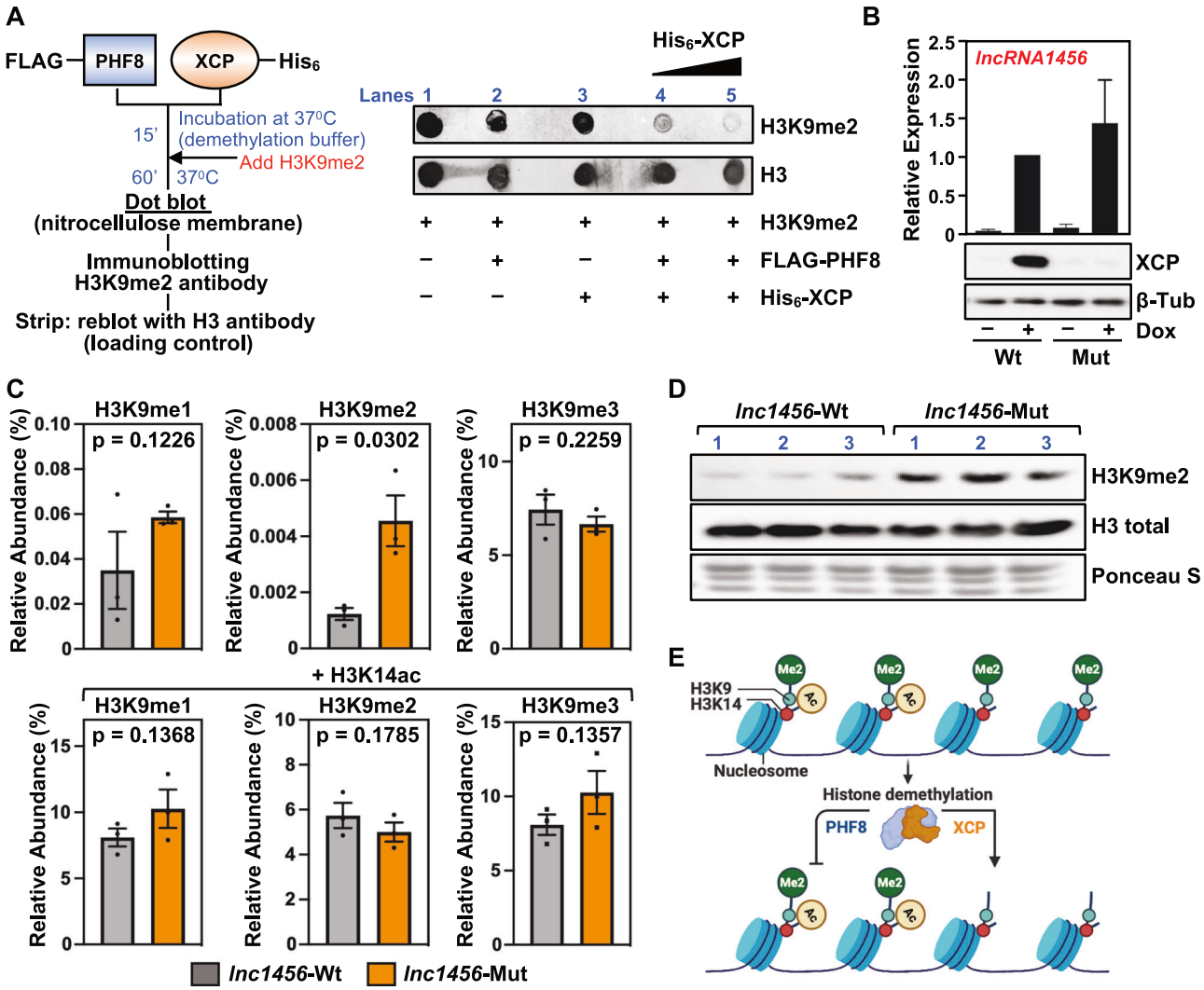

Fig. 7 XCP directly modulates the demethylase activity of PHF8. A Diagram showing the flow of in vitro demethylase reactions using recombinant FLAG-PHF8, His_6-XCP and H3K9me2 peptide (*left*). Dot bot showing the demethylase activity of PHF8 on H3K9me2 peptide in the presence or absence of XCP (*right*). Total Histone H3 was used as a loading control for dot blot. **B** RT-qPCR (*top*) and Western blot (*bottom*) showing dox-inducible ectopic expression of full length *lncRNA1456* in its wild-type form (Wt), which expresses the XCP peptide, or an ATG mutant (Mut) version which abolishes XCP peptide production. For RT-qPCR, RNA levels were quantified by normalizing to housekeeping gene *RPL19* mRNA. Each bar represents the mean + SEM; $n = 3$. For Western blot, β-tubulin was used a loading control. **C** Bar graphs showing global changes in H3K9 methylation modifications assayed by mass spectrometry of MCF-7 histones in the absence (*top*) or presence (*bottom*) of neighboring H3K14ac. Each bar represents the mean ± SEM; $n = 3$. Significance calculated by 1-sided Welch's t test. **D** Western blot analysis of MCF-7 histone preparation used for mass spectrometry analysis (**C**), corroborating levels of H3K9me2. Total Histone H3 was used as a loading control. **E** Diagram summarizing model: XCP, encoded by *lncRNA1456*, interacts with PHF8, a histone demethylase, and directly regulates its activity to modulate gene expression. Diagram generated using BioRender. [See also Fig. S7].

XCP functions as an epigenomic regulator via interaction with and coregulation of PHF8

XCP localizes to the nucleus, suggesting a role in the regulation of gene expression, and partners with PHF8, suggesting a role in chromatin regulation. Similar to a recently published study, where a micropeptide is capable of activating the enzymatic activity of GSK-3β [57], XCP is another example of a small polypeptide that can directly modulate the activity of an enzyme (PHF8). We found that XCP interacts with and is required for the binding of PHF8 to chromatin. Surprisingly, XCP also stimulates the histone demethylase activity of PHF8, a unique role for a lncRNA-encoded polypeptide. These studies identify a functional partnership between two X chromosome-encoded chromatin regulators that allows them to collectively enhance demethylation of H3K9me2 to support chromatin-dependent gene regulation.

Although homology searches indicated similarities to proteins containing DNA-binding domains, XCP itself does not contain amino acid sequences similar to those domains. This observation led us to hypothesize that XCP interacts with proteins that have DNA-binding capabilities. In this regard, we identified PHF8, which contains a plant homeodomain, a known DNA-binding domain. Notably, similar to XCP, PHF8 expression is higher in the testis and is upregulated in luminal subtype of breast cancer, while its expression is downregulated in aggressive breast cancers, including basal subtypes. These similarities between XCP and PHF8 strongly indicate their cooperative roles in regulating gene expression.

PHF8 is a histone demethylating enzyme and an epigenomic modulator that has been shown to demethylate H3K9me2, a repressive chromatin modification, thereby activating transcription and regulating gene expression. Although our study focuses

on breast cancer, it is noteworthy that PHF8 has been linked to X-chromosome-linked intellectual disability (XLID) affecting learning and memory [58], craniofacial development in zebrafish [59], neuronal differentiation [60], and X-linked mental retardation [42]. It would be interesting to explore whether XCP may contribute to PHF8 regulation in the brain.

Through its interactions with PHF8, XCP functions to support PHF8-driven epigenomic modification and gene expression (Fig. 7E). This is analogous to other coactivators that interact with epigenetic modulators to regulate gene expression; for example, specific coactivators interact with the histone acetyltransferase p300 to enhance acetylation at promoters and promote transcription [61]. However, we also observed that depletion of XCP (siRNA-mediated knockdown of *lncRNA1456*) reduced PHF8's binding to chromatin but did not enhance its binding to H3K9me2 peptide in an in vitro assay (Figs. S6 and S7). This discrepancy could be attributed to the varying effects of XCP on its interaction with PHF8 in the presence or absence of chromatin. We also observe a global, *lncRNA1456*-dependent change in H3K9me2 levels (Fig. 7C, D), however, this does not translate into global corresponding changes in gene expression. This may be explained by the fact that H3K9me2 is primarily associated with constitutive heterochromatin [62], where gene activation generally requires coordinated input from additional regulatory factors, such as activating transcription factors, rather than histone modification changes alone [63–65]. Thus, the impact of altered H3K9me2 on transcription is likely highly context-dependent. Within cells, XCP associates not only with PHF8 but also with the PHF8 complex, which may contain other chromatin-modulating factors. Also, it is plausible that PHF8 is recruited to target chromatin regions through its DNA-binding domain, followed by an enhanced demethylase activity mediated by a chromatin-modulating complex containing XCP. Additionally, it is possible that XCP is recruited to chromatin prior to binding of PHF8, which facilitates its enzymatic activity. Furthermore, as in MCF-7 cells, XCP may also interact with PHF8 in MDA-MB-231 cells to regulate genes that suppress cell growth. This interaction likely functions through a context-dependent mechanism involving transcriptional regulatory complexes that activate anti-proliferative gene programs and related cellular processes. However, further detailed studies are required to understand this process.

Previous studies have shown that coregulators can modulate the enzymatic activity of histone modifiers, such as acetyltransferases and methyltransferases [61]. Specifically, XCP directly interacts with and modulates the demethylase activity of PHF8, enhancing demethylation of H3K9me2. Although functional lncRNAs can function independently of protein translation, ectopic expression of a mutant form of *lncRNA1456*, which cannot be translated into XCP, did not retain the same functionality as wild-type *lncRNA1456*. Thus, key functions of *lncRNA1456* require the protein coding of XCP. Given the low expression levels of PHF8 in cells, it is plausible that it relies on interactions with other proteins like XCP, resulting in a synergistic effect that enhances transcriptional regulation efficiency.

Mechanistically, we show that XCP is a chromatin-associated protein that binds to a subset of PHF8-binding regions; this interaction regulates the demethylase activity of PHF8, thereby modulating the gene expression of target genes in breast cancer cells (Fig. 7E). Our study demonstrates how lncRNA-encoded peptides/proteins can drive cancer-specific phenotypes and serve as potential biomarkers and/or targets for therapeutic intervention.

DATA AVAILABILITY

All of the RNA-seq, ChIP-seq and MS data sets used in this study can be accessed through the NCBI's Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/) (RRID:SCR_005012) using accession number GSE288868 for RNA-seq, and GSE287423 for ChIP-seq, or the MS Interactive Virtual Environment (MassIVE) repository (https://massive.ucsd.edu/) (RRID:SCR_013665) using the accession number MSV000096645.

REFERENCES

1. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489:101–8.
2. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics. 2011;27:i275–82.
3. Scheidler CM, Kick LM, Schneider S. Ribosomal peptides and small proteins on the rise. Chembiochem. 2019;20:1479–86.
4. Huarte M. The emerging role of lncRNAs in cancer. Nat Med. 2015;21:1253–61.
5. Barczak W, Carr SM, Liu G, Munro S, Nicastri A, Lee LN, et al. Long non-coding RNA-derived peptides are immunogenic and drive a potent anti-tumour response. Nat Commun. 2023;14:1078.
6. Wu P, Mo Y, Peng M, Tang T, Zhong Y, Deng X, et al. Emerging role of tumor-related functional peptides encoded by lncRNA and circRNA. Mol Cancer. 2020;19:22.
7. Zhang Y, Wang X, Hu C, Yi H. Shiny transcriptional junk: lncRNA-derived peptides in cancers and immune responses. Life Sci. 2023;316:121434.
8. Zheng C, Wei Y, Zhang P, Xu L, Zhang Z, Lin K, et al. CRISPR/Cas9 screen uncovers functional translation of cryptic lncRNA-encoded open reading frames in human cancer. *J* Clin Invest. 2023;133:e159940.
9. Sandmann CL, Schulz JF, Ruiz-Orera J, Kirchner M, Ziehm M, Adami E, et al. Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. Mol Cell. 2023;83:994–1011.
10. Matsumoto A, Pasut A, Matsumoto M, Yamashita R, Fung J, Monteleone E, et al. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. Nature. 2017;541:228–32.
11. Chugunova A, Loseva E, Mazin P, Mitina A, Navalayeu T, Bilan D, et al. LINC00116 codes for a mitochondrial peptide linking respiration and lipid metabolism. Proc Natl Acad Sci USA. 2019;116:4940–5.
12. Jackson R, Kroehling L, Khitun A, Bailis W, Jarret A, York AG, et al. The translation of non-canonical open reading frames controls mucosal immunity. Nature. 2018;564:434–8.
13. Lu S, Zhang J, Lian X, Sun L, Meng K, Chen Y, et al. A hidden human proteome encoded by 'non-coding' genes. Nucleic Acids Res. 2019;47:8111–25.
14. Zheng W, Guo Y, Zhang G, Bai J, Song Y, Song X, et al. Peptide encoded by lncRNA BVES-AS1 promotes cell viability, migration, and invasion in colorectal cancer cells via the SRC/mTOR signaling pathway. PLoS One. 2023;18:e0287133.
15. Ye M, Gao R, Chen S, Bai J, Chen J, Lu F, et al. FAM201A encodes small protein NBASP to inhibit neuroblastoma progression via inactivating MAPK pathway mediated by FABP5. Commun Biol. 2023;6:714.
16. Tan Z, Zhao L, Huang S, Jiang Q, Wei Y, Wu JL, et al. Small peptide LINC00511-133aa encoded by LINC00511 regulates breast cancer cell invasion and stemness through the Wnt/beta-catenin pathway. Mol Cell Probes. 2023;69:101913.
17. Li W, Yu Y, Zhou G, Hu G, Li B, Ma H, et al. Large-scale ORF screening based on LC-MS to discover novel lncRNA-encoded peptides responding to ionizing radiation and microgravity. Comput Struct Biotechnol J. 2023;21:5201–11.
18. Kim DS, Camacho CV, Setlem R, Kim K, Malladi S, Hou TY, et al. Functional characterization of lncRNA152 as an angiogenesis-inhibiting tumor suppressor in triple-negative breast cancers. Mol Cancer Res. 2022;20:1623–35.
19. Sun M, Gadad SS, Kim DS, Kraus WL. Discovery, annotation, and functional analysis of long noncoding RNAs controlling cell-cycle gene expression and proliferation in breast cancer cells. Mol Cell. 2015;59:698–711.
20. Zhong S, Joung JG, Zheng Y, Chen YR, Liu B, Shao Y, et al. High-throughput illumina strand-specific RNA sequencing library preparation. Cold Spring Harb Protoc. 2011;2011:940–9.
21. Volders PJ, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J, et al. An update on LNCipedia: a database for annotated human lncRNA sequences. Nucleic Acids Res. 2015;43:D174–80.
22. Quinlan AR. BEDTools: The Swiss-Army tool for genome feature analysis. Curr Protoc Bioinforma. 2014;47:11 12 11–34.
23. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16:276–7.
24. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004;32:D115–9.
25. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. Methods Mol Biol. 2016;1374:23–54.
26. Gibson BA, Conrad LB, Huang D, Kraus WL. Generation and characterization of recombinant antibody-like ADP-ribose binding proteins. Biochemistry. 2017;56:6305–16.

27. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. Nat Biotechnol. 2017;35:319–21.
28. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 2015;43:D1079–85.
29. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.
30. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.
31. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
33. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012;22:1813–31.
34. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nat Protoc. 2012;7:1728–40.
35. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44:W160–5.
36. Huang D, Edwards AD, Gong X, Kraus WL. Functional Analysis of Histone ADP-Ribosylation In Vitro and in Cells. Methods Mol Biol. 2023;2609:157–92.
37. Shechter D, Dormann HL, Allis CD, Hake SB. Extraction, purification and analysis of histones. Nat Protoc. 2007;2:1445–57.
38. Garcia BA, Mollah S, Ueberheide BM, Busby SA, Muratore TL, Shabanowitz J, et al. Chemical derivatization of histones for facilitated analysis by mass spectrometry. Nat Protoc. 2007;2:933–8.
39. Huang D, Camacho CV, Setlem R, Ryu KW, Parameswaran B, Gupta RK, et al. Functional interplay between histone H2B ADP-ribosylation and phosphorylation controls adipogenesis. Mol Cell. 2020;79:934–49.e914.
40. Xiang R, Ma L, Yang M, Zheng Z, Chen X, Jia F, et al. Increased expression of peptides from non-coding genes in cancer proteomics datasets suggests potential tumor neoantigens. Commun Biol. 2021;4:496.
41. Ovcharenko I, Nobrega MA, Loots GG, Stubbs L. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. Nucleic Acids Res. 2004;32:W280–6.
42. Kleine-Kohlbrecher D, Christensen J, Vandamme J, Abarrategui I, Bak M, Tommerup N, et al. A functional link between the histone demethylase PHF8 and the transcription factor ZNF711 in X-linked mental retardation. Mol Cell. 2010;38:165–78.
43. Yu L, Wang Y, Huang S, Wang J, Deng Z, Zhang Q, et al. Structural insights into a novel histone demethylase PHF8. Cell Res. 2010;20:166–73.
44. Loenarz C, Ge W, Coleman ML, Rose NR, Cooper CD, Klose RJ, et al. PHF8, a gene associated with cleft lip/palate and mental retardation, encodes for an Nepsilon-dimethyl lysine demethylase. Hum Mol Genet. 2010;19:217–22.
45. Suganuma T, Workman JL. Features of the PHF8/KIAA1718 histone demethylase. Cell Res. 2010;20:861–2.
46. Zhu Z, Wang Y, Li X, Wang Y, Xu L, Wang X, et al. PHF8 is a histone H3K9me2 demethylase regulating rRNA synthesis. Cell Res. 2010;20:794–801.
47. Schmitt AM, Chang HY. Long noncoding RNAs in cancer pathways. Cancer Cell. 2016;29:452–63.
48. Xue B, He L. An expanding universe of the non-coding genome in cancer biology. Carcinogenesis. 2014;35:1209–16.
49. Pauli A, Valen E, Schier AF. Identifying (non-)coding RNAs and small peptides: challenges and opportunities. Bioessays. 2015;37:103–12.
50. Baena-Angulo C, Platero AI, Couso JP. Cis to trans: small ORF functions emerging through evolution. Trends Genet. 2025;41:119–31.
51. Tong G, Martinez TF. Ribosome profiling reveals hidden world of small proteins. Trends Genet. 2025;41:101–3.
52. Ruiz Cuevas MV, Hardy MP, Holly J, Bonneil E, Durette C, Courcelles M, et al. Most non-canonical proteins uniquely populate the proteome or immunopeptidome. Cell Rep. 2021;34:108815.
53. Chothani S, Ho L, Schafer S, Rackham O. Discovering microproteins: making the most of ribosome profiling data. RNA Biol. 2023;20:943–54.
54. Betran E, Thornton K, Long M. Retroposed new genes out of the X in Drosophila. Genome Res. 2002;12:1854–9.
55. Bernardo GM, Bebek G, Ginther CL, Sizemore ST, Lozada KL, Miedler JD, et al. FOXA1 represses the molecular phenotype of basal breast cancer cells. Oncogene. 2013;32:554–63.
56. Kouros-Mehr H, Bechis SK, Slorach EM, Littlepage LE, Egeblad M, Ewald AJ, et al. GATA-3 links tumor differentiation and dissemination in a luminal breast cancer model. Cancer Cell. 2008;13:141–52.
57. Zhang Z, Li F, Dai X, Deng J, Wang Y, Zhang S, et al. A novel micropeptide miPEP205 suppresses the growth and metastasis of TNBC. Oncogene. 2025;44:513–29.
58. Chen X, Wang S, Zhou Y, Han Y, Li S, Xu Q, et al. Phf8 histone demethylase deficiency causes cognitive impairments through the mTOR pathway. Nat Commun. 2018;9:114.
59. Qi HH, Sarkissian M, Hu GQ, Wang Z, Bhattacharjee A, Gordon DB, et al. Histone H4K20/H3K9 demethylase PHF8 regulates zebrafish brain and craniofacial development. Nature. 2010;466:503–7.
60. Qiu J, Shi G, Jia Y, Li J, Wu M, Li J, et al. The X-linked mental retardation gene PHF8 is a histone demethylase involved in neuronal differentiation. Cell Res. 2010;20:908–18.
61. Johnson AB, O'Malley BW. Steroid receptor coactivators 1, 2, and 3: critical regulators of nuclear receptor activity and steroid receptor modulator (SRM)-based cancer therapy. Mol Cell Endocrinol. 2012;348:430–9.
62. Padeken J, Methot SP, Gasser SM. Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance. Nat Rev Mol Cell Biol. 2022;23:623–40.
63. Jozwik KM, Chernukhin I, Serandour AA, Nagarajan S, Carroll JS. FOXA1 directs H3K4 monomethylation at enhancers via recruitment of the methyltransferase MLL3. Cell Rep. 2016;17:2715–23.
64. Messier TL, Boyd JR, Gordon JA, Stein JL, Lian JB, Stein GS. Oncofetal epigenetic bivalency in breast cancer cells: H3K4 and H3K27 tri-methylation as a biomarker for phenotypic plasticity. J Cell Physiol. 2016;231:2474–81.
65. Messier TL, Gordon JA, Boyd JR, Tye CE, Browne G, Stein JL, et al. Histone H3 lysine 4 acetylation and methylation dynamics define breast cancer subtypes. Oncotarget. 2016;7:5094–109.

ACKNOWLEDGEMENTS

We thank Yasmin M. Vasquez for assistance with ChIP and members of the Kraus lab for their helpful comments and support. We also acknowledge and thank the following UT Southwestern core facilities: the Live Cell Imaging Core for microscopy support (Dr. Katherine Luby Phelps; 1S10OD021684-01), the Next Generation Sequencing Core for deep sequencing services (Dr. Ralf Kittler), the Proteomics Core Facility for MS (Dr. Andrew Lemoff), and the Tissue Management Shared Resource at the Simmons Cancer Center (NCI; 5P30CA142543) for immunohistochemical support. S.S.G. is a CPRIT scholar in cancer research and is supported by a first-time faculty recruitment award from the Cancer Prevention and Research Institute of Texas (CPRIT; RR170020). This work was initiated with support from a grant from the Cancer Prevention and Research Institute of Texas (CPRIT; RP190235) and funds from the Cecil H. and Ida Green Center for Reproductive Biology Sciences Endowment to W.L.K. S.S.G. is also supported by grants from the American Cancer Society (RSG-22-170-01-RMC), NIH 1R16GM149497, and CPRIT-TREC (RP230420).

AUTHOR CONTRIBUTIONS

SSG and WLK conceived this project and oversaw its execution. SSG and CVC designed and performed the experiments. SSG performed all biochemical experiments. XG prepared samples for the histone MS and MT analyzed MS data. VSM developed a bioinformatics pipeline to detect ORFs and analyzed TCGA, GTEx, and initial mass-spec data sets. CVC performed the in vivo experiments, analysis of human tissue arrays, and ChIPs. AN, VSM, TN, and SK analyzed ChIP-seq data. AN, AS, and TN analyzed RNA-seq data. YP performed pathological review of human tissue arrays. SSG and CVC prepared initial drafts of the figures and text, which were edited and finalized by WLK.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

All animal experiments were performed in compliance with the Institutional Animal Care and Use Committee (IACUC) at the UT Southwestern Medical Center (APN 2015-101155). All procedures were carried out in accordance with institutional and national ethical guidelines for animal research. All human breast tissue samples were obtained from a commercial repository (TissueArray.com) as non-identifiable samples, therefore no informed consent was required.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41388-026-03740-w.

Correspondence and requests for materials should be addressed to W. Lee Kraus.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.