Supplementary Information

RNA-mediated gene fusion in mammalian cells

Sachin Kumar Gupta, Liming Luo, and Laising Yen*

Department of Pathology & Immunology, Department of Molecular and Cellular Biology, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX 77030, USA

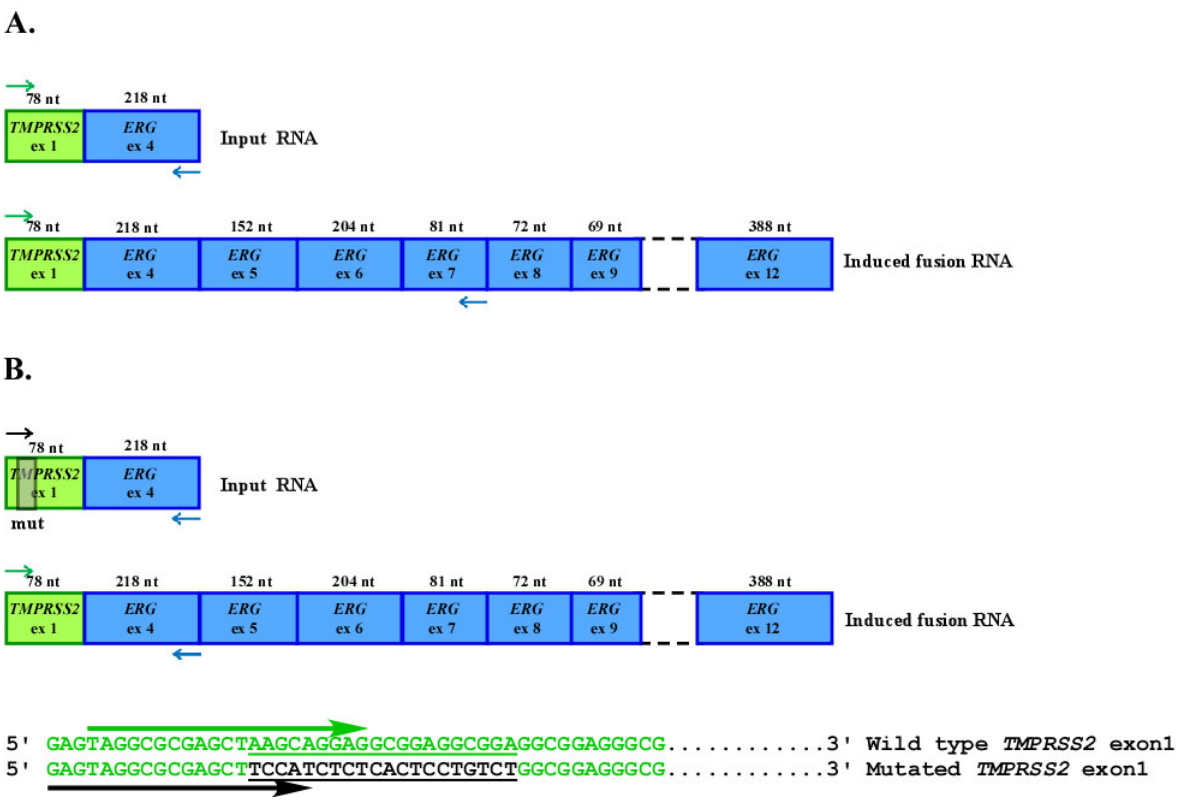


Fig. S1. Primer design used to distinguish induced endogenous fusion RNAs from exogenous input RNAs.

A. The endogenous full-length *TMPRSS2-ERG* fusion RNA most commonly found in prostate cancer consisting of *TMPRSS2* exon 1 (78 nt, uc002yzj.3) spliced to *ERG* exons 4-12 (1289 nt, uc021wjd.1). In Fig. 1A in the main text, the exogenous input RNAs expressed from plasmids consists of *TMPRSS2* exon-1 (78 nt) fused to *ERG* exon-4 (218 nt). The RT-PCR results shown in Fig. 1A and 1B (middle panel) in the main text were performed using a forward primer on *TMPRSS2* exon-1 paired with a reverse primer on *ERG* exon-7, thereby specifically amplifying the endogenous fusion RNA transcribed from the *TMPRSS2-ERG* fusion gene, but not from transfected plasmids or input RNAs because both lack *ERG* exon-7. **B.** Because RNA-induced fusion is a low frequency event, we developed a more efficient PCR method to detect induced endogenous fusion RNA using a primer pair targeting *TMPRSS2* exon-1 and *ERG* exon-4. To distinguish induced endogenous fusion RNAs from exogenous input RNAs, mutations were introduced at nucleotide positions 16-35 of the input RNAs (gray box in *TMPRSS2* exon-1), allowing specific amplifications of induced endogenous fusion RNAs or exogenous input RNA using specific forward primers. The lower panel shows the wild type *TMPRSS2* exon-1 sequence in green. The mutated area in input RNA is shown in black and underlined. The primer used to specifically recognize induced fusion RNA is shown as the green arrow and was used in most of the figures in the main text. The primer for input RNA is denoted by the black arrow and was used in Fig. 1B, 1C, 1D, 1E, Fig. 2B, and Fig. 3B in the main text.

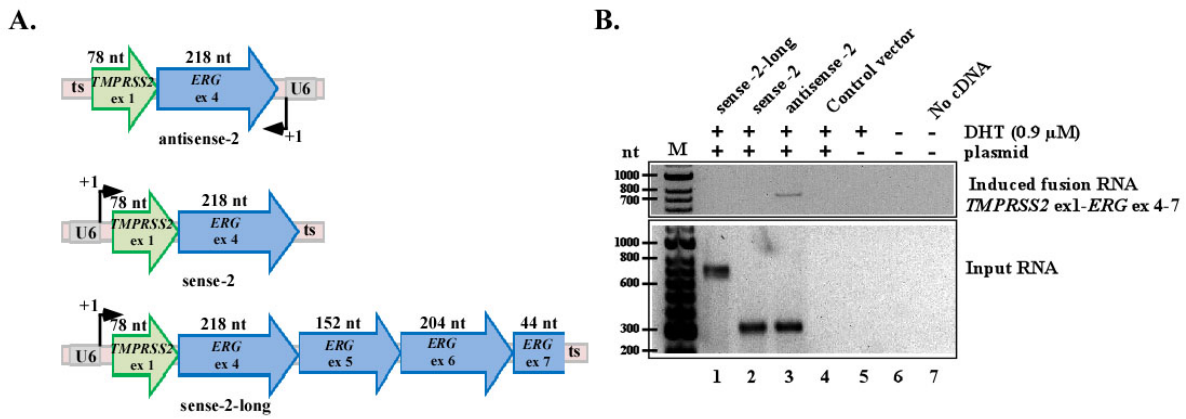
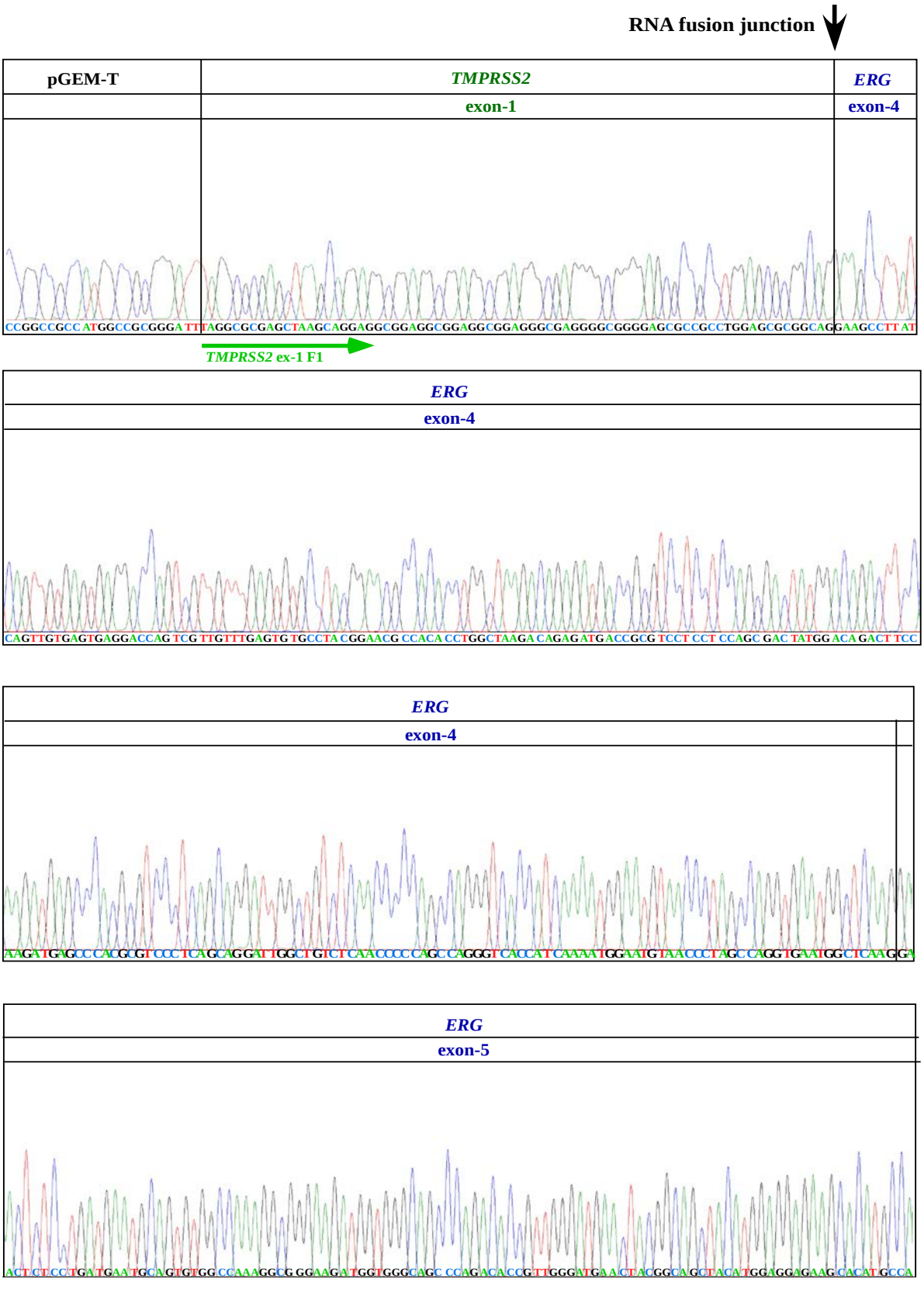
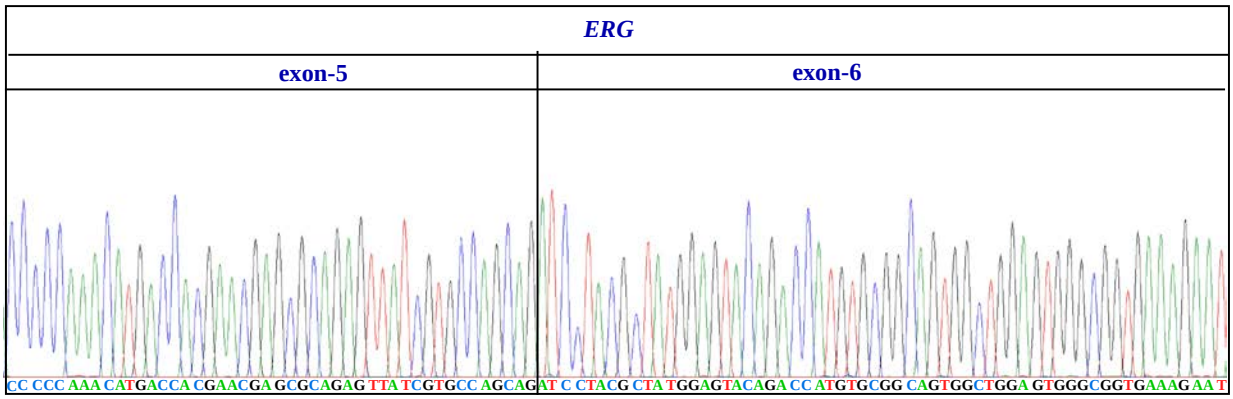
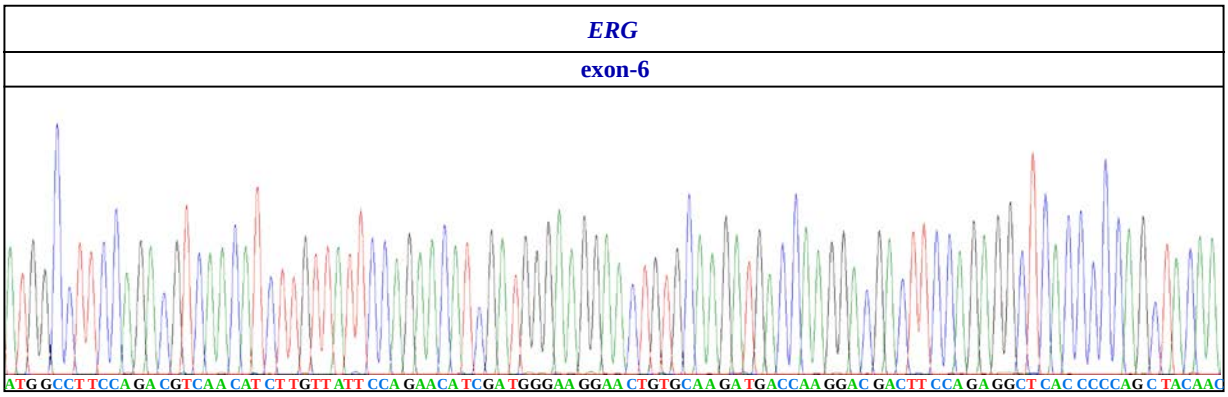


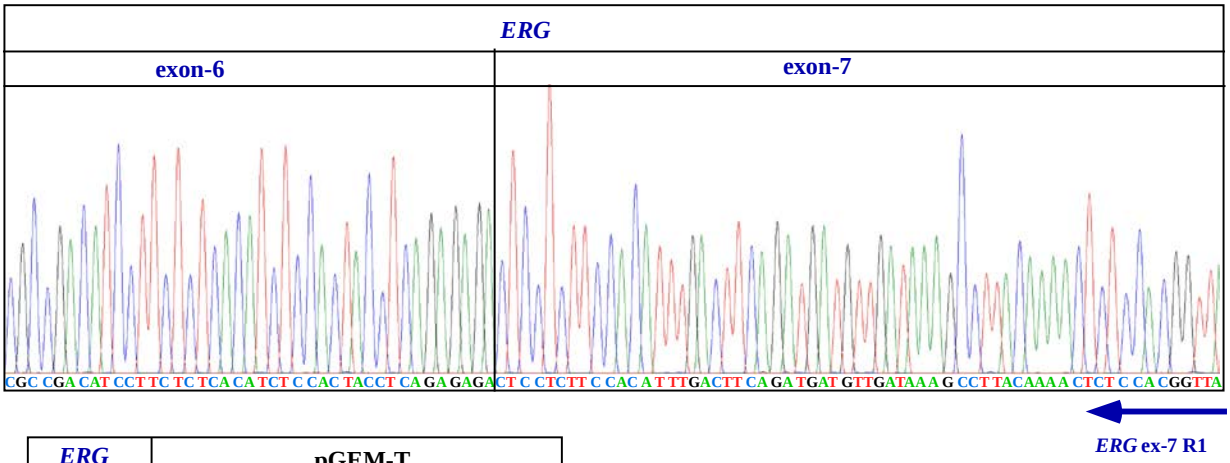
Fig. S2. Sense input RNAs that mimic trans-spliced products failed to induce fusion transcripts.

A. Schematics of the designed input RNAs consisting of *TMPRSS2* (uc002yzj.3) and *ERG* (uc021wjd.1) exons. Upper panels: antisense-2 and sense-2 input RNA consisting of *TMPRSS2* exon-1 (78 nt) and *ERG* exon-4 (218 nt). Lower panel: sense-2-long RNA consisting of *TMPRSS2* exon-1 (78 nt) with *ERG* exons-4, -5, -6 and partial -7 (618 nt). ts: transcriptional stop "TTTTTT" for the U6 promoter. **B.** RT-PCR detection of induced fusion transcript (top gel) and input RNA (bottom gel). For induced fusion transcript, RT-PCR was performed using a forward primer on *TMPRSS2* exon-1 paired with a reverse primer on the 3' end of *ERG* exon-7 (see Fig. S1A), thereby specifically amplifying the endogenous fusion RNA transcribed from the *TMPRSS2-ERG* fusion gene, but not from transfected plasmids or input RNAs because both lack the 3' end of *ERG* exon-7. Control vector lacking input RNA sequence (lane 4), DHT treatment only (lane 5), and PCR reactions lacking cDNA (lane 7) were used as negative controls. Antisense RNA is capable of inducing the fusion transcript (lane 3), whereas sense RNAs failed to induce fusion transcripts regardless of their lengths (lanes 1 and 2).


RNA fusion junction
pGEM-T
TMPRSS2
exon-1
ERG
exon-4
TMPRSS2 ex-1 F1
ERG
exon-4
ERG
exon-4
ERG
exon-5


ERG
exon-5
exon-6
CC CCC AAA CATGACCA CGAACGA GCGCAGAG TTA TCGTGCC AGCAG
AT C CTACG CTA TGGAGTACAGA CC ATGTGCGG CAGTGGCTGGA GTGGGCGGTGAAAG AA T


ERG
exon-6
ATG GCCT TCCA GA CGTCAA CAT CT TGTT ATT CCA GAACA TCGA TGGGAA GGAA CTGTGCAA GA TGACCAA GGAC GACTT CCA GA GGCT CAC CCCCAG C TACAAC


ERG
exon-6
exon-7
CGC CGA CAT CCT TC TC TCA CA TCT C CAC TACCT CA GA GAGA CT C CTCTT C CAC A T TTGACTT CA GA TGAT GTTGATAAA G CCT TACAAAA CTCT C CA CGGTTA
ERG ex-7 R1

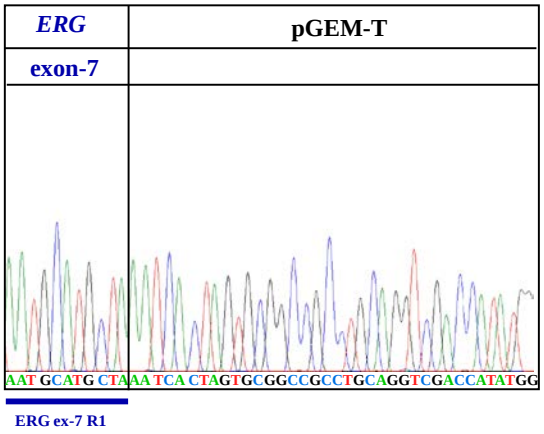

ERG
pGEM-T
exon-7
AAT GCATG CTA AA TCA CTAGTGCGGCCGCCTGCAGGTCGACCATATGG
ERG ex-7 R1

Fig. S3. Sanger sequencing confirmation of the induced *TMPRSS2-ERG* fusion transcript in LNCaP cells mediated by antisense-2 input RNA.

Induced fusion transcripts were converted to cDNA using oligo dT primers and PCR was performed using primers targeting *TMPRSS2* exon-1 and *ERG* exon-7 as described in Fig. S1. The RT-PCR assay relies on the primer targeting *ERG* exon-7 to specifically amplify the induced fusion transcript but not the input RNA, which lacks *ERG* exon-7 sequence. The locations of the primers used to amplify the PCR product are shown as the green and blue arrows for *TMPRSS2* exon-1 and *ERG* exon-7, respectively. The RNA fusion junction between *TMPRSS2* exon-1 and *ERG* exon-4 is indicated by the black arrow. The pGEM-T plasmid was used for cloning the cDNA inserts. Sanger sequencing chromatograms confirmed that the induced fusion transcript contains *TMPRSS2* exon-1 fused to *ERG* exons-4, -5, -6, and -7, and that these exons were joined by splicing at the annotated splice sites as would be expected of mature endogenous *TMPRSS2-ERG* fusion mRNA. This result confirms that expression of an input RNA can lead to specific fusion transcript in human cells.

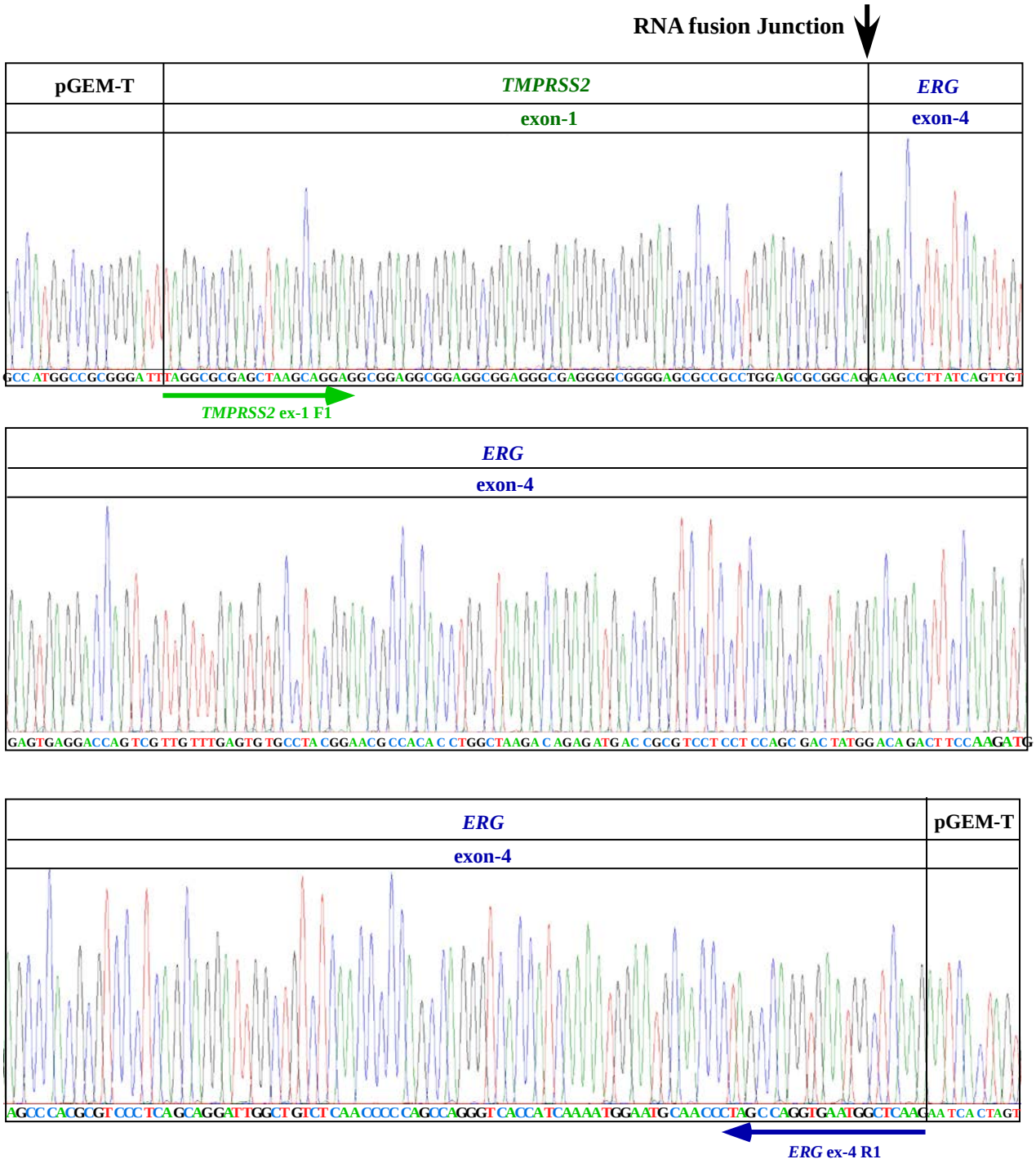


Fig. S4. Sanger sequencing confirmation of the induced *TMPRSS2-ERG* fusion transcripts mediated by antisense-5 input RNA and amplified using an optimized primer pair.

Induced fusion transcripts were converted to cDNA using oligo dT primers. RT-PCR was performed with primers targeting *TMPRSS2* exon-1 and *ERG* exon-4, which led to higher

amplification efficiency. This RT-PCR assay relies on the primer targeting *TMPRSS2* exon-1 to specifically amplify the induced fusion transcript but not the input RNA, which contains the mutated region described in Fig S1B. The locations of the primers are shown as green and blue arrows for *TMPRSS2* exon-1 and *ERG* exon-4, respectively. The RNA fusion junction between *TMPRSS2* exon-1 and *ERG* exon-4 is indicated by the black arrow. The pGEM-T plasmid was used for cloning the cDNA inserts. Sanger sequencing chromatograms confirmed that the induced fusion transcript contained *TMPRSS2* exon-1 fused to *ERG* exon-4, as would be expected of mature endogenous *TMPRSS2-ERG* fusion mRNA. We developed this primer pair to optimize the efficiency of detecting low levels of induced endogenous *TMPRSS2-ERG* fusion mRNA, and these primers were used for most of the RT-PCR assays.

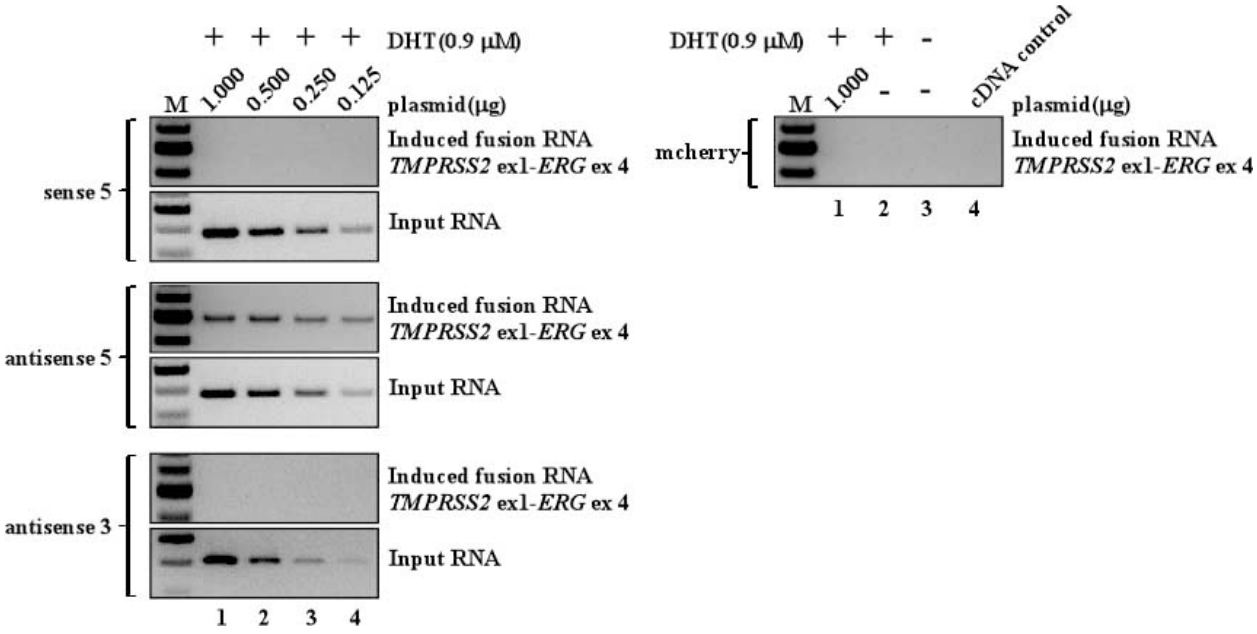


Fig. S5. Antisense-5 but not its corresponding sense-5 induces *TMPRSS2-ERG* gene-fusion.

Different amount of plasmids (1.000μg, 0.500μg, 0.250μg and 0.125μg) expressing either sense-5 or antisense-5 were transfected to produce different amount of corresponding input RNAs in LNCaP cells. To maintain transfection efficiency, mCherry control plasmid was added to each transfection to make the final amount of plasmid to 1.0 μg. Induction was done for three days in the presence of 0.9 μM DHT. RT-PCR was performed to detect the levels of both induced *TMPRSS2-ERG* transcript and input RNA. RT-PCR results showed that even very low amount of antisense-5 input RNA is sufficient to induce *TMPRSS2-ERG* gene fusion (middle panel, lane 4) whereas very high level of sense-5 input RNA fails to induce *TMPRSS2-ERG* gene fusion (upper panel, lane 1). Thus, it is not the amount of the input RNA but the orientation of the input RNA (antisense vs. sense) that is important for fusion induction. Antisense-3, which is unable to induce fusion, was used as a negative control. mCherry alone was also transfected as a negative control. Additional controls (right panel): a vector expressing mCherry lacking input RNA sequence (lane 1), DHT treatment only (lane 2), and PCR reaction lacking cDNA (lane 4).

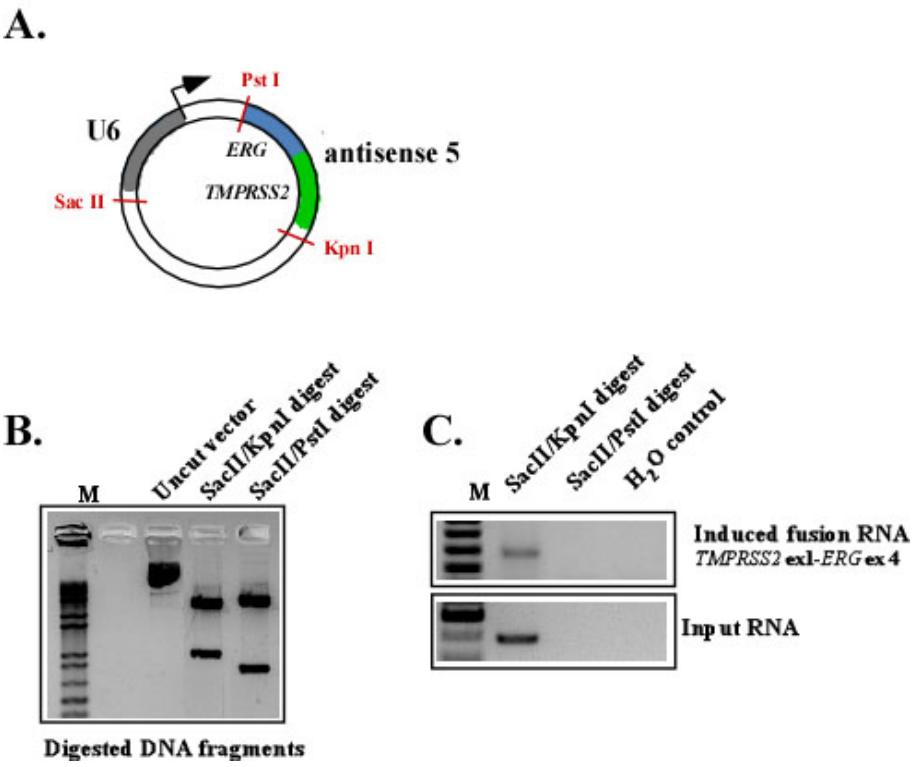


Fig. S6. Expressed input RNA, but not the DNA plasmid, induces fusion transcript.

A. Schematic of the expression plasmid containing a U6 promoter and antisense-5 sequence. Unique restriction sites are shown in red. The arrow indicates the transcriptional start. The plasmid was digested by SacII/KpnI, leaving U6 with antisense-5, or with SacII/PstI, separating U6 from antisense-5. **B.** Confirmation of complete digestion was visualized by agarose gel. Uncut plasmid vector was used as control. **C.** Induction of *TMPRSS2-ERG* fusion after transfection of digested plasmid fragments. The *TMPRSS2-ERG* fusion transcript was induced only when the U6 promoter was attached to antisense-5 sequence, indicating that it is the input RNA expressed by the U6 promoter, not the transfected DNA plasmid per se, that leads to fusion induction.

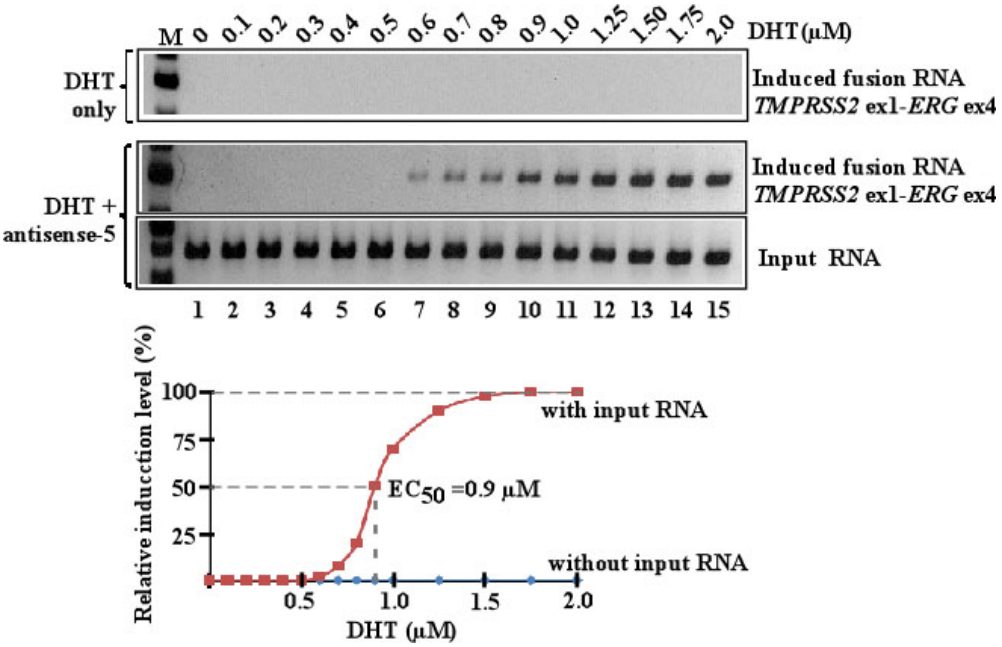


Fig. S7. DHT dosage effect in facilitating RNA-mediated gene fusion as assayed by one round RT-PCR.

To determine the DHT concentrations that facilitate RNA-mediated gene fusion detectable by one round RT-PCR, we generated a dose-response curve using LNCaP cells transfected with antisense-5 and treated with different DHT concentrations for 3 days. Relative RT-PCR intensities of induced fusion transcript bands as generated by one round RT-PCR were plotted against DHT concentrations. In the presence of antisense-5, fusion induction is detectable at ~0.6 μM DHT, reaches its maximum at 1.5 μM DHT, with the effective concentration that yields 50% of maximal induction (EC_{50}) at ~0.9 μM. In the absence of antisense-5, no induction was observed in the entire DHT range tested up to 2.0 μM (upper panel), indicating that DHT alone is ineffective in inducing *TMPRSS2-ERG* fusion. To streamline our experimental procedures, we used 0.9 μM DHT (the EC_{50} concentration) for 3 days followed by one round of RT-PCR as our standard assay in the main text. However, it is important to note that all key experiments were also performed under physiologically relevant DHT concentrations (<100 nM), and in those cases three-rounds of nested PCR were used to reveal the lowest amount of DHT required. As shown in Fig. 1D and Fig. 6A in the main text, gene fusion events induced by antisense-5 occurred at physiologically relevant DHT concentrations as low as 20 nM, and gene fusion events induced by endogenous *AZI1* mRNA occurred at as low as 40 nM DHT.

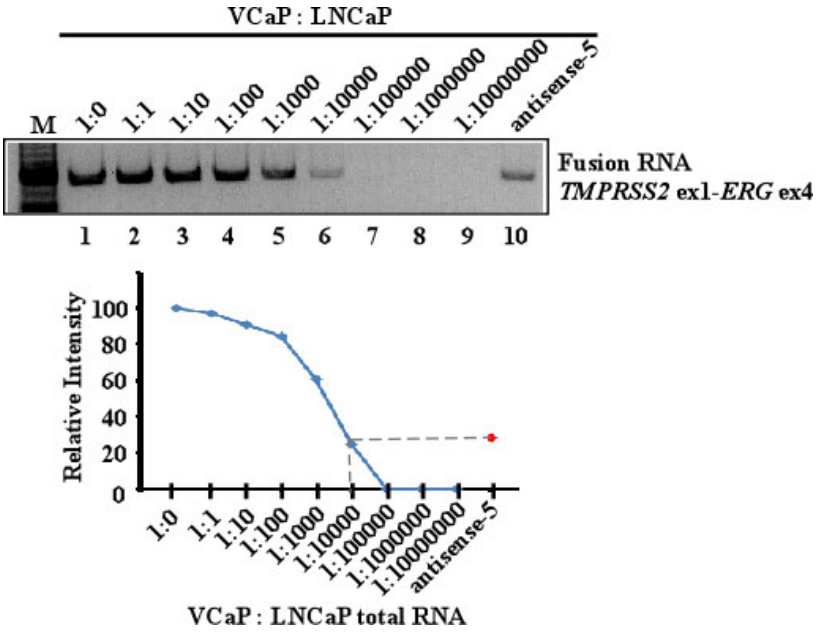


Fig. S8. Estimation of the percentage of LNCaP cells transfected with antisense-5 that carry the induced *TMPRSS2-ERG* fusion.

To estimate the percentage of LNCaP cells that underwent *TMPRSS2-ERG* gene fusion mediated by antisense-5 at 0.9 µM DHT (the EC_{50}), we mixed various amounts of total RNA from VCaP cells that express *TMPRSS2-ERG* (Mertz et al., 2007; Teles Alves et al., 2013) with the total RNA from untransfected LNCaP cells in ratios of 1:1 up to $1:10^7$. A standard dilution curve was determined using the relative RT-PCR intensity of the *TMPRSS2-ERG* fusion transcript. The relative intensity of induction by antisense-5 in LNCaP cells, as assayed under the same RT-PCR conditions, is shown as a red dot that corresponds to an equivalent dilution of $1:10^3$ to $1:10^4$, suggesting that approximately 1 in 10^3 or 10^4 LNCaP cells is positive of *TMPRSS2-ERG* fusion.

	TMPRSS2 (-strand)	*ERG* (-strand)	Tm
STEMS			
A	chr21:41508022-41508035	chr21:38445526-38445539	44°C
B	chr21:41508030-41508040	chr21:38463727-38463737	40°C
C	chr21:41507658-41507669	chr21:38457001-38457012	40°C
D	chr21:41507876-41507888	chr21:38451230-38451242	44°C
E	chr21:41499266-41499277	chr21:38481641-38481652	30°C
F	chr21:41507207-41507213	chr21:38482731-38482737	24°C
G	chr21:41503846-41503852	chr21:38455751-38455757	16°C
EXONS			
TMPRSS2 exon-1	chr21:41508081-41508158		
TMPRSS2 exon-2	chr21:41498119-41498189		
ERG exon-3		chr21:38575662-38575747	
ERG exon-4		chr21:38445404-38445621	

Fig. S9. Genome coordinates and Tm for the putative genomic stems described in Fig. 2C and 2D of the main text.

Genome coordinates are based on the GRCh38/hg38 version. Coordinates of the exons (*TMPRSS2* exon-1 and -2; *ERG* exon-3 and -4) are also shown.

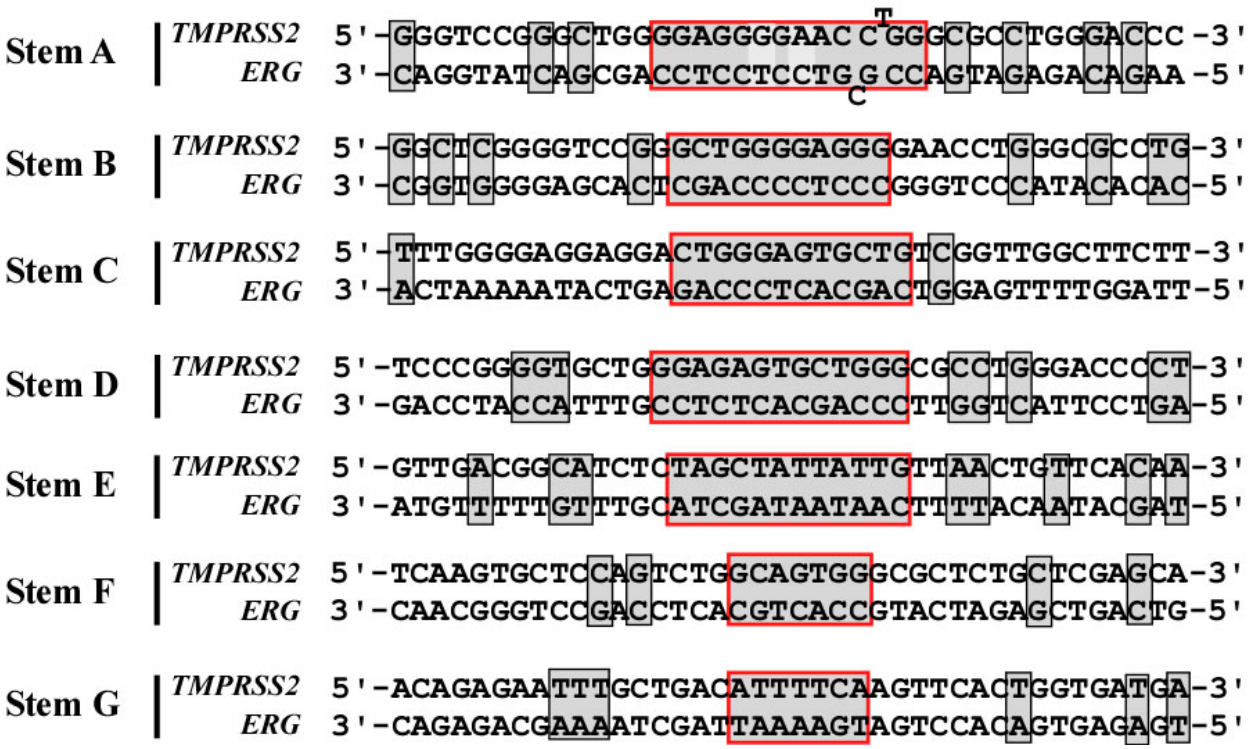


Fig. S10. Complementary base pairing of the genomic DNA sequences comprising the stems and their flanking sequences.

The sense strand of genomic sequences from *TMPRSS2* and *ERG* that form the putative stems (in red boxes) with various degrees of stability as identified by BLAST analyses. Regions showing perfect complementary base pairing (A-T and G-C) are shaded in gray.

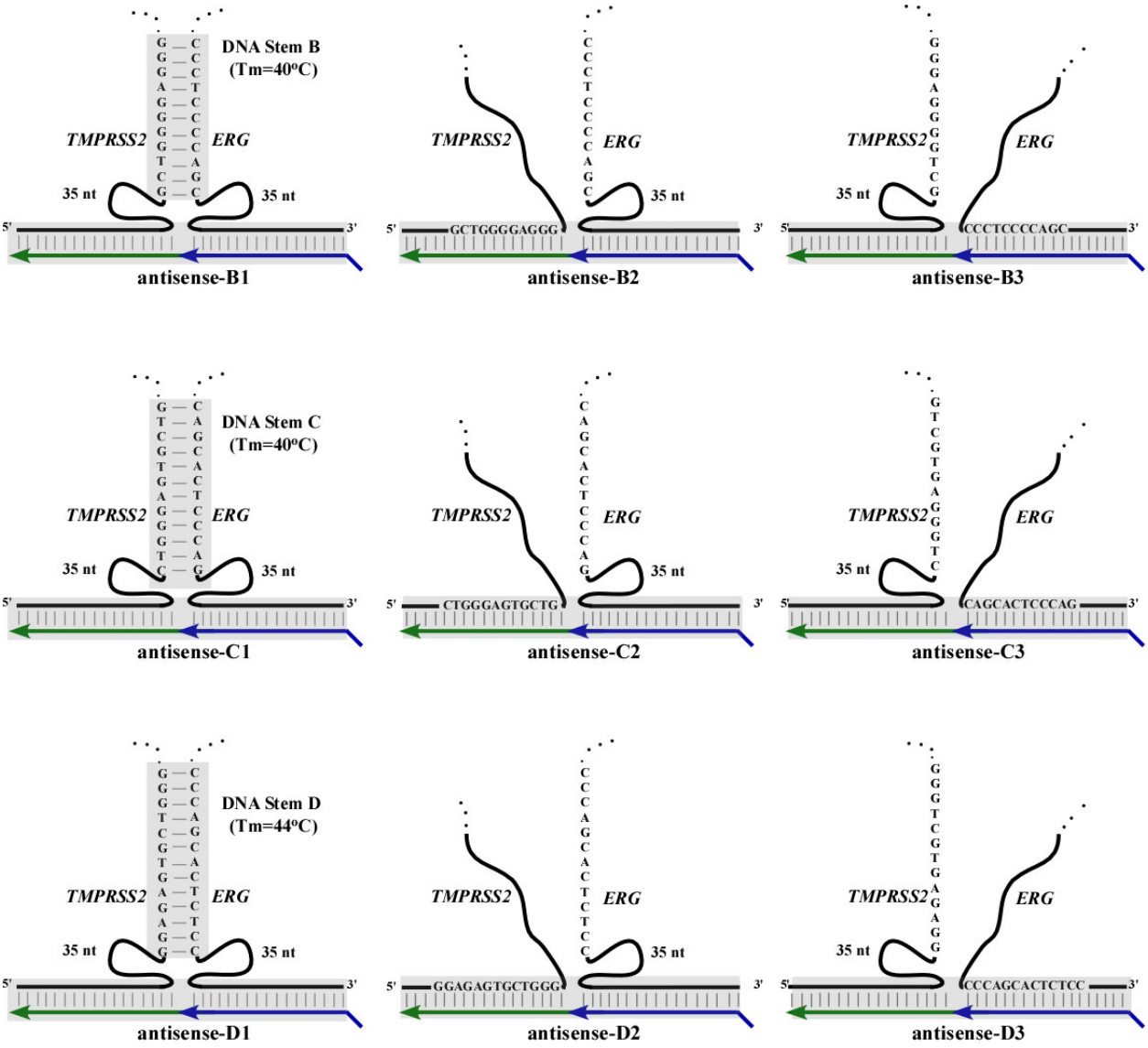


Fig. S11. Antisense input RNAs designed to disrupt the genomic stems B, C, and D.

Stem B (upper row). Stem C (middle row). Stem D (lower row). Left column: Illustrations show the three-way junction formed by genomic DNA and corresponding antisense input RNA (antisense-B1, C1, and D1). Middle column: Disruption of stems via the *TMPRSS2* side using tailor-made input RNAs (antisense-B2, C2, and D2) that directly hybridize to the *TMPRSS2* stem sequence. Right column: Disruption of stems via the *ERG* side using tailor-made input RNAs (antisense-B3, C3, and D3) that directly hybridize to the *ERG* stem sequence. Shaded regions indicate base pairing. The *TMPRSS2* portion of the input RNA is shown in green and the *ERG* portion is in blue.

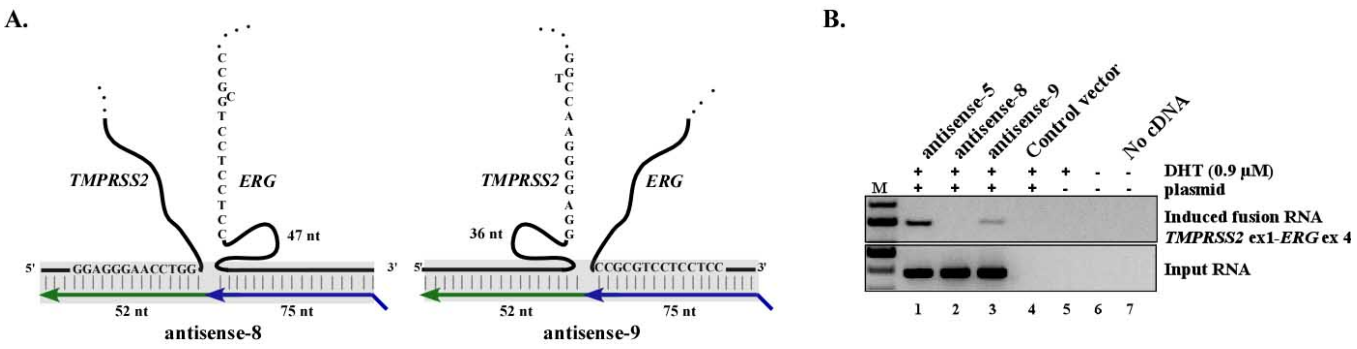


Fig. S12. Antisense input RNAs designed to disrupt the genomic stems A.

A. Disruption of the putative genomic stem using tailor-made input RNAs that directly hybridize to one side of the stem A. **B.** RT-PCR assays of fusion transcripts induced by the input RNAs illustrated in A. Interfering with three-way junction formation eliminated (lane 2) or significantly reduced the induction (lane 3). Controls include: vector lacking input RNA sequence (lane 4), DHT treatment only (lane 5), and PCR reaction lacking cDNA (lane 7).

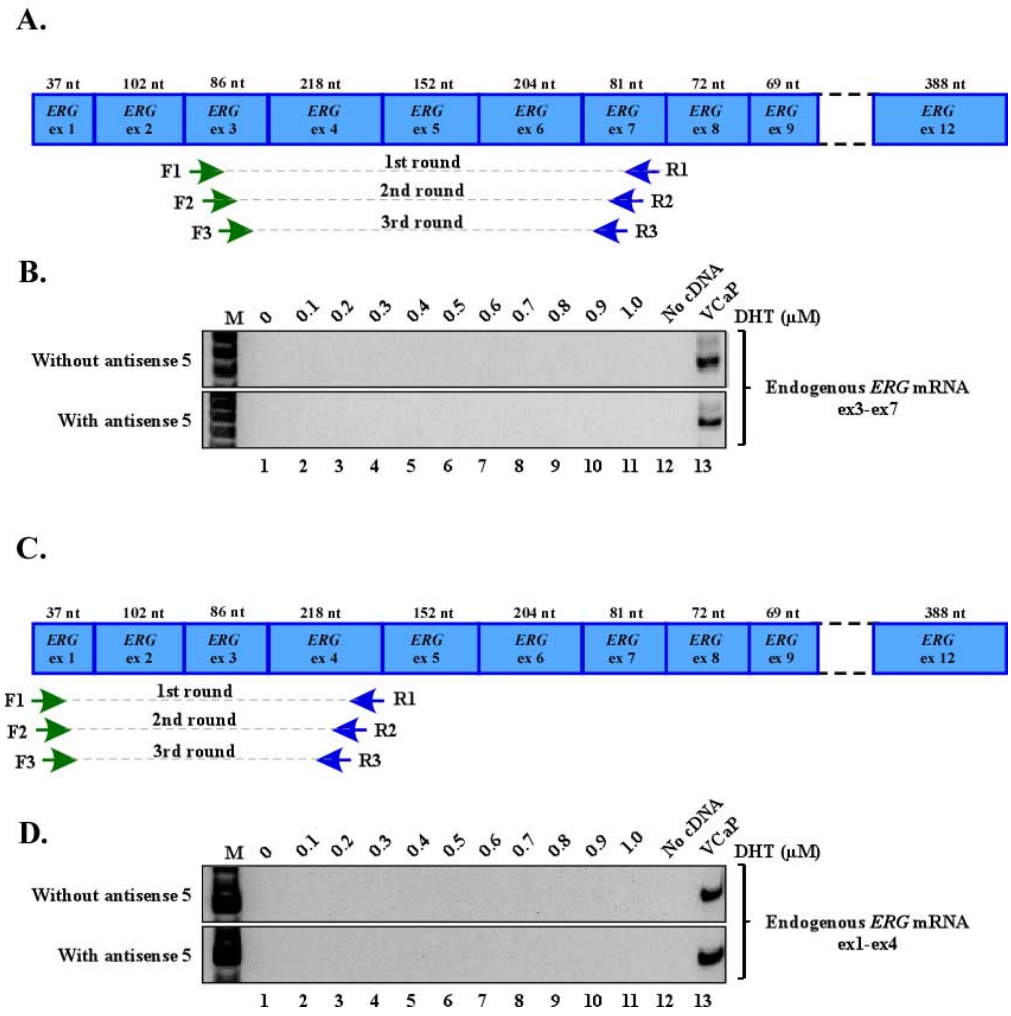


Fig. S13. Endogenous *ERG* mRNA is not detected in LNCaP cells by three-rounds of nested RT-PCR.

A. Nested primer pairs used to detect low levels of endogenous *ERG* mRNA from LNCaP cells. Forward primers for *ERG* exon-3 are shown in green whereas reverse primers for *ERG* exon-7 are shown in blue (primer sequences are listed in the Materials and Methods). **B.** Endogenous *ERG* mRNA was not detected in LNCaP cells in the presence or absence of DHT or antisense-5 (lane 1 to 11) by three-rounds of nested RT-PCR. cDNA prepared from VCaP cells which express *ERG* mRNA was used as positive control (lane 13). **C** and **D.** Similarly, using forward primers on *ERG* exon-1 and reverse primers on *ERG* exon-4 also failed to detect endogenous *ERG* mRNA by three-rounds of nested RT-PCR (lane 1 to 11, primer sequences are listed in the Materials and Methods). These primer sets were chosen because they specifically amply endogenous *ERG* mRNA but not *TMPRSS2-ERG* fusion transcript that contains *ERG* exon-3 to exon-12.

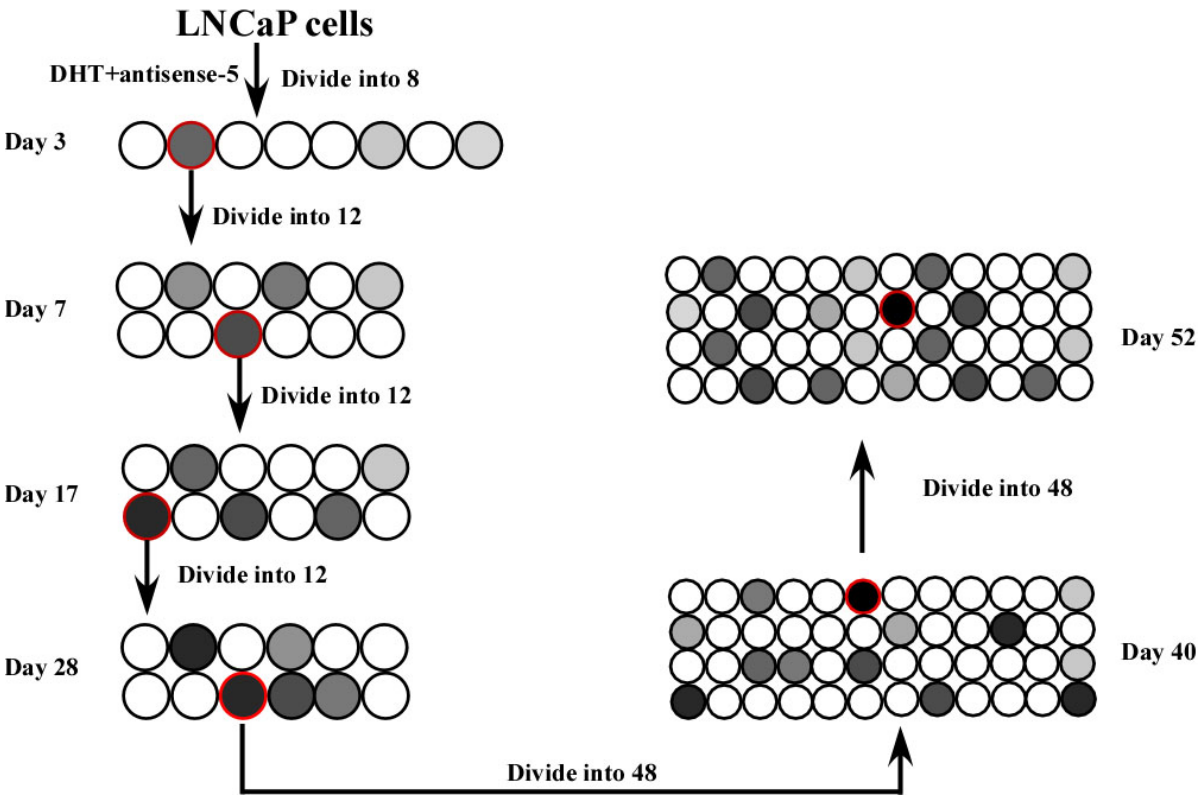


Fig. S14. Procedure used to propagate and enrich the induced LNCaP population for 52 days in the absence of DHT.

LNCaP cells were transfected on day 1 with antisense-5 and incubated for 3 days with DHT. Cells were then divided into sub-populations. Half of the cells from each sub-population were harvested for RT-PCR assays to detect the induced fusion transcript whereas the other half was used to propagate the population. The sub-population containing the highest intensity of induced fusion transcripts (circled in red) was selected for continuous propagation for the next round of division and RT-PCR assays. The process was continued for 52 days. RT-PCR was performed to detect both induced fusion transcripts and antisense input RNA during each division as shown in Fig. 3B.

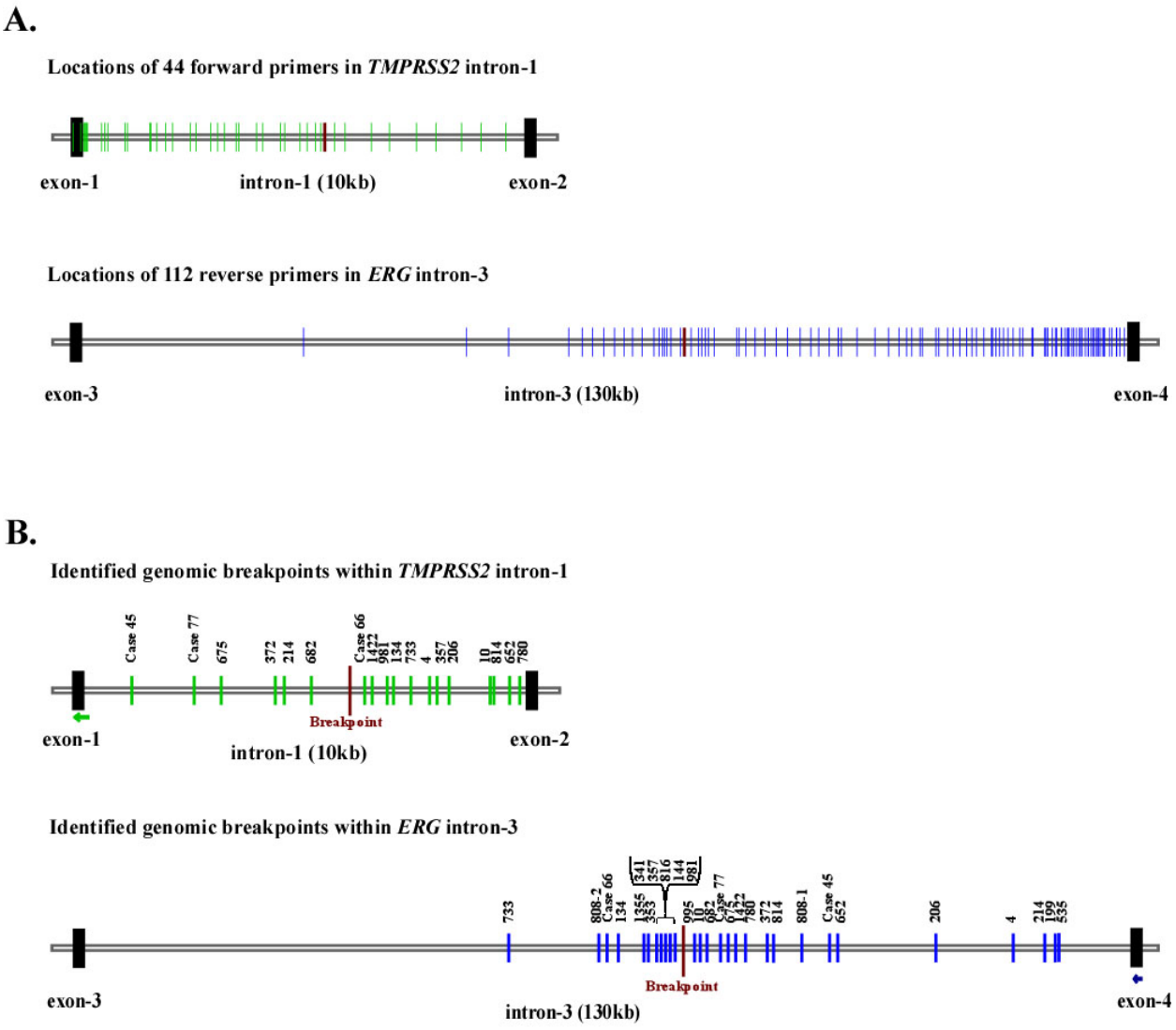
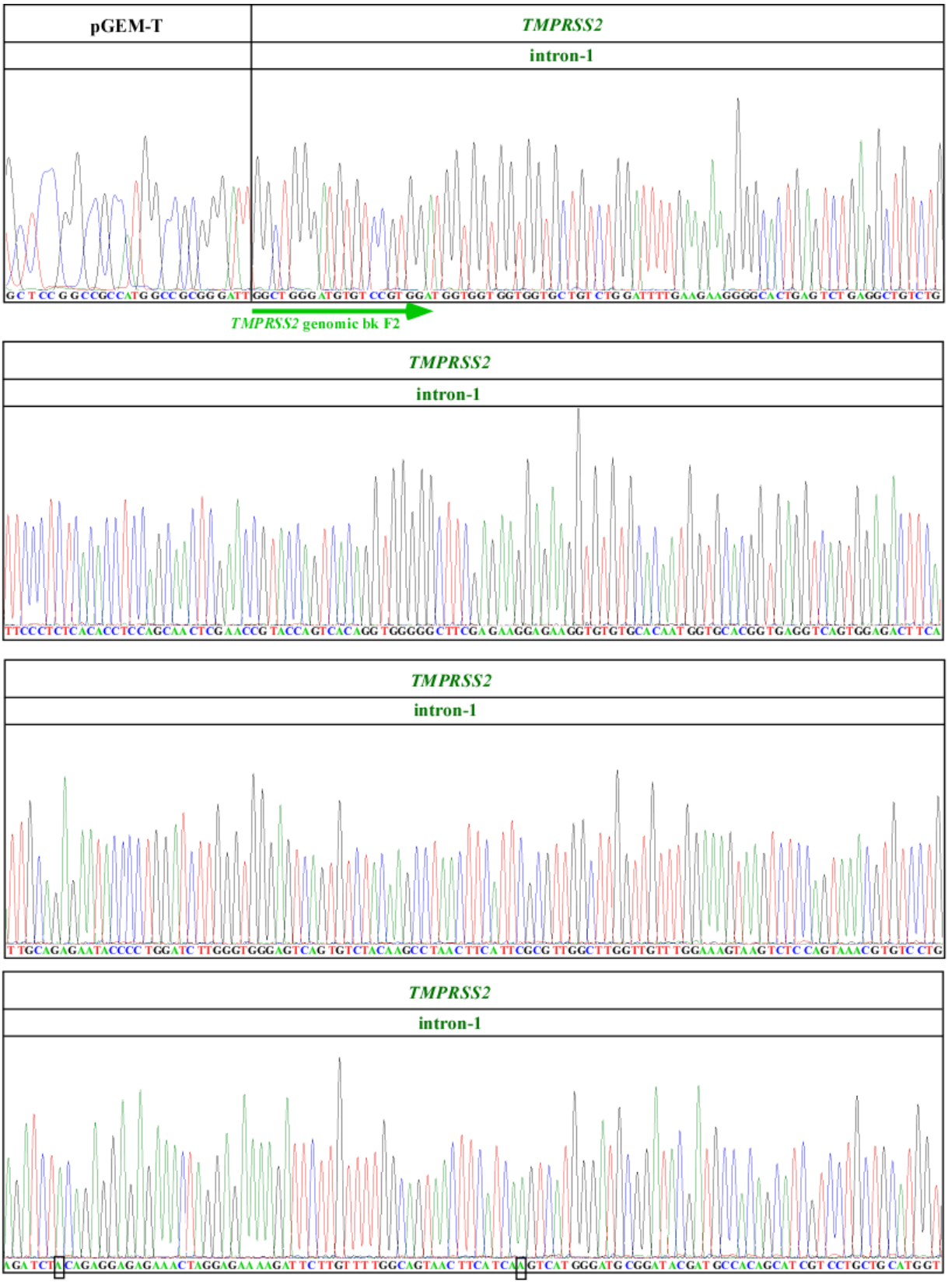
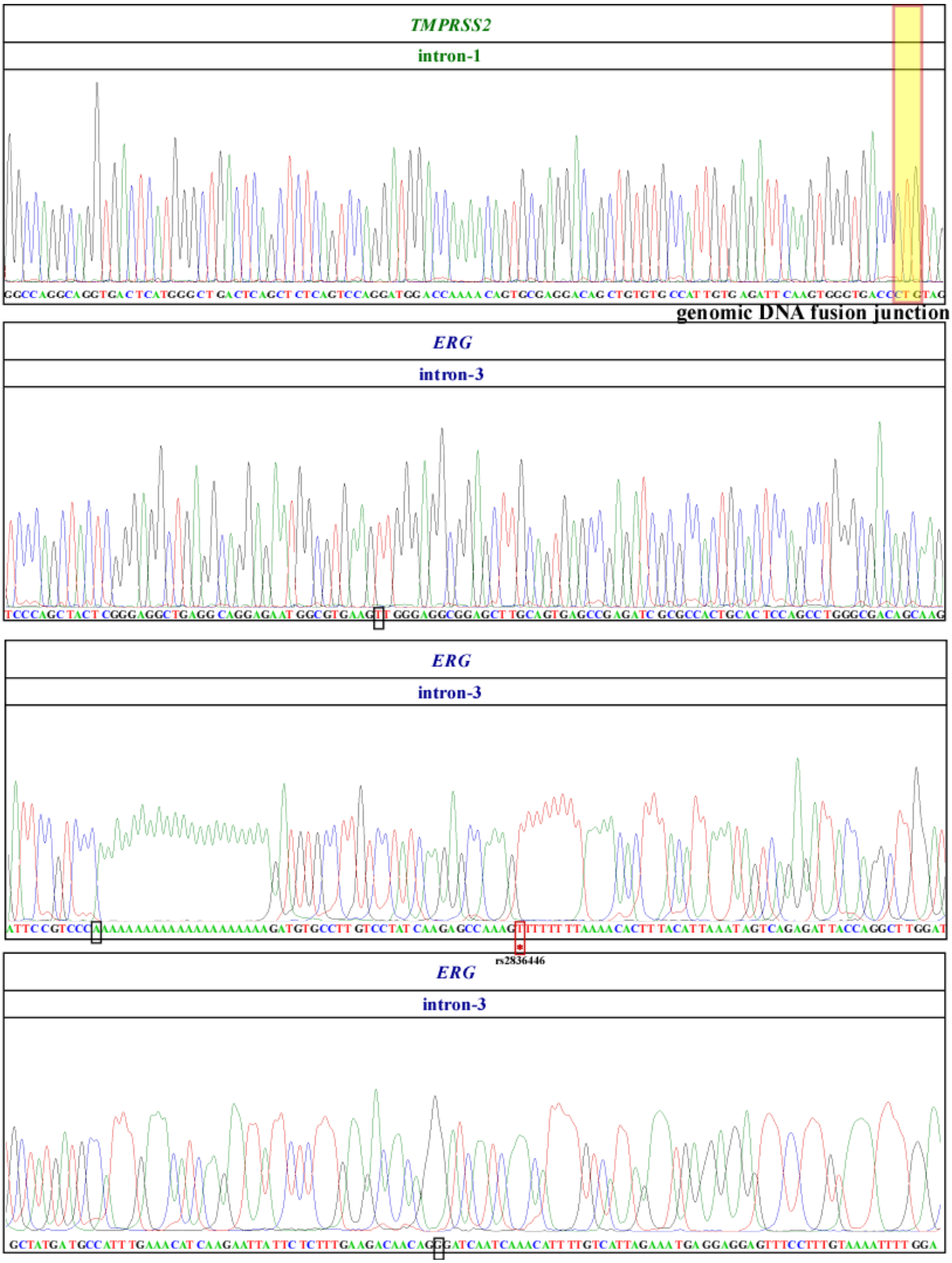


Fig. S15. Locations targeted by primers for mapping the genomic breakpoint and locations of identified *TMPRSS2-ERG* genomic breakpoints.

A. The locations of 44 forward primers (green) spacing across *TMPRSS2* intron-1 (~10 kb) and 112 reverse primers (blue) spacing across *ERG* intron-3 (~130 kb). *ERG* intron-3 is significantly larger than that of *TMPRSS2* intron-1. Each vertical line represents a target location by a primer or a set of nested primers. The primers targeting *ERG* intron-3 are designed to concentrate in the region near exon-4 and in the hot spot regions previously identified from prostate cancer patients. The specific primers that amplify the genomic breakpoint shown in Fig. 3D in the main text are labeled as red. **B.** Locations of identified *TMPRSS2-ERG* genomic breakpoints. Locations of previously identified *TMPRSS2-ERG* genomic breakpoints from prostate cancer patients (Weier et al.) are shown in green in *TMPRSS2* intron-1 and blue in *ERG* intron-3. The number on top of each location is the ID number of a patient. The genomic breakpoint identified in our study is shown in red. Short arrows indicate sites targeted by antisense-5.


pGEM-T
TMPRSS2
intron-1
TMPRSS2 genomic bk F2
TMPRSS2
intron-1
TMPRSS2
intron-1
TMPRSS2
intron-1


TMPRSS2
intron-1
GGCCAGGCAGGTGACTCATGGGCTGACTCAGCTCTCAGTCCAGGATGGACCAAAACAGTGCGAGGACAGCTGTGTGCCATTGTGAGATTCAAGTGGGTGACCCTGTAG
genomic DNA fusion junction
ERG
intron-3
TCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAAGTTGGGAGGCGGAGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGCAAG
ERG
intron-3
ATTCCGTCCCAAAAAAAAAAAAAAAAAAAAGATGTGCCTTGTCCTATCAAGAGCCAAAGTTTTTTTTAAAACACTTTACATTAAATAGTCAGAGATTACCAGGCTTGGAT
rs2836446
ERG
intron-3
GCTATGATGCCATTTGAAACATCAAGAATTATTCTCTTTGAAGACAACAGGGATCAATCAAACATTTTGTCATTAGAAATGAGGAGGAGTTTCCTTTGTAAAATTTTGGA

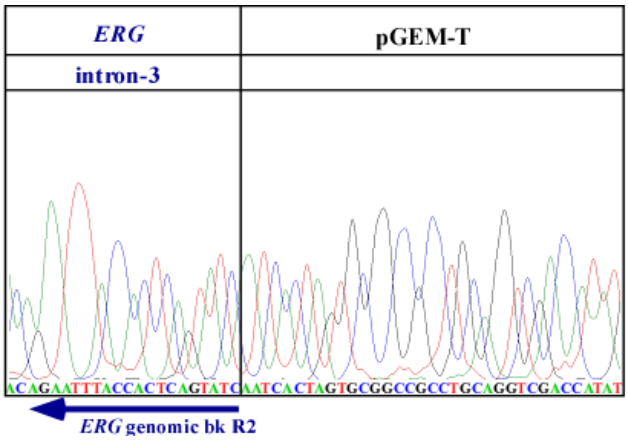


Fig. S16. Sanger sequencing confirmed that the induced *TMPRSS2-ERG* fusion is due to gene fusion resulting from genomic arrangement.

LNCaP cells transfected with antisense-5 were enriched for the *TMPRSS2-ERG* fusion transcript as described in Fig. S14. Genomic DNA PCR was performed on the enriched LNCaP population which yielded a gene fusion band of ~862 bp as shown in Fig. 3D in the main text. Sanger sequencing chromatograms confirmed that the 862 bp band contains ~500 bp of *TMPRSS2* intron-1 joined to ~362 bp of *ERG* intron-3. The locations of the primers used to amplify the PCR product are shown as the green and blue arrows for *TMPRSS2* intron-1 and *ERG* intron-3, respectively under the chromatograms. The genomic breakpoint between *TMPRSS2* intron-1 and *ERG* intron-3 is indicated by the yellow shaded region that contains a three nucleotide 'CTG' microhomology. The pGEM-T plasmid was used for cloning the genomic DNA PCR product. Five mutations shown in black boxes are PCR artifacts because they are absent in other clones that were sequenced in parallel. The mutation denoted by the asterisk (*) in the red box is present in all sequenced clones and is a known Single Nucleotide Polymorphism (rs2836446).

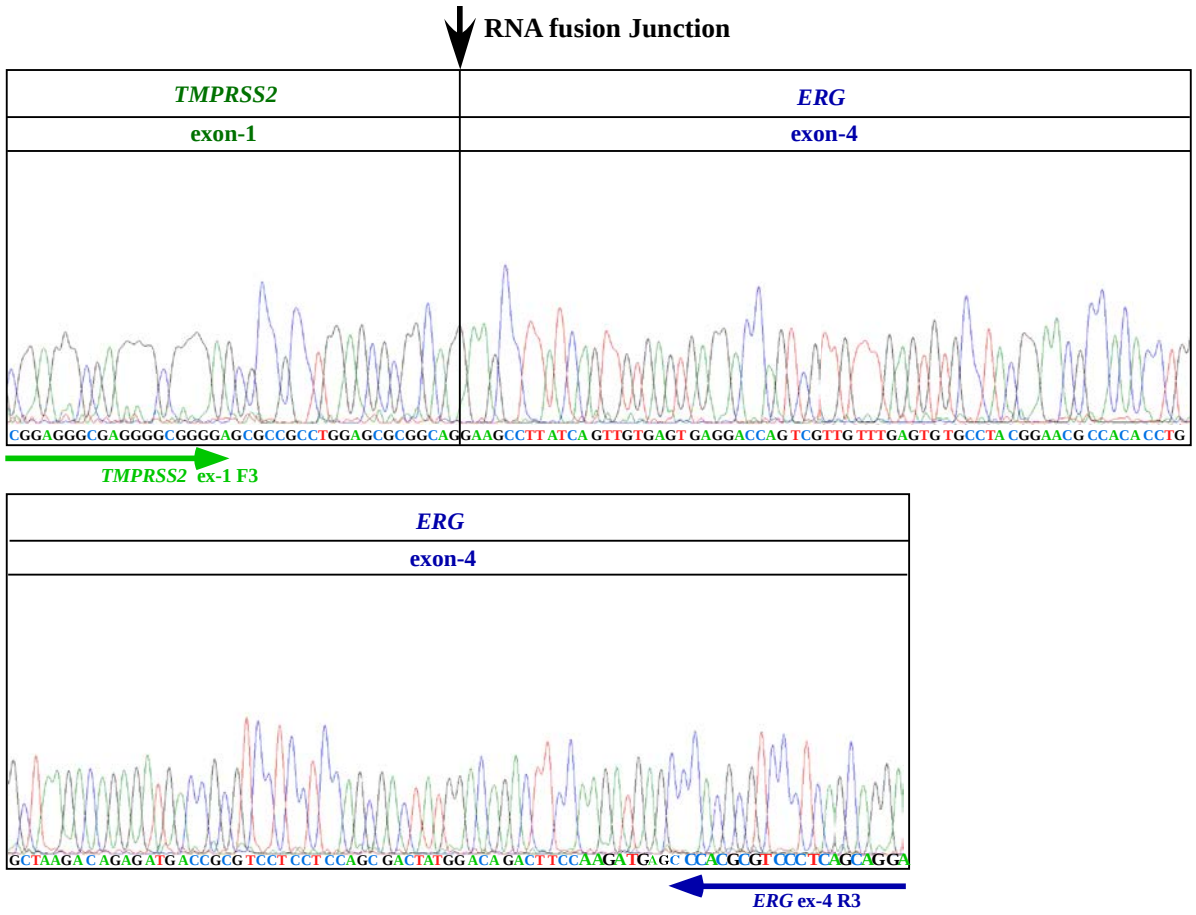


Fig. S17. Sanger sequencing confirmation of the induced *TMPRSS2-ERG* fusion transcript in PNT1A cells mediated by antisense-5 RNA.

Sanger sequencing confirmed that the induced fusion RNA in normal prostate epithelium cells (as shown in Fig. 3F in the main text) is the same fusion transcript containing *TMPRSS2* exon-1 fused to *ERG* exon-4 as would be expected of mature endogenous *TMPRSS2-ERG* fusion mRNA. This indicates that induction of the *TMPRSS2-ERG* fusion transcript by input RNA is permissible in normal prostate epithelial cells prior to malignant transformation. The locations of the primers used to amplify the PCR product are shown as the green and blue arrows for *TMPRSS2* exon-1 and *ERG* exon-4, respectively. The RNA fusion junction between *TMPRSS2* exon-1 and *ERG* exon-4 is indicated by the black arrow.

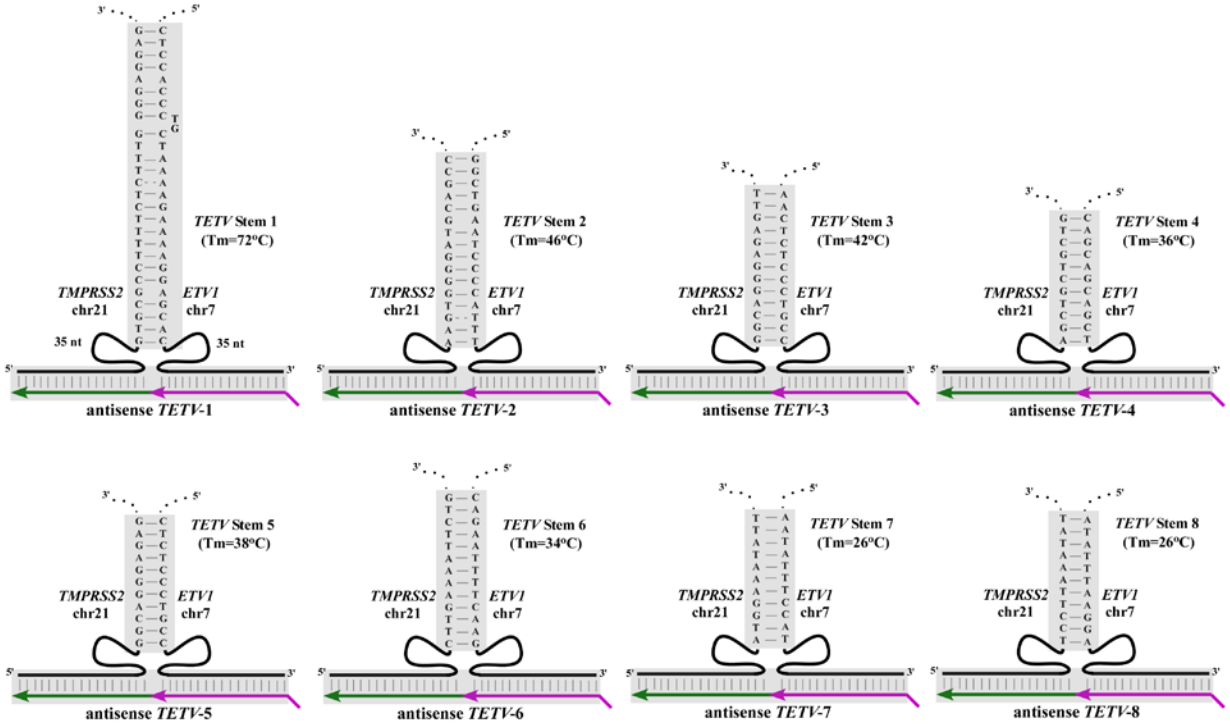
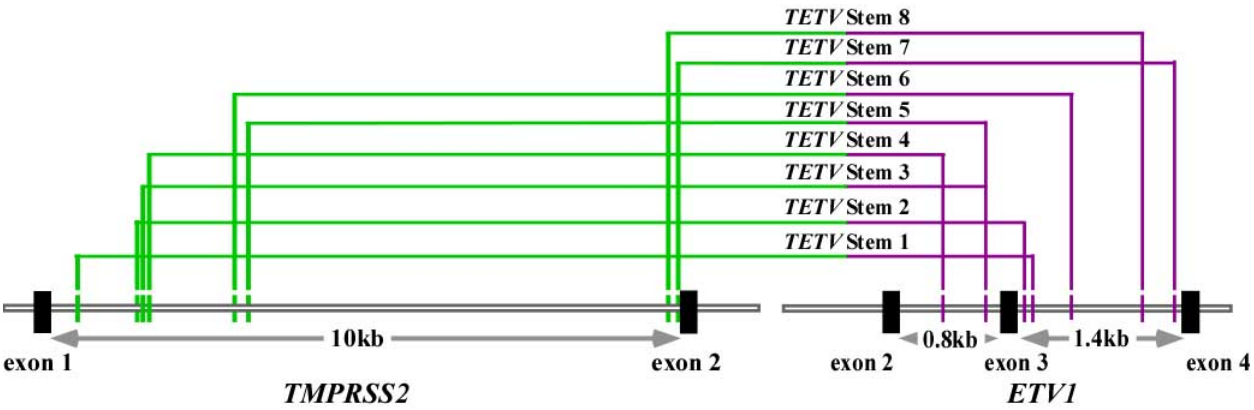


Fig. S18. Schematics of the putative three-way junctions formed between *TMPRSS2* and *ETV1* genomic DNA and the corresponding antisense input RNA.

We used BLAST alignment to identify eight stems that could be formed with varying degrees of stability by the sense genomic *TMPRSS2* sequence paired with the sense genomic *ETV1* sequence. Matching antisense RNAs were then designed to forge the three-way junction with these putative intron stems. Shaded regions indicate base pairing between antisense input RNA (in green and purple) and genomic DNA (in black). *TMPRSS2* is on chromosome 21 whereas *ETV1* is on chromosome 7. For both, the genomic minus strand is the sense strand.



	TMPRSS2 (-strand)	*ETV* (-strand)	Tm
STEMS			
TETV stem 1	chr21:41507672-41507697	chr7:13988013-13988040	72°C
TETV stem 2	chr21:41506698-41506713	chr7:13988054-13988069	46°C
TETV stem 3	chr21:41506649-41506661	chr7:13988373-13988385	42°C
TETV stem 4	chr21:41506584-41506594	chr7:13988732-13988742	36°C
TETV stem 5	chr21:41505067-41505077	chr7:13988373-13988383	38°C
TETV stem 6	chr21:41505211-41505223	chr7:13987620-13987632	34°C
TETV stem 7	chr21:41498215-41498225	chr7:13986643-13986753	26°C
TETV stem 8	chr21:41498359-41498369	chr7:13986021-13987031	26°C
EXONS			
TMPRSS2 exon-1	chr21:41508081-41508158		
TMPRSS2 exon-2	chr21:41498119-41498189		
ETV1 exon-2		chr7:13989090-13989139	
ETV1 exon-3		chr7:13988140-13988173	
ETV1 exon-4		chr7:13986638-13986685	

Fig. S19. Genome coordinates and Tm for the putative *TETV* genomic stems described in S18.

Upper panel: Locations of putative *TETV* stems 1, 2, 3, 4, 5, 6, 7, and 8 identified by BLAST alignment. All stem sequences are located in *TMPRSS2* intron-1 (chromosome 21) paired with sequences in *ETV1* intron-2 or -3 (chromosome 7). Genome coordinates are based on GRCh38/hg38 version. Coordinates of the exons (*TMPRSS2* exons-1 and -2; *ETV1* exons-2, -3, and -4) are also shown.

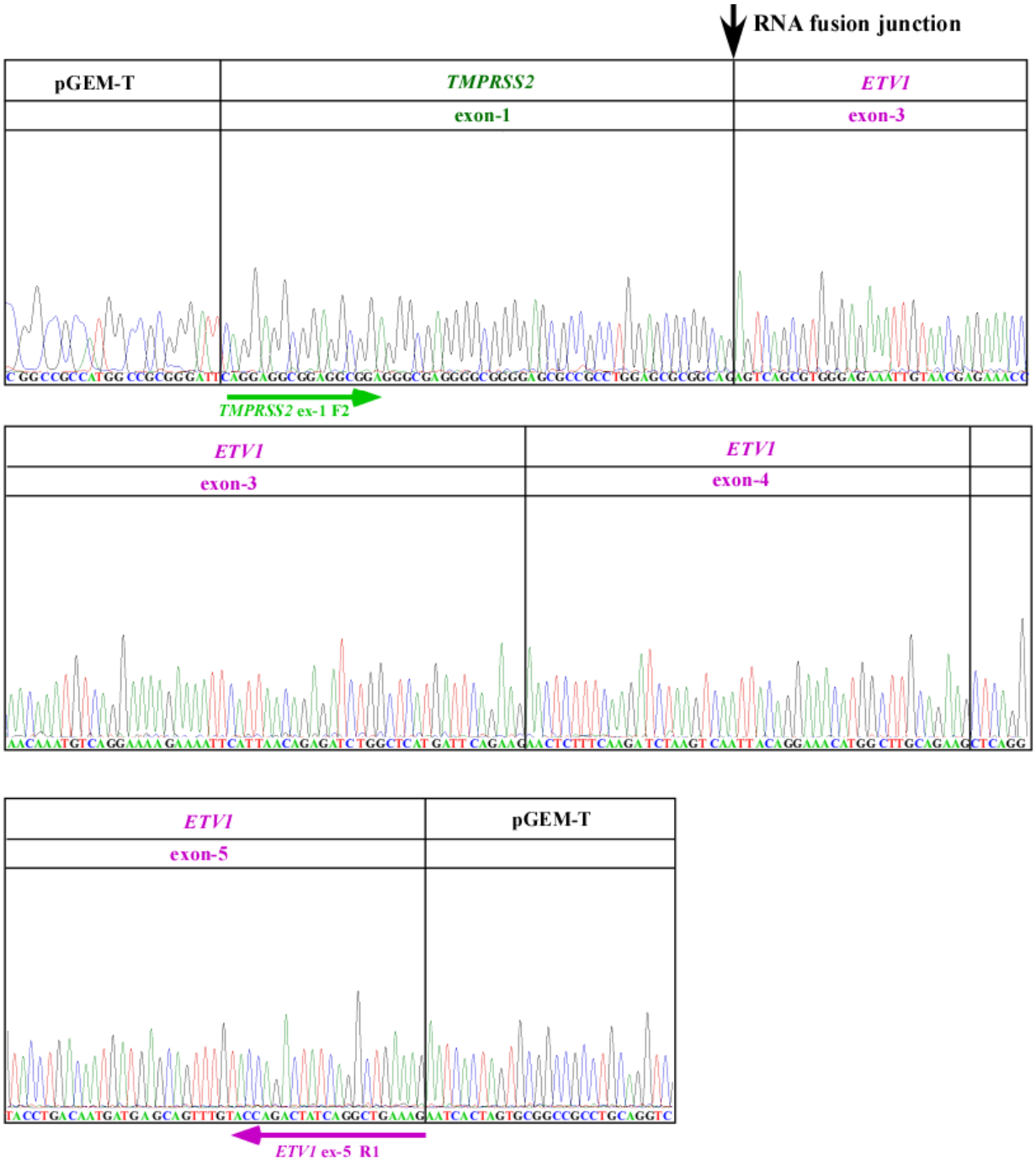


Fig. S20. Sanger sequencing confirmed the induced *TMPRSS2-ETV1* fusion transcripts in LNCaP cells.

Sanger sequencing chromatograms confirmed that the induced RNA contains *TMPRSS2* exon-1 fused to *ETV1* exon-3 as would be expected of mature endogenous *TMPRSS2-ETV1* fusion

mRNA. These results suggest that RNA-mediated gene fusion in mammalian cells is not restricted to *TMPRSS2* and *ERG* but appears to be generally permissible regardless of whether it is an intra-chromosomal (*TMPRSS2-ERG)* or inter-chromosomal (*TMPRSS2-ETV1)* fusion. The induced fusion transcript was converted to cDNA using oligo dT primers and PCR was performed using primers targeting *TMPRSS2* exon-1 and *ETV1* exon-5. The locations of the primers are shown as the green and purple arrow for *TMPRSS2* exon-1 and *ETV1* exon-5, respectively. The RNA fusion junction between *TMPRSS2* exon-1 and *ETV1* exon-3 is indicated by the black arrow.

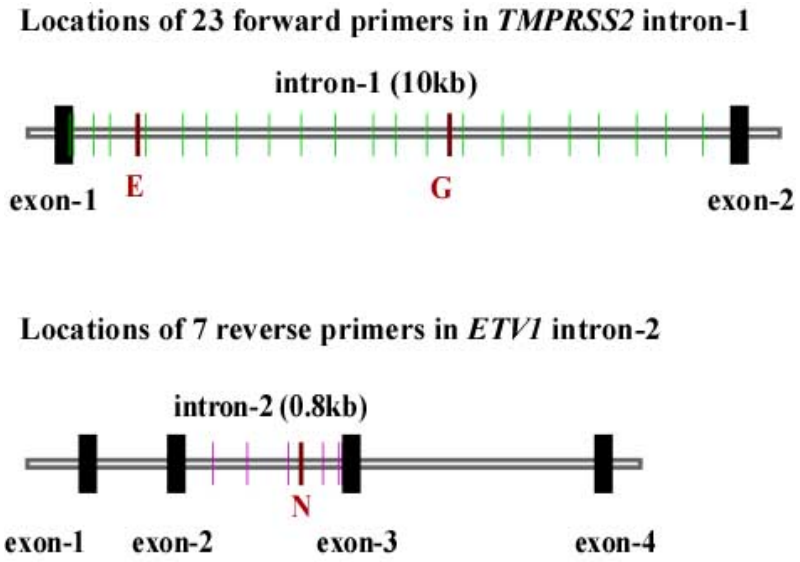
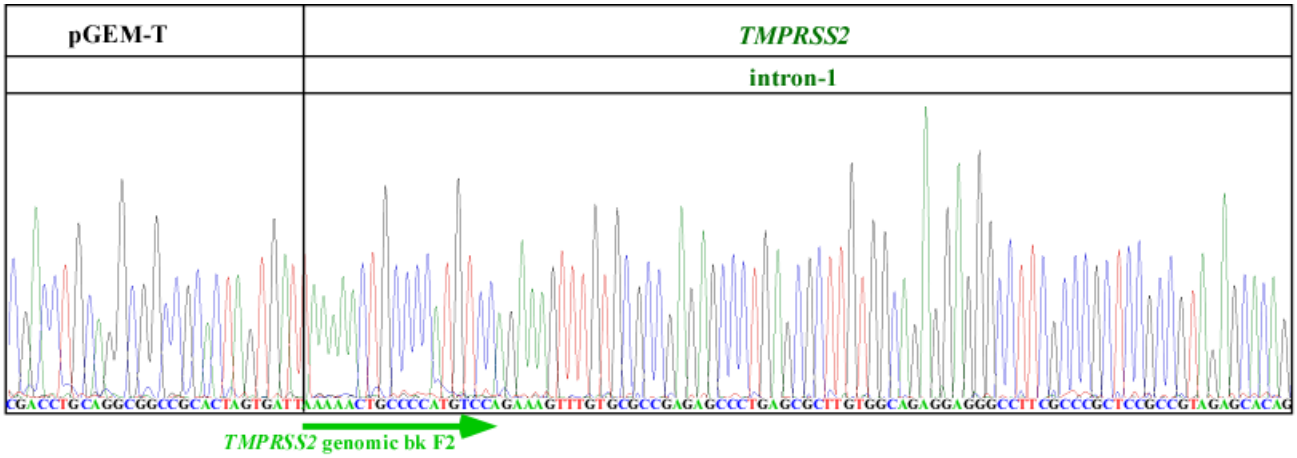
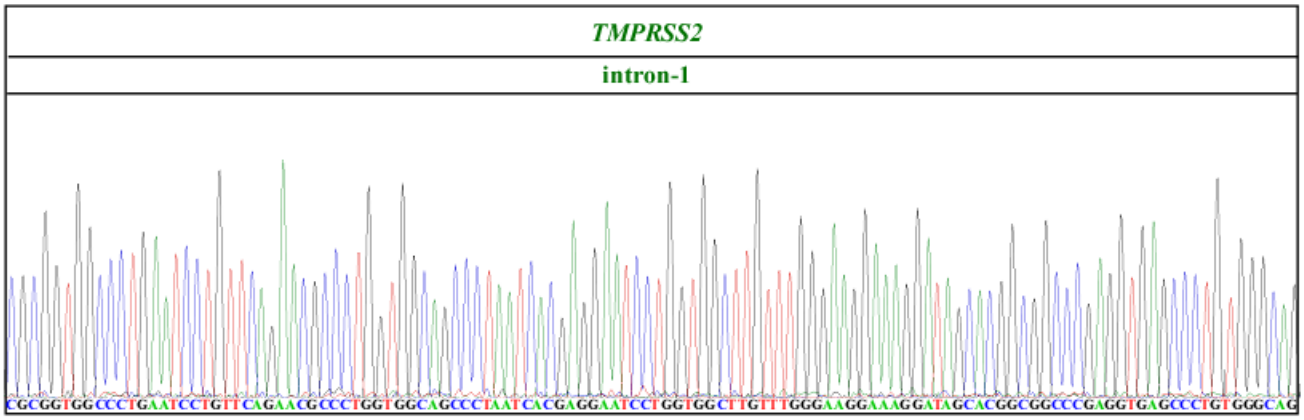
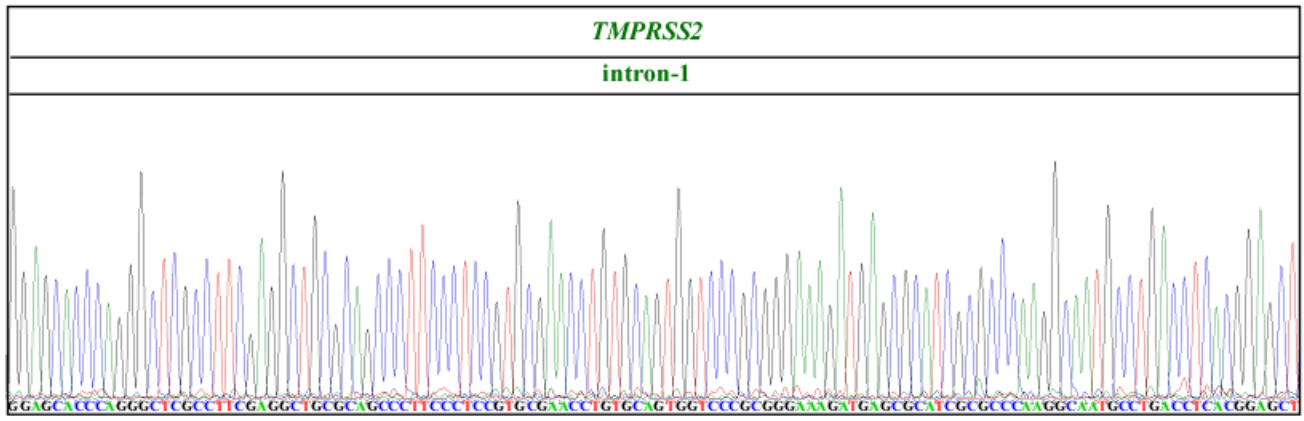


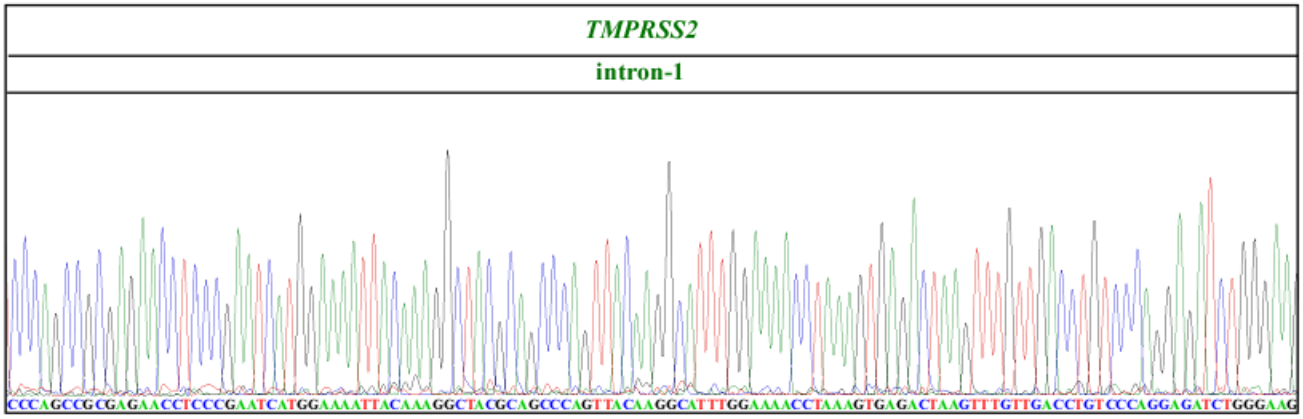
Fig. S21. Locations targeted by primers for mapping the *TMPRSS2-ETV1* genomic breakpoints.

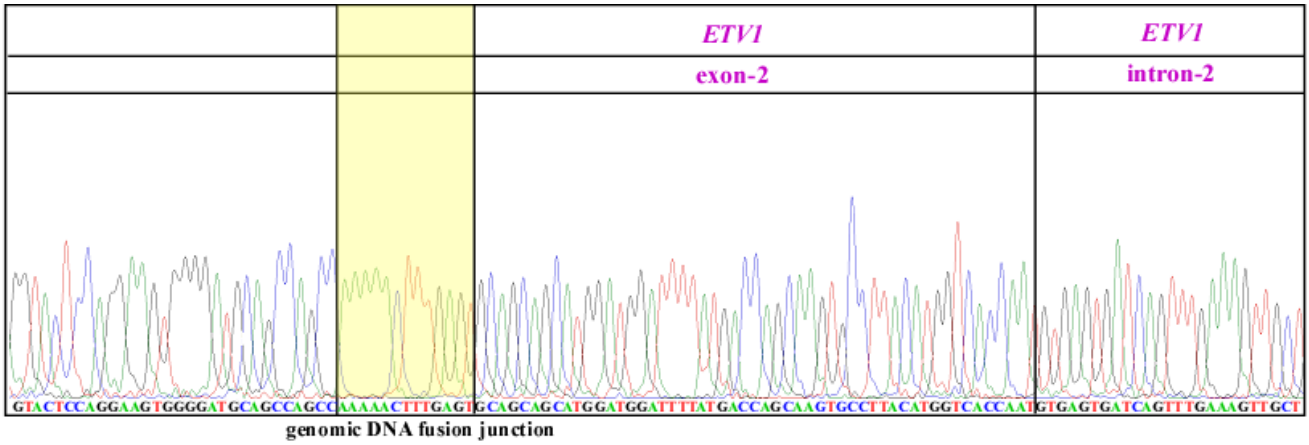
The locations of 23 forward primers (green) spacing across *TMPRSS2* intron-1 (~10 kb) and 7 reverse primers (purple) spacing across *ETV1* intron-2 (~0.8 kb). Each vertical line represents a target location by a primer or a set of nested primers. The specific primers that amplify the genomic breakpoint shown in Fig. 4E in the main text are labeled as red.


pGEM-T
TMPRSS2
intron-1
TMPRSS2 genomic bk F2

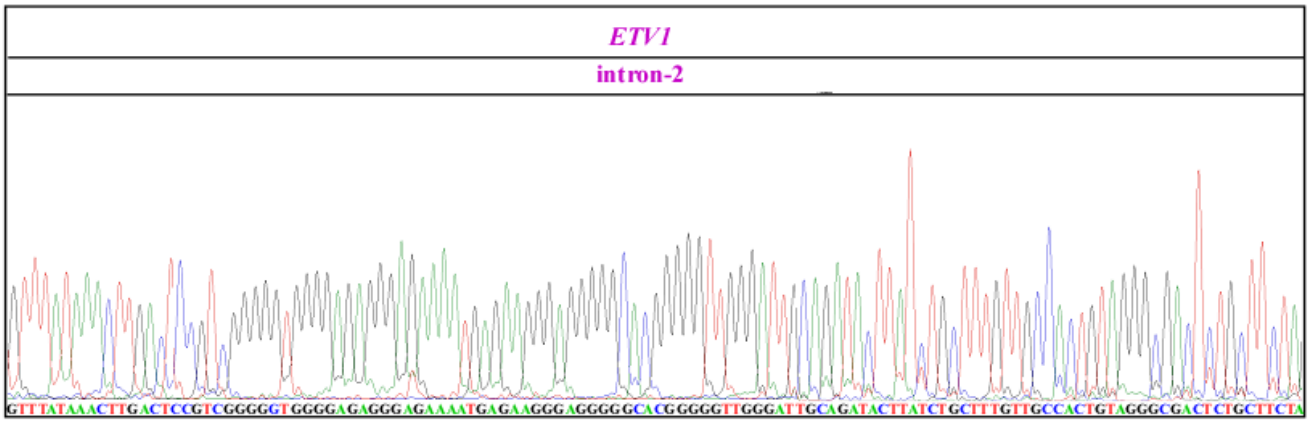

TMPRSS2
intron-1

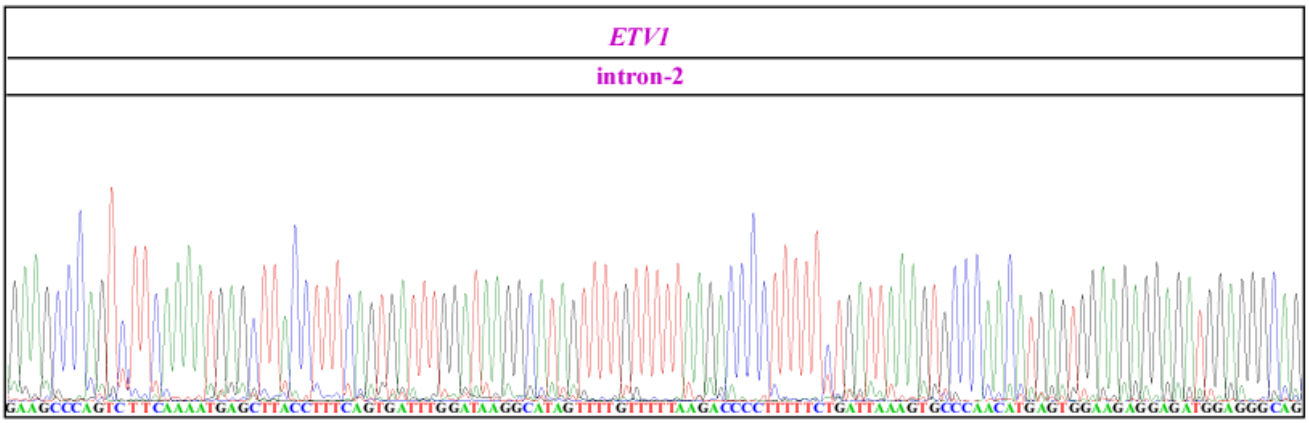

TMPRSS2
intron-1

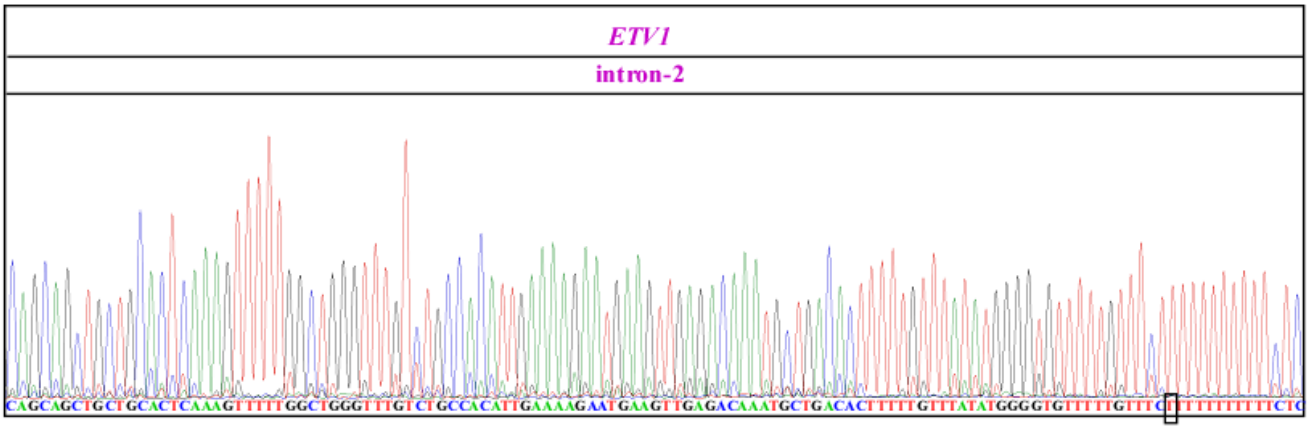

TMPRSS2
intron-1



genomic DNA fusion junction







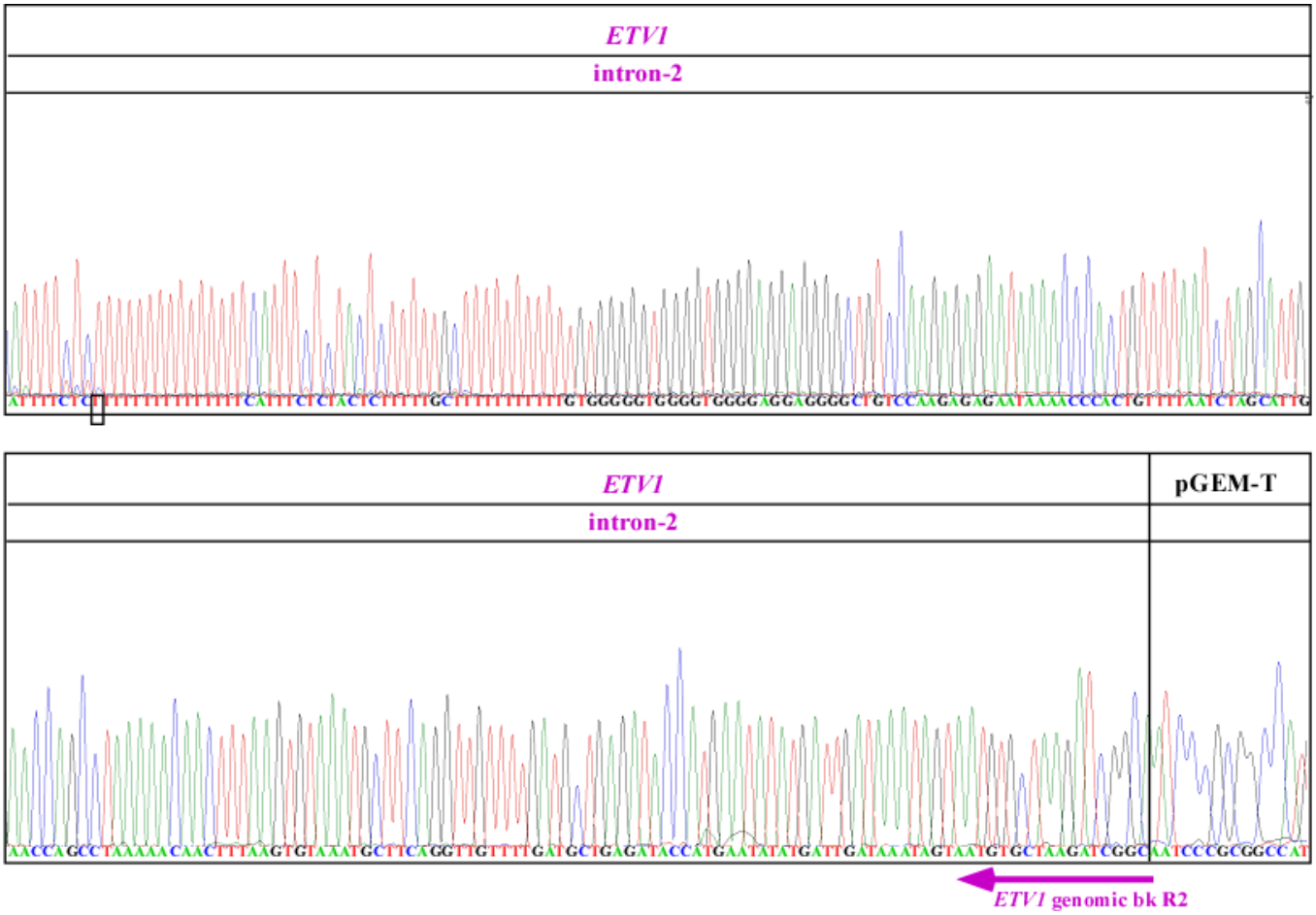
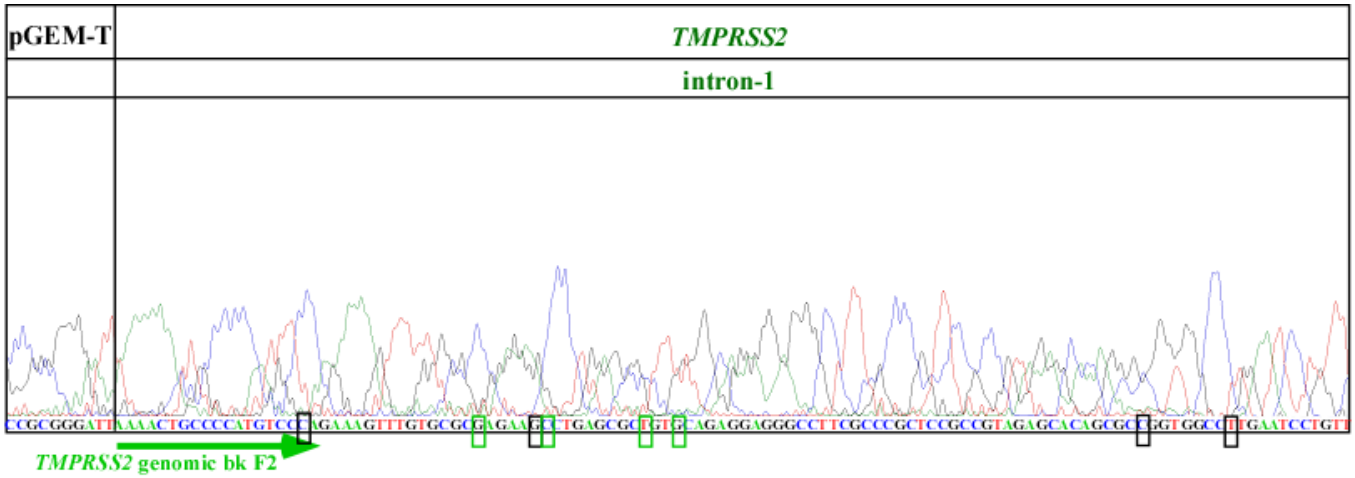
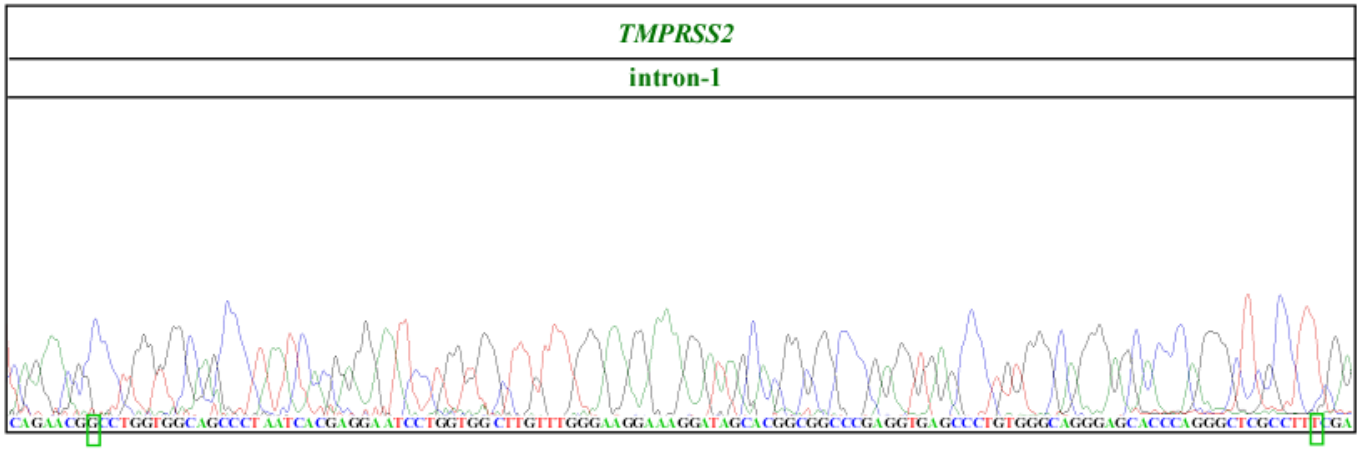
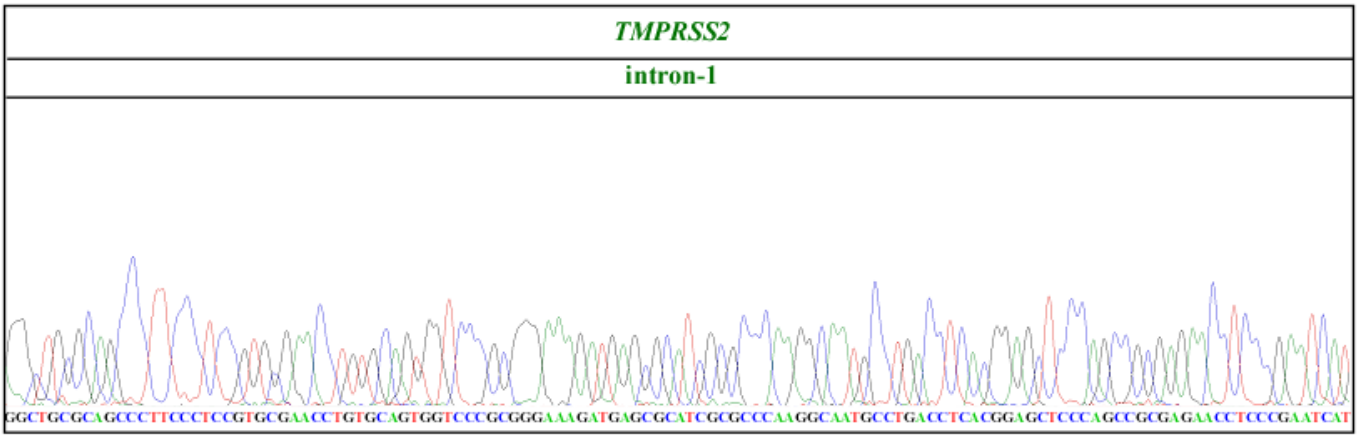


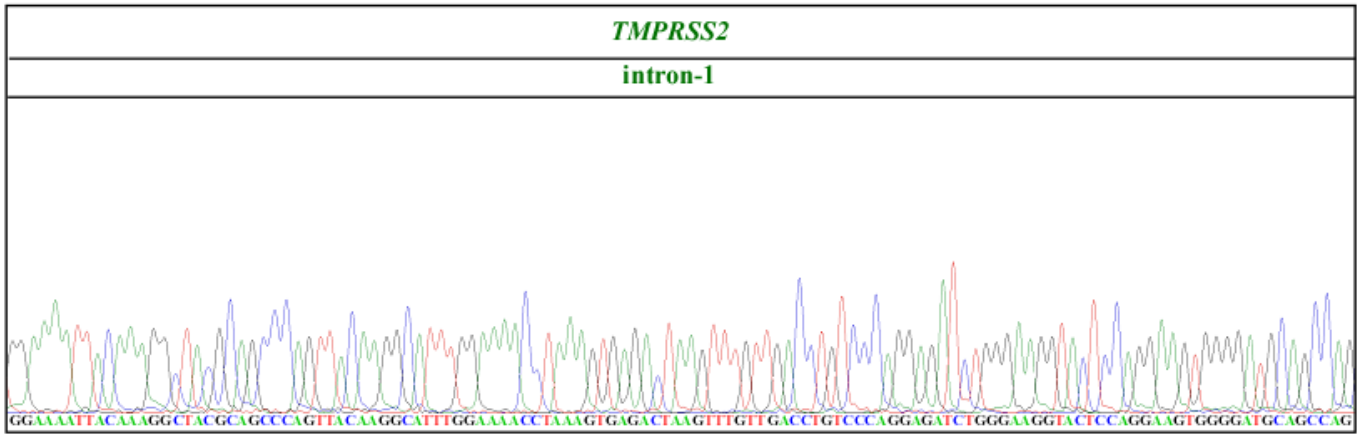
Fig. S22. Sanger sequencing confirmed that the induced *TMPRSS2-ETV1* band x is due to gene fusion resulting from genomic arrangement.

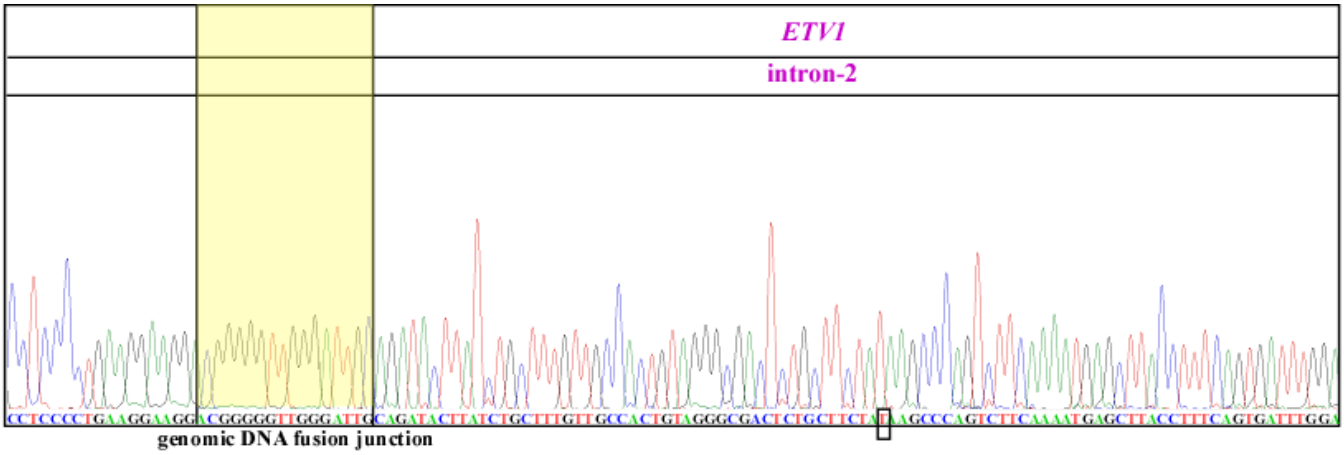
LNCaP cells transfected with antisense-TETV-1 were enriched for the *TMPRSS2-ETV1* fusion transcript using a procedure similar to that described in Fig. S14. Genomic DNA PCR was performed on the enriched LNCaP population which yielded three gene fusion bands of ~1150 bp (band x), ~1044 bp (band y), ~1043 bp (band z), as shown in Fig. 4E in the main text. Sanger sequencing chromatograms confirmed that the 1150 bp band x contains ~485 bp of *TMPRSS2* intron-1 joined to ~665 bp of *ETV1* intron-3. The locations of the primers used to amplify the PCR product are shown as the green and purple arrows for *TMPRSS2* intron-1 and *ETV1* intron-3, respectively. Region of microhomology at the breakpoints are boxed by solid lines. The pGEM-T plasmid was used for cloning the genomic DNA PCR product. Mutations shown in black boxes are PCR artifacts because they are absent in other clones that were sequenced in parallel.

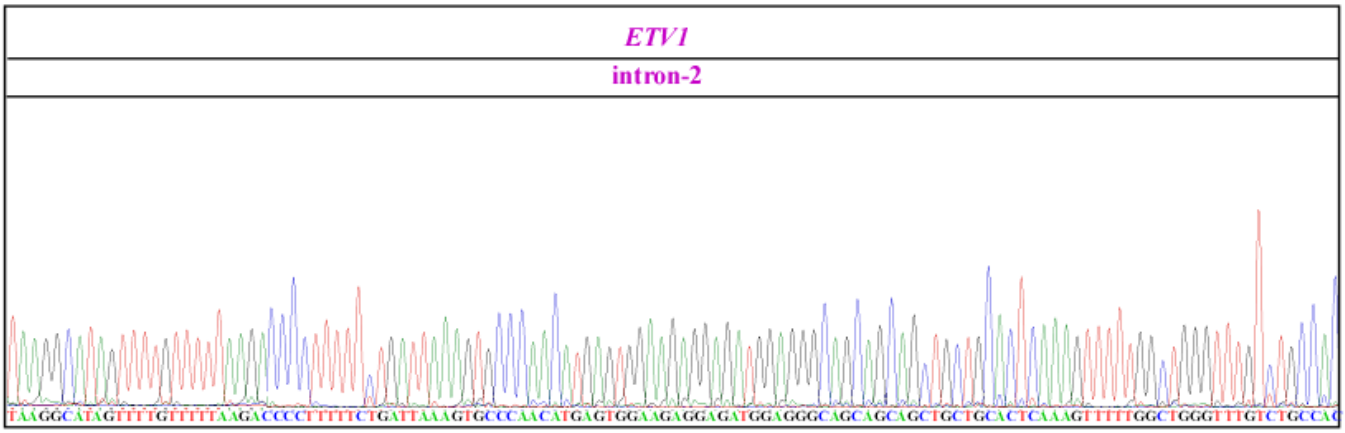

pGEM-T
TMPRSS2
intron-1
TMPRSS2 genomic bk F2

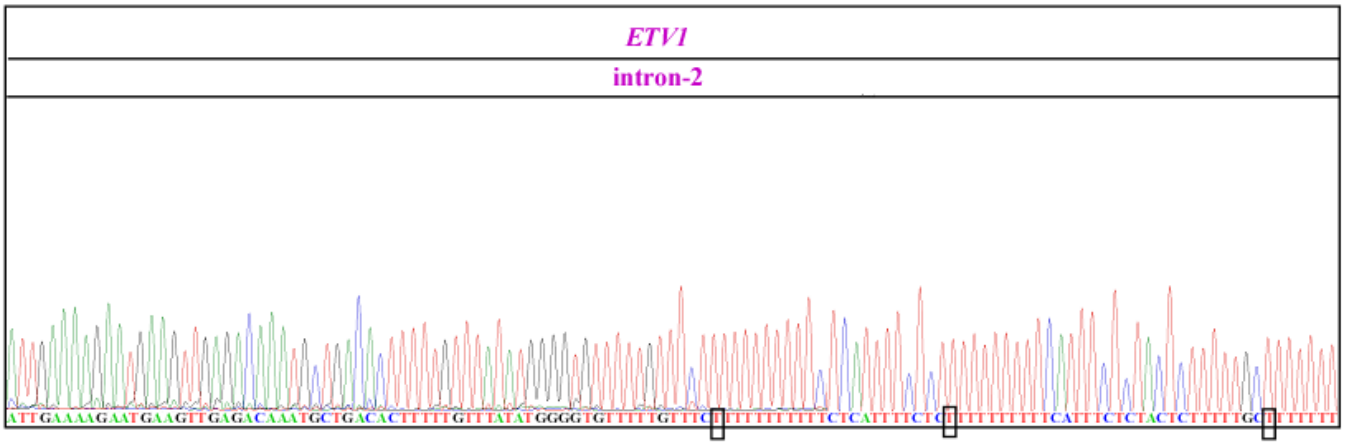

TMPRSS2
intron-1

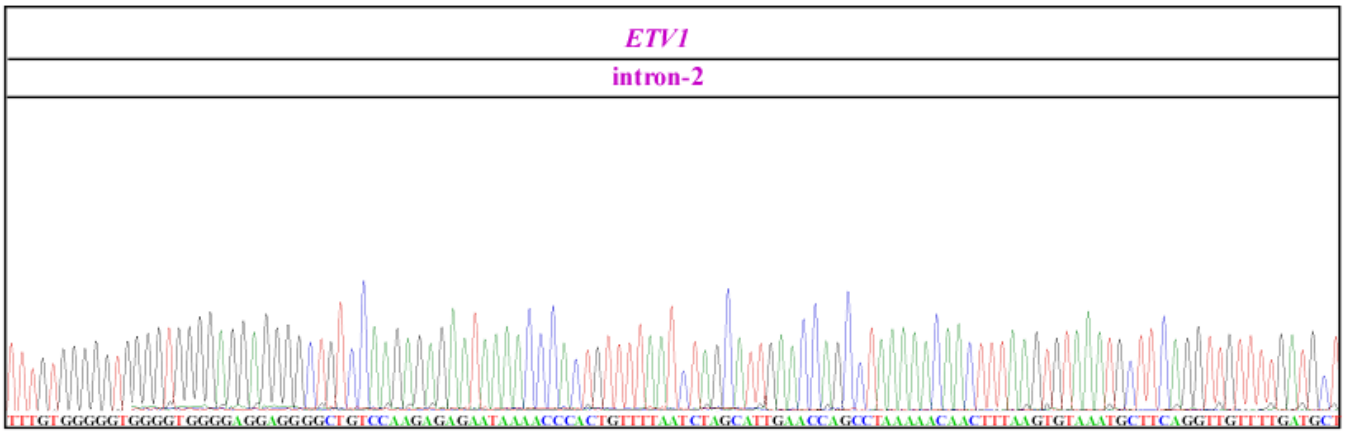

TMPRSS2
intron-1


TMPRSS2
intron-1


ETV1
intron-2
genomic DNA fusion junction


ETV1
intron-2


ETV1
intron-2


ETV1
intron-2

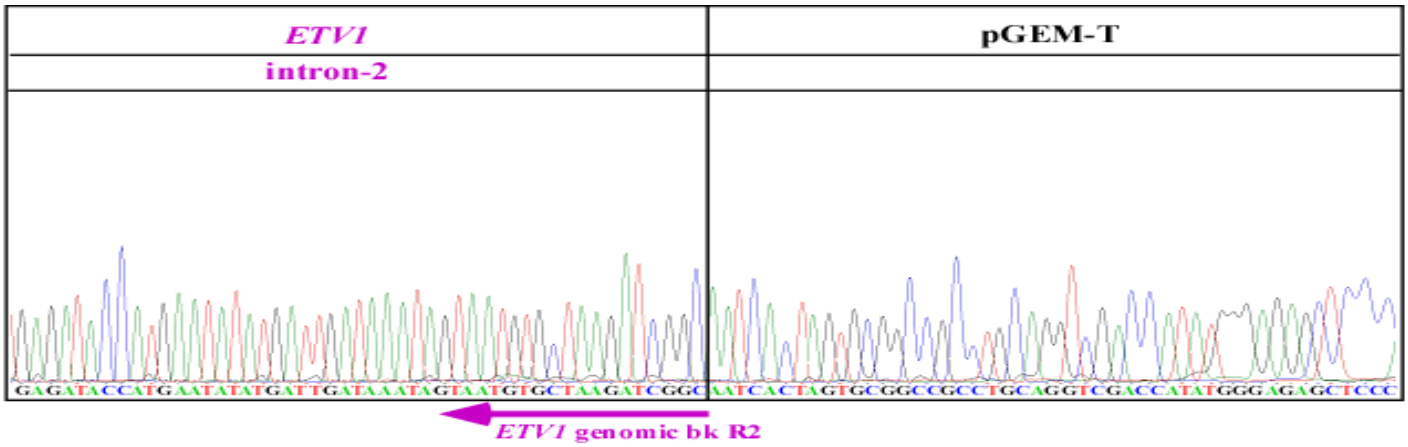
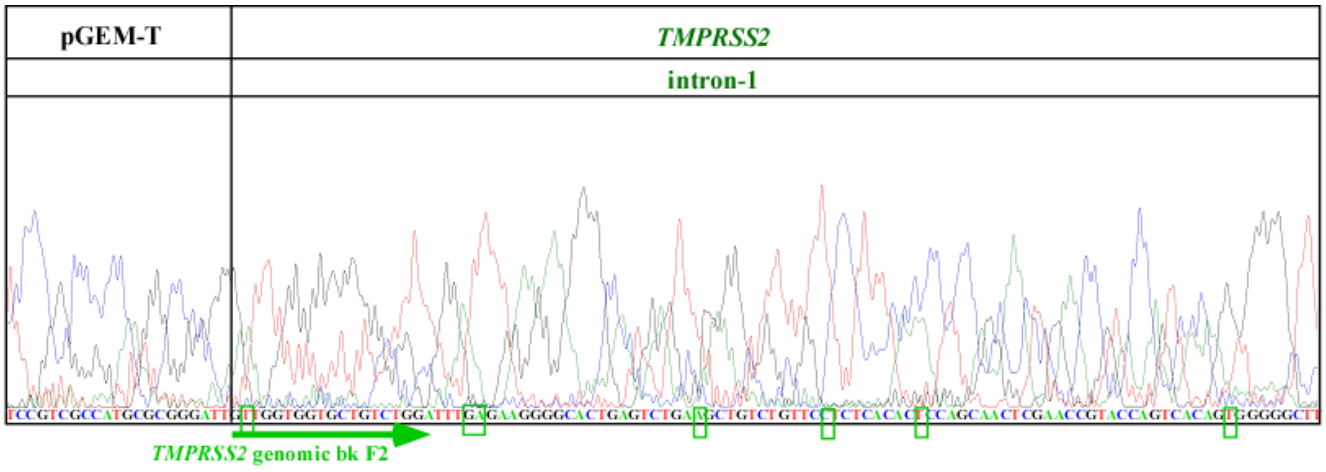
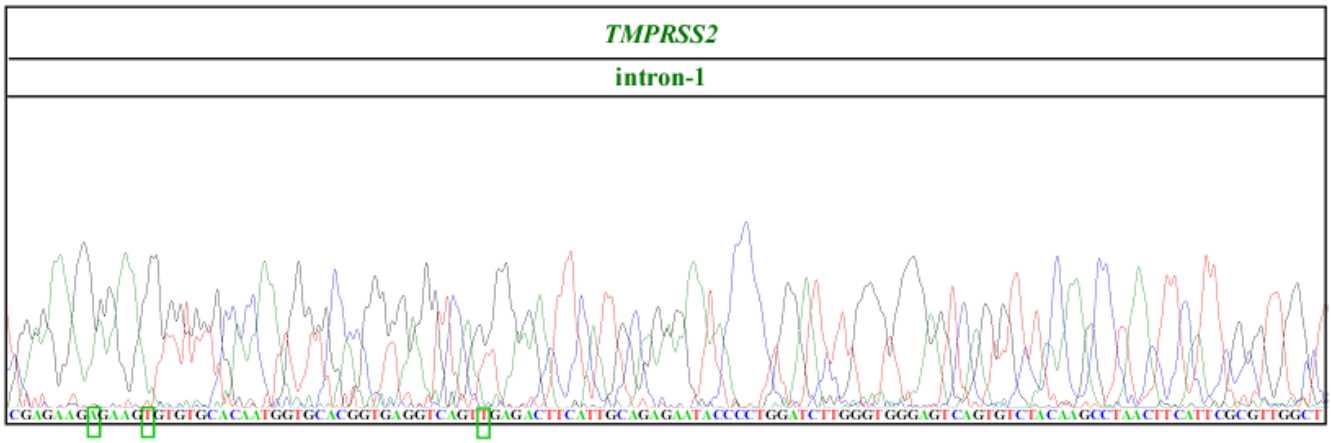
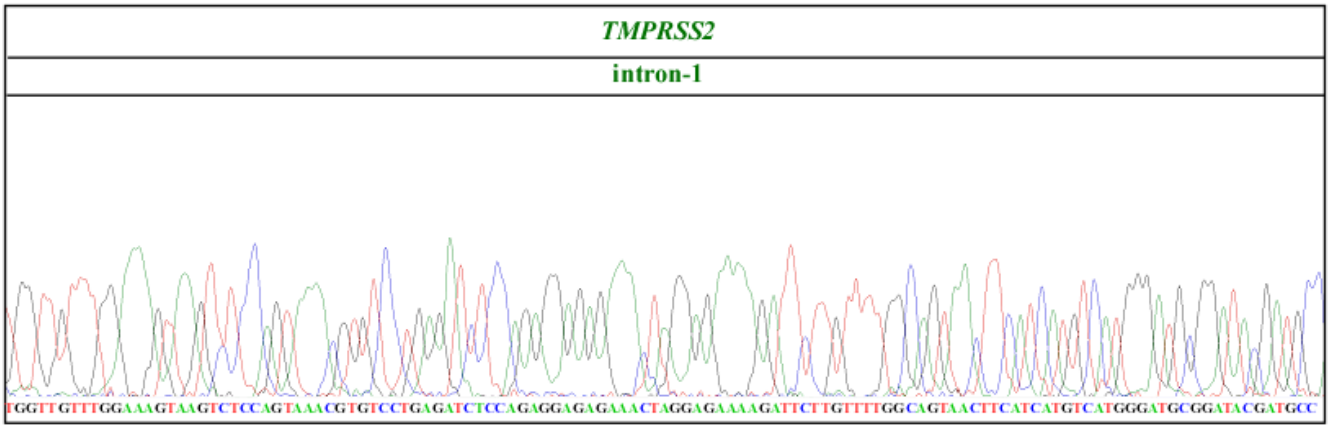


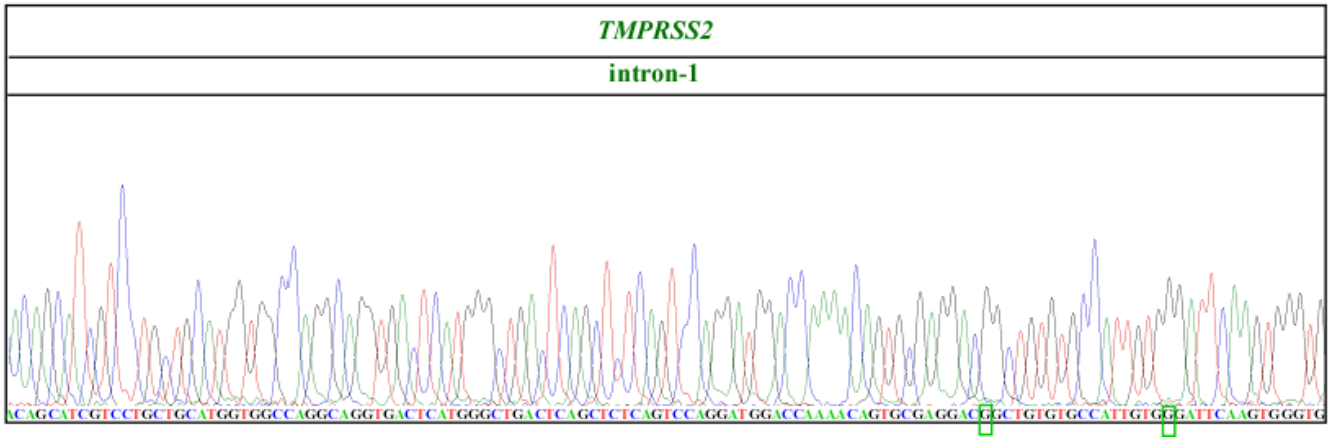
Fig. S23. Sanger sequencing confirmed that the induced *TMPRSS2-ETV1* band y is due to gene fusion resulting from genomic arrangement.

Similarly, Sanger sequencing was performed on the genomic PCR fusion bands ~1044 bp (band y) shown in Fig. 4E in the main text. The resulting chromatograms confirmed that the 1044 bp band y contains ~516 bp of *TMPRSS2* intron-1 joined to ~528 bp of *ETV1* intron-3. The locations of the primers used to amplify the PCR product are shown as the green and purple arrows for *TMPRSS2* intron-1 and *ETV1* intron-3, respectively. Region of microhomology and indels at the breakpoints are boxed by solid lines. The pGEM-T plasmid was used for cloning the genomic DNA PCR product. Mutations shown in small black boxes are PCR artifacts because they are absent in other clones that were sequenced in parallel, whereas mutations shown in small green boxes are either PCR introduced or not verified because of poor sequencing quality in other clones that were sequenced in parallel.

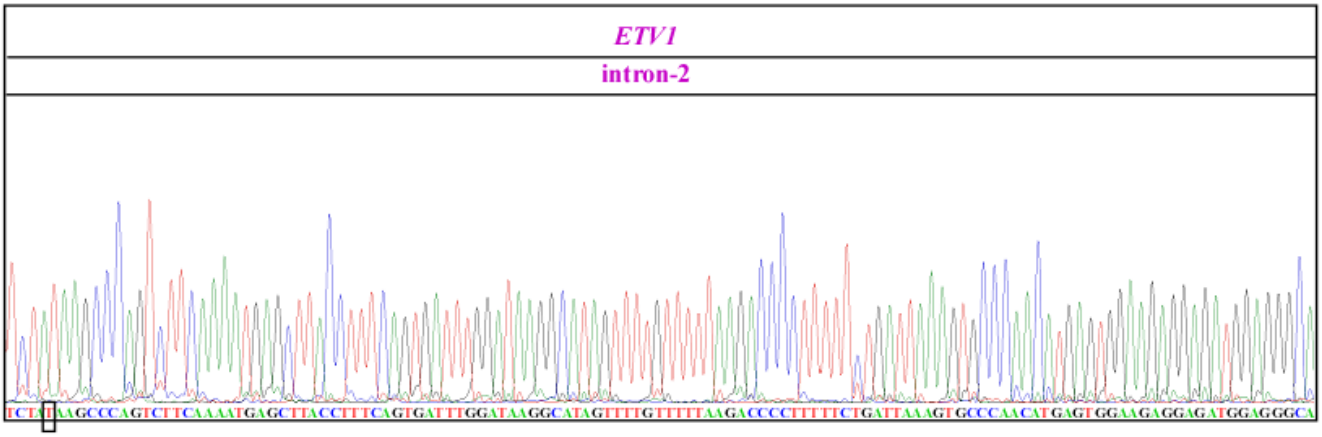

pGEM-T
TMPRSS2
intron-1
TMPRSS2 genomic bk F2

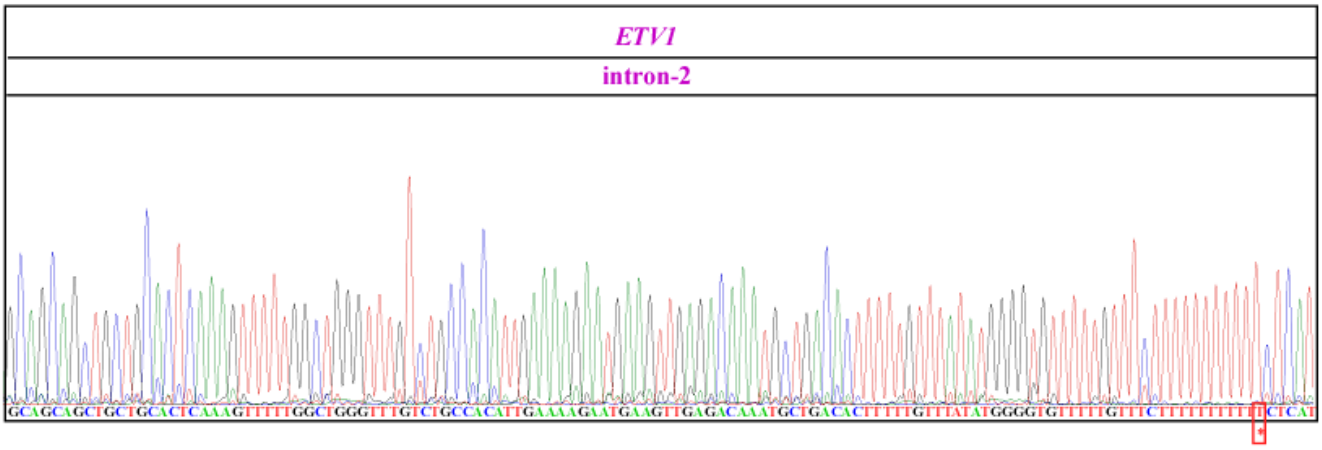

TMPRSS2
intron-1

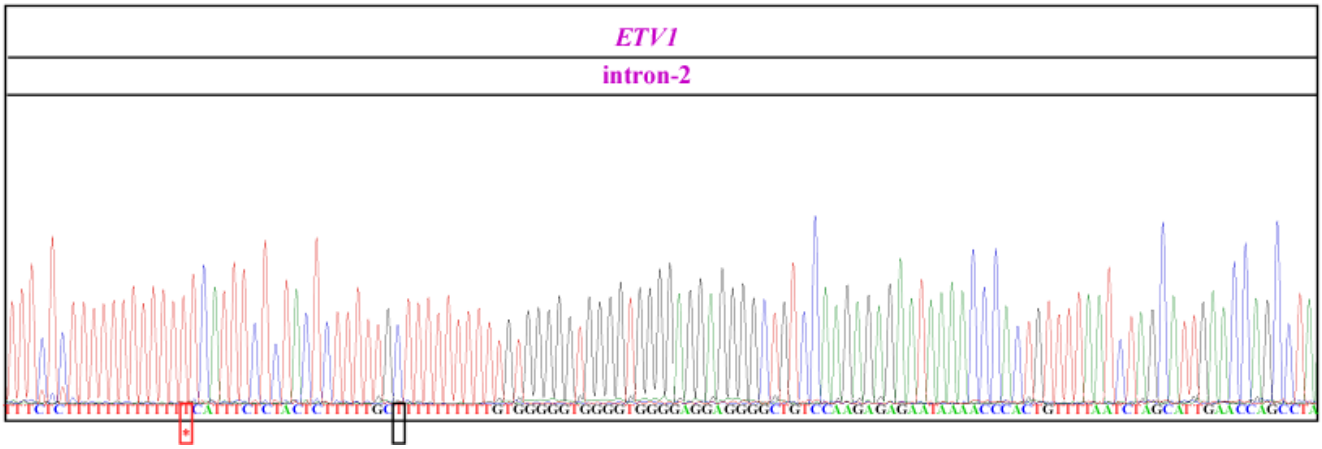

TMPRSS2
intron-1


TMPRSS2
intron-1

genomic DNA fusion junction







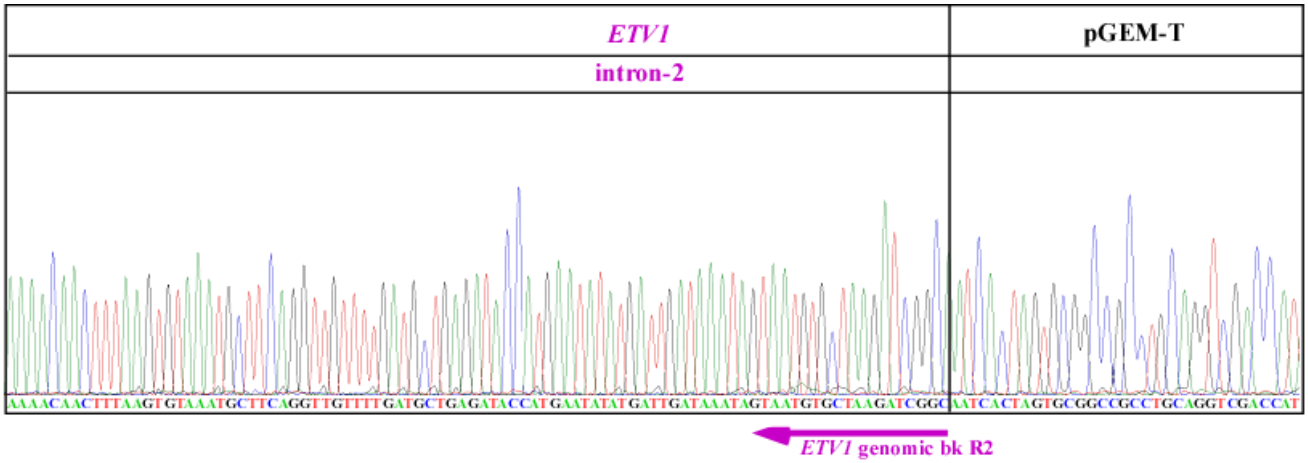


Fig. S24. Sanger sequencing confirmed that the induced *TMPRSS2-ETV1* band z is due to gene fusion resulting from genomic arrangement.

Similarly, Sanger sequencing was performed on the genomic PCR fusion bands ~1043 bp (band z) shown in Fig. 4E in the main text. The resulting chromatograms confirmed that the 1043 bp band z contains ~514 bp of *TMPRSS2* intron-1 joined to ~529 bp of *ETV1* intron-3. The locations of the primers used to amplify the PCR product are shown as the green and purple arrows for *TMPRSS2* intron-1 and *ETV1* intron-3, respectively. Region of microhomology and indels at the breakpoints are boxed by solid lines. The pGEM-T plasmid was used for cloning the genomic DNA PCR product. Mutations shown in small black boxes are PCR artifacts because they are absent in other clones that were sequenced in parallel, whereas mutations shown in small green boxes are either PCR introduced or not verified because of poor sequencing quality in other clones that were sequenced in parallel.

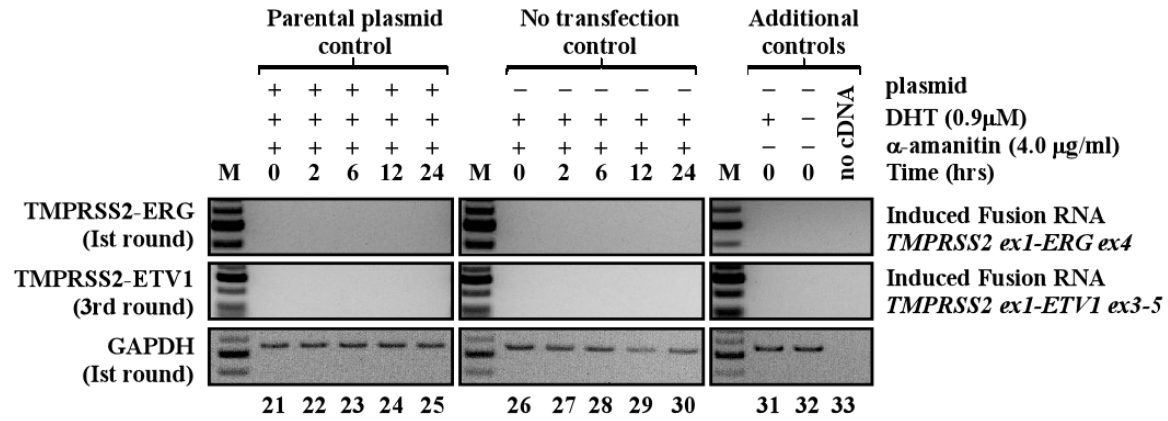


Fig. S25. The disparity between antisense versus sense input RNA is due to transcriptional conflict.

The input RNAs were expressed by U6 (a pol-III promoter) for one day, followed by α-amanitin-mediated inhibition of pol-II transcription for various time periods (0, 2, 6, 12 and 24 hrs) to shut down parental gene transcription. α-amanitin was then removed to resume cellular transcription and the induction by sense vs. antisense input RNA were compared. The corresponding sense input RNAs that previously failed to induce fusion, began to induce *TMPRSS2-ERG* (Fig. 5B) after 12 hours of α-amanitin treatment, and *TMPRSS2-ETV1* fusion (Fig. 5B) after 24 hours of α-amanitin treatment, respectively.

As controls, transfecting cells with a parental plasmid containing no input RNA sequences (lower panel, left column), cells without transfection (lower panel, center column), and DHT and α-amanitin controls (lower panel, right column), also failed to induce fusion. In addition, GAPDH is used as internal control for the amount of RNA in each lane.

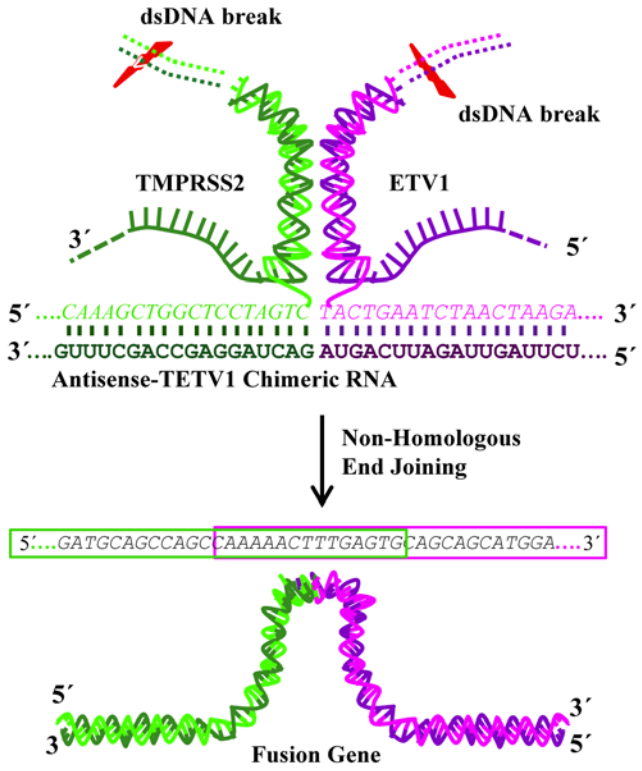


Fig. S26. A model of RNA-mediated gene fusion in mammalian cells.

Our data support a model where the initiator RNA with antisense-TETV1 chimeric sequence invades chromosomal DNA of *TMPRSS2* and *ETV1* to stabilize a transient RNA/DNA duplex (reminiscent of an R-loop) using DNA sequences located in two distant genes. Resolution of such an RNA/DNA duplex by DNA break/repair mechanisms might yield the final gene fusion through recombination in regions prone to DNA breaks. Base-pairing of antisense-TETV1 RNA sequence with *TMPRSS2* and *ETV1* genes are illustrated. For simplicity, only the partial antisense-TETV1 RNA sequence is shown (upper panel). Genomic break point “x” identified in Fig. 4E is shown in lower panel.

Movie S1 & S2. Examples of *TMPRSS2-ETV1* fusion gene identified by 3D-microscopy FISH image reconstruction.

The enriched *TMPRSS2-ETV1* LNCaP cell population was fixed with formaldehyde followed by hybridization with FISH probes against *TMPRSS2* gene (red) and *ETV1* gene (green). Images were acquired on a GE Healthcare DeltaVision image restoration microscope with an Olympus UPLS Apo 100x/1.4 NA and 2k x 2k EDGE/sCMOS camera. Z stacks were acquired with spacing of 0.25 um. Images were deconvolved using conservative algorithm in Resolve3D SoftWorx (version 6.5.2). Max projection and volume visualization was also performed with the same software. S1 represents 3-D along x-axis and S2 along y-axis.

Materials and Methods

LNCaP cell culture

LNCaP cells were routinely cultured in RPMI 1640 medium (RPM1 1640, 1X, with L-glutamine, #10-040-CV, CORNING cellgro) containing 10% fetal bovine serum (premium grade FBS, #1500-500, Seradigm) and 1% penicillin/streptomycin (#15140-122, Gibco) in a 5% CO_2 humidified incubator. For experiments involving the induction of fusion gene by input RNA, regular fetal bovine serum in the culture medium was replaced by Charcoal:Dextran stripped fetal bovine serum (catalog#100-119, Gemini Bioproducts) to remove hormones present in serum. LNCaP cells were cultured in this special medium for 24 hrs prior to plasmid transfection.

PNT1A cell culture

PNT1A cells were routinely cultured in RPMI 1640 medium containing 10% fetal bovine serum (premium grade FBS, #1500-500, Seradigm) and 1% penicillin/streptomycin (#15140-122, Gibco) in a 5% CO_2 humidified incubator.

Transient transfection of plasmids for expressing the input RNAs

Twenty hours prior to transfection, LNCaP cells were seeded in 12-wells plate (BioLite 12 Well Multidish, #130185, Thermo Fisher Scientific) with a density of $5x10^5$cells/well and 1 ml/well of culture medium containing Charcoal:Dextran stripped fetal bovine as described above. Transfection was performed using Turbofect transfection reagent (Thermo Scientific, #R0531) according to manufacturer's protocol. Briefly, 1µg of a particular plasmid was first diluted in 100µl of the serum-free DMEM followed by immediate mixing by pipetting. 4µl of the transfection reagent was then added to the diluted DNA followed by mixing and incubation for 20 min. The DNA/transfection reagent mixture was then added drop wise to a well containing LNCaP cells in 1ml medium.

For transfection in PNT1A cells, $5x10^5$ cells/well were plated in 12-wells plate in 1 ml/well of cultured medium 24 hrs prior to transfection. Transfection was performed using the same formula described for LNCaP cells. For repetitive transfections, initially transfected PNT1A cell population were split every three days, half was processed for RT-PCR assay and half was seeded again in a new well for the next transfection.

DHT preparation and treatment

DHT (Dihydrotestosterone) was purchased from Sigma Adlrich (5α-Androstan-17β-ol-3-one, #A8380). Concentrated stock of 1500µM was prepared by dissolving 4.3566 mg of DHT powder in 10 ml of 100% ethanol (200 proof ethanol, Koptec, #V1016) and then aliquoted in 1ml tubes and stored at -80ºC.

For treating cultured cells, concentrated DHT stock was diluted as 10x working solutions (for example, for 0.9 μM final concentration, 10x is prepared as 9.0 μM) with the appropriate complete culture medium and used immediately. Complete media for LNCaP cells: RPMI 1640 + 10% Charcoal:Dextran stripped fetal bovine serum + 1% penicillin/streptomycin. Complete media for PNT1A: RPMI 1640 + 10% fetal bovine serum + 1% penicillin/streptomycin. Six hrs post transfection, 111μl of fresh 10x DHT working solutions was added to each well of 12-wells plate containing 1ml medium and transfected cells.

For long-term treatment, medium was changed with fresh DHT every three days.

RNA isolation

Total RNA from cultured cells was extracted using High Pure RNA isolation Kit according to manufacturer's instructions (#11828665001, Roche). Briefly, cells were suspended in 200μl of PBS buffer and were then lysed with 400μl of lysis buffer. The sample was then passed through the filter assembly resulting in the binding of the nucleic acids to the filter. The filter containing nucleic acids was then incubated with DNase I dissolved in DNase incubation buffer to degrade genomic and plasmid DNAs. The column was then rinsed with wash buffer and total RNA then eluted in a new tube for further analysis.

For detection of residual genomic and plasmid DNA, eluted RNA was subject to PCR reaction with primers specific to intron regions of house-keeping gene GAPDH, and with primers specific to plasmid transfected. Total RNA was converted to cDNA only if it is validated as free of DNA contamination.

Reverse transcription reaction

1 μg of total RNA was used for each reverse transcription reaction according to manufacturer instruction (superscript III RT, # 18080-051, Invitrogen). RNA was converted to cDNA either with Oligo dT primer (for induced fusion transcripts) or with random hexamers (for input RNAs expressed by U6 promoter). After the addition of dNTPs, the mixture was denatured at 65°C for 5 minutes. This was followed by the addition of a master-mix containing 1× superscript buffer, 10 mM DTT, 5 mM Magnesium chloride, RNaseOUT and Superscript III reverse transcriptase. Reactions were carried out at 50°C for 50 minutes and then terminated by incubation at 85°C for 5 minutes. cDNA was then treated with RNase-H for 20 minutes at 37°C to degrade RNA in DNA/RNA hybrid. 1 μl of cDNA was used as template for each subsequent PCR reaction.

RT-PCR for detecting induced fusion transcripts

The majority of induced fusion RNAs in this manuscript were detected using one-round RT-PCR. The following cases were assayed using three-round nested PCR: (1) the results of DHT treatment at physiological concentrations as shown in Fig. 1D and 6A, (2) the induction of *TMPRSS2-ERG* fusion transcript in non-malignant PNT1A cells as shown in Fig. 3F, (3)

the specificity of input RNAs assayed in Fig. 4C, and (4) endogenous *ERG* level detection in LNCaP cells in Fig. S13. The following cases were assayed using two-round nested PCR: the induction of *TMPRSS2-ETV1* fusion RNA as shown in Fig. 4A and 4B, and the detection of *TMPRSS2-ETV1* fusion in the enriched population in Fig. 4D.

PCR was done with a standard three-step protocol using REDTaq DNA polymerase (#D5684-1KU, Sigma) according to manufacturer instruction.

Reaction was set as follows:

PCR reaction:

Forward primer :	1.0µl (from 10µM stock, Sigma)
Reverse primer :	1.0µl (from 10µM stock, Sigma)
10x reaction buffer:	5.0µl (comes with REDTaq, Sigma)
dNTPs:	1.0µl (from 10mM stock, #11969064001, Roche)
DMSO:	1.5µl (#154938, Sigma-Aldrich)
cDNA:	1.0µl (from 20µl stock prepared from 1 µg RNA)
Autoclaved Milli-Q water:	38.5µl
REDTaq DNA polymerase:	1.0µl (#D5684-1KU, Sigma)
Total volume:	50µl

Standard one-round PCR conditions for *TMPRSS2-ERG*:

Pre-denaturation	94^0C, 4 min	
Denaturation	94^0C, 30 sec	32 cycles for induced fusion RNA
Annealing	58^0C, 45 sec	27 cycles for input RNA
Extension	72^0C, 60 sec	
Final Extension	72^0C, 5 min	

PCR conditions for three-round nested PCR for *TMPRSS2-ERG*:

1st round : PCR with *TMPRSS2* ex-1 F1 and *ERG* ex-4 R1 on 1µl of cDNA

Pre-denaturation	94^0C, 4 min	
Denaturation	94^0C, 30 sec	
Annealing	58^0C, 45 sec	32 cycles for induced fusion RNA
Extension	72^0C, 60 sec	
Final Extension	72^0C, 5 min	

2nd round : PCR with *TMPRSS2* ex-1 F2 and *ERG* ex-4 R2 on 1µl of 1st round product, PCR conditions same as 1st round.

3rd round : PCR with *TMPRSS2* ex-1 F3 and *ERG* ex-4 R3 on 1µl of 2nd round product, PCR conditions same as 1st round.

PCR conditions for two-round nested PCR for *TMPRSS2-ETV1*:

1st Round: Top down PCR with *TMPRSS2* ex-1 F1 and *ETV1* ex-6 R1

Denaturation	94^0C, 30 sec	2 cycles
Annealing	* , 45 sec	
Extension	72^0C, 60 sec	

*2 cycles at each temperature: 62^0C, 61^0C, 60^0C…….49^0C followed by

Denaturation	94^0C, 30 sec	20 cycles
Annealing	48 ^{0}C, 45 sec	
Extension	72^0C, 60 sec	

2nd Round: PCR with *TMPRSS2* ex-1 F2 and *ETV1* ex-5 R1 on 1µl of 1st round.

Pre-denaturation	94^0C, 4 min	
Denaturation	94^0C, 30 sec	32 cycles for Induced fusion RNA
Annealing	57^0C, 45 sec	
Extension	72^0C, 60 sec	
Final Extension	72^0C, 5 min	

3rd round (for Fig. 4D): PCR with *TMPRSS2* ex-1 F3 and *ETV1* ex-5 R2 on 1µl of 2nd round, PCR conditions same as 2nd round.

Long range PCR for detecting genomic DNA fusion junction

Nested long-range PCRs according to the manufacturer's protocols using LA PCR kit (Takara, # RR002M). 200 ng of genomic DNA was used in each reaction and PCR was performed with annealing and extension at 68°C for 20 minutes. 1µl from the above reaction (1st round PCR) was used as template for the 2nd round PCR.

For the genomic breakpoint identified in this manuscript, 1st round long range PCR was done using primers *TMPRSS2* genomic bk-F1 and *ERG* genomic bk-R1 shown in primer list below. 1µl from the above reaction (1st round PCR) was used as template for the 2nd round PCR using inner primers *TMPRSS2* genomic bk-F2 and *ERG* genomic bk-R2.

Cloning and Sanger sequencing of induced fusion transcripts

PCR amplified cDNA bands were excised from the gel and eluted using QIAquick Gel Extraction Kit (#28706, Qiagen). The eluted bands were then cloned to pGEM-T vector (pGEM-T vector system I, # A3600) following manufacturer instruction. Sanger sequencing was performed using the service of Beckman Coulter Genomics.

Tm calculations

Melting temperature (Tm) of putative genomic DNA stems were calculated using the following formula (Rychlik and Rhoads, 1989):

$$Tm\ (°C) = 4 \times (\text{number of G's and C's}) + 2 \times (\text{number of A's and T's})$$

A high energy G·T and A·C wobble pair known to have Watson-Crick like geometry in DNA double helix (Kimsey and Al-Hashimi, 2014; Watson and Crick, 1953) are considered as having the same stability as an A·T pair.

Fluorescent in situ hybridization (FISH)

Enriched population carrying *TMPRRS2-ETV1* fusion events were first grown on 18 mm round #1 coverglass in a 12-well cell culture plate at the initial density of 200-400k/well. Cells were then fixed with 4% (vol./vol.) formaldehyde followed by denaturation of DNA with 0.1 N HCl for 5 min and with 70% formamide at 85°C for 7 min. Hybridization of target DNA with probes were done at 37°C for 16hr in a humidified chamber. Cells were then washed, stained with DAPI and imaged with microscope. FISH probes for *TMPRSS2* (RP11-35C4, red) and *ETV1* (RP11-769K2, green) were purchased from Empire Genomics.

α-Amanitin assay

Twenty hours prior to transfection, LNCaP cells were seeded in 12-wells plate (BioLite 12 Well Multidish, #130185, Thermo Fisher Scientific) with a density of $5x10^5$cells/well and transfection was performed using Turbofect transfection reagent (Thermo Scientific, #R0531) as described earlier. DHT was added at the final concentration of 0.9µM six hours post transfection. Following overnight incubation, cells were then treated with 4µg/ml α-amanitin for various time periods (0, 2, 6, 12 and 24 hours). Cells were then revived in fresh medium containing 0.9µM DHT without α-amanitin and RT-PCR was performed for either *TMPRSS2-ERG* or *TMPRSS2-ETV1* fusion.

RNase H treatment

LNCaP cells were transfected with input chimeric RNA expression plasmid together with a second plasmid that expresses wildtype RNaseH (WT:pICE-RNaseHI-WT-NLS-mCherry, Addgene No. 60365) in ratio 2:3 (400 ng antisense5/antisense-TETV-1:600ng RNase H). As a control, an inactive mutant (MUT: pICE-RNaseHI-D10R-E48R-NLS-mCherry, Addgene No. 60367) that lacks the ability to degrade RNA was used for head-to-head comparisons.

Input RNA sequences

Sense-1

+1tcagatcgcctggagacgccatccacgctgttttgacctccatagaagacaccgggacc
gatccagcctcccctcgaagctgatcctgagaacttcaggctcctgggcaacgtgctggtc
tgtgtgctggcccatcactttggcaaagaattcGAGTAGGCGCGAGCTAAGCAGGAGGCGG
AGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAGGAAGCCTTATC
AGTTGTGAGTGAGGACCAGTCGTTGTTTGAGTGTGCCTACGGAACGCCACACCTGGCTAAG
ACAGAGATGACCGCGTCCTCCTCCAGCGACTATGGACAGACTTCCAAGATGAGCCCACGCG
TCCCTCAGCAGGATTGGCTGTCTCAACCCCCAGCCAGGGTCACCATCAAAATGGAATGTAA
CCCTAGCCAGGTGAATGGCTCAAGaagcttatcgataccgtcgacctcgagggcccagatc
taattcaccccaccagtgcaggctgcctatcagaaagtggtggctggtgtggctaatgccc
tggcccacaagtatcactaagctcgctttcttgctgtccaatttctattaaaggttccttt
gttccctaagtccaactactaaactgggggatattatgaagggccttccggagcatctgga
ttctgcctaataaaaaacatttattttcattgc*aaaaaa……a*

: *TMPRSS2* : *ERG* **+1**: Transcription start aataaa: poly(A) signal *aaaaaa……a*: poly(A) tail gaattc: EcoRI aagctt: HindIII

Sense-2

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagGAGTAGGCGCG
AGCTAAGCAGGAGGCGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCG
CGGCAGGAAGCCTTATCAGTTGTGAGTGAGGACCAGTCGTTGTTTGAGTGTGCCTACGGAA
CGCCACACCTGGCTAAGACAGAGATGACCGCGTCCTCCTCCAGCGACTATGGACAGACTTC
CAAGATGAGCCCACGCGTCCCTCAGCAGGATTGGCTGTCTCAACCCCCAGCCAGGGTCACC
ATCAAAATGGAATGTAACCCTAGCCAGGTGAATGGCTCAAG

: *TMPRSS2* : *ERG* **+1**: Transcription start ctgcag: PstI

Sense-2 long

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagGAGTAGGCGCG
AGCTAAGCAGGAGGCGGAGGCGGAGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCG
CGGCAGGAAGCCTTATCAGTTGTGAGTGAGGACCAGTCGTTGTTTGAGTGTGCCTACGGAA
CGCCACACCTGGCTAAGACAGAGATGACCGCGTCCTCCTCCAGCGACTATGGACAGACTTC
CAAGATGAGCCCACGCGTCCCTCAGCAGGATTGGCTGTCTCAACCCCCAGCCAGGGTCACC
ATCAAAATGGAATGTAACCCTAGCCAGGTGAATGGCTCAAGGAACTCTCCTGATGAATGCA
GTGTGGCCAAAGGCGGGAAGATGGTGGGCAGCCCAGACACCGTTGGGATGAACTACGGCAG
CTACATGGAGGAGAAGCACATGCCACCCCCAAACATGACCACGAACGAGCGCAGAGTTATC
GTGCCAGCAGATCCTACGCTATGGAGTACAGACCATGTGCGGCAGTGGCTGGAGTGGGCGG
TGAAAGAATATGGCCTTCCAGACGTCAACATCTTGTTATTCCAGAACATCGATGGGAAGGA
ACTGTGCAAGATGACCAAGGACGACTTCCAGAGGCTCACCCCCAGCTACAACGCCGACATC
CTTCTCTCACATCTCCACTACCTCAGAGAGACTCCTCTTCCACATTTGACTTCAGATGATG
TTGATAAAGCCTTA

: *TMPRSS2* : *ERG* **+1**: Transcription start ctgcag: PstI

Antisense-1

+1tcagatcgcctggagacgccatccacgctgttttgacctccatagaagacaccgggacc
gatccagcctcccctcgaagctgatcctgagaacttcaggctcctgggcaacgtgctggtc

tgtgtgctggcccatcactttggcaaagaattcCTTGAGCCATTCACCTGGCTAGGGTTAC
ATTCCATTTTGATGGTGACCCTGGCTGGGGGTTGAGACAGCCAATCCTGCTGAGGGACGCG
TGGGCTCATCTTGGAAGTCTGTCCATAGTCGCTGGAGGAGGACGCGGTCATCTCTGTCTTA
GCCAGGTGTGGCGTTCCGTAGGCACACTCAAACAACGACTGGTCCTCACTCACAACTGATA
AGGCTTCCTGCCGCGCTCCAGGCGGCGCTCCCCGCCCCTCGCCCTCCGCCTCCGCCTCCGC
CTCCTGCTTAGCTCGCGCCTACTCaagcttatcgataccgtcgacctcgagggcccagatc
taattcaccccaccagtgcaggctgcctatcagaaagtggtggctggtgtggctaatgccc
tggcccacaagtatcactaagctcgctttcttgctgtccaatttctattaaaggttccttt
gttccctaagtccaactactaaactgggggatattatgaagggccttccggagcatctgga
ttctgcctaataaaaaacatttattttcattgc*aaaaaa……a*

: *TMPRSS2* : *ERG* **+1**: Transcription start aataaa: poly(A) signal *aaaaaa……a*: poly(A) tail gaattc: EcoRI aagctt: HindIII

Antisense-2

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCTTGAGCCATT
CACCTGGCTAGGGTTACATTCCATATTGATGGTGACCCTGGCTGGGGGTTGAGACAGCCAA
TCCTGCTGAGGGACGCGTGGGCTCATCTTGGAAGTCTGTCCATAGTCGCTGGAGGAGGACG
CGGTCATCTCTGTCTTAGCCAGGTGTGGCGTTCCGTAGGCACACTCAAACAACGACTGGTC
CTCACTCACAACTGATAAGGCTTCCTGCCGCGCTCCAGGCGGCGCTCCCCGCCCCTCGCCC
TCCGCCTCCGCCTCCGCCTCCTGCTTAGCTCGCGCCTACTC

: *TMPRSS2* : *ERG* **+1**: Transcription start

A: T to A change to inactivate the cryptic transcription termination by U6 promoter
ctgcag: PstI

Antisense-3

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCTTGAGCCATT
CACCTGGCTAGGGTTACATTCCATATTGATGGTGACCCTGGCTGGGGGTTGAGACAGCCAA
TCCCGGACCCCGAGCCGGGACCCTGGTACCGGCGCCGCTCACCTGCCGCGCTCCAGGCGGC
GCTCCCCGCCCCTCGCCCTCCGCCagacaggagtgagagatggaagctcgcgcctactc

: *TMPRSS2* : *ERG* : Mutated region of *TMPRSS2* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter
ctgcag: PstI

Antisense-4

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCTTGAGCCATT
CACCTGGCTAGGGTTACATTCCATATTGATGGTGACCCTGGCTGGGGGTTGAGACAGCCAA
TCCGACCCTGGTACCGGCGCCGCTCACCTGCCGCGCTCCAGGCGGCGCTCCCCGCCCCTCG
CCCTCCGCCagacaggagtgagagatggaagctcgcgcctactc

: *TMPRSS2* : *ERG* : Mutated region of *TMPRSS2* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Antisense-5

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCTTGAGCCATT
CACCTGGCTAGGGTTACATTCCATATTGATGGTGACCCTGGCTGGGGGTTGAGACAGCCAA
TCCGCCGCTCACCTGCCGCGCTCCAGGCGGCGCTCCCCGCCCCTCGCCCTCCGCCagacag
gagtgagagatggaagctcgcgcctactc

: *TMPRSS2* : *ERG* : Mutated region of *TMPRSS2* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter

ctgcag: PstI

Antisense-6

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCTTGAGCCATT
CACCTGGCTAGGGTTACATTCCATATTGATGGTGACCCTGGCTGGGGGTTGAGACAGCCAA
TCCCCAGGCGGCGCTCCCCGCCCCTCGCCCTCCGCCagacaggagtgagagatggaagctc
gcgcctactc

: *TMPRSS2* : *ERG* : Mutated region of *TMPRSS2* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Antisense-7

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCTTGAGCCATT
CACCTGGCTAGGGTTACATTCCATATTGATGGTGACCCTGGCTGGGGGTTGAGACAGCCAA
TCCCGGACCCCGAGCCGGGACCCTGGTACCGGCGCCGCTCACCTGCCGCGCTCCAagacag
gagtgagagatggaagctcgcgcctactc

: *TMPRSS2* : *ERG* : Mutated region of *TMPRSS2* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Antisense-8

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCTTGAGCCATT
CACCTGGCTAGGGTTACATTCCATATTGATGGTGACCCTGGCTGGGGGTTGAGACAGCCAA
TCCCCAGGTTCCCCTCCCCAGCCCGGACCCCGAGCCGGGACCCTGGTACCGGCGCagacag
gagtgagagatggaagctcgcgcctactc

: *TMPRSS2* : *ERG* : Mutated region of *TMPRSS2* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Antisense-9

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagGAGACAGCCAA
TCCTGCTGAGGGACGCGTGGGCTCATCTTGGAAGTCTGTCCATAGTCGCTGGAGGAGGACG
CGGGCCGCTCACCTGCCGCGCTCCAGGCGGCGCTCCCCGCCCCTCGCCCTCCGCCagacag
gagtgagagatggaagctcgcgcctactc

: *TMPRSS2* : *ERG* : Mutated region of *TMPRSS2* **+1**: Transcription start ctgcag: PstI

Antisense-5A

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagGCCGCTCACCT
GCCGCGCTCCAGGCGGCGCTCCCCGCCCCTCGCCCTCCGCCagacaggagtgagagatgga
agctcgcgcctactc

: *TMPRSS2* : Mutated region of *TMPRSS2* **+1**: Transcription start ctgcag: PstI

Antisense-5B

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCTTGAGCCATT
CACCTGGCTAGGGTTACATTCCATATTGATGGTGACCCTGGCTGGGGGTTGAGACAGCCAA
TCC

: *ERG* **+1**: Transcription start ctgcag: PstI A: T to A change to inactivate the cryptic transcription termination by U6 promoter

Antisense-B1

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagTGGCCTGAGCC
TTGAAGAATGGGGTGTACTGGGTAAATCAAAATGGTGGGGAGCATTTCCAGATGGAGAAAC
TGCCTCACCTGCCGCGCTCCAGGCGGCGCTCCCCGCCCCTCGCCCTCCGCCTCCG

: *TMPRSS2* : *ERG* **+1**: Transcription start ctgcag: PstI

Antisense-B2

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagTGGCCTGAGCC
TTGAAGAATGGGGTGTACTGGGTAAATCAAAATGGTGGGGAGCATTTCCAGATGGAGAAAC
TGCCCCTCCCCAGCCCGGACCCCGAGCCGGGACCCTGGTACCGGCGCCGCTCACC

: *TMPRSS2* : *ERG* **+1**: Transcription start ctgcag: PstI

Antisense-B3

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagTGGGGAGCATT
TCCAGATGGAGAAACTGCAAGGAAAAGCATAGAAGTGGGGCCACCCCTCGTGAGCTGGGGA
GGGCTCACCTGCCGCGCTCCAGGCGGCGCTCCCCGCCCCTCGCCCTCCGCCTCCG

: *TMPRSS2* : *ERG* **+1**: Transcription start ctgcag: PstI

Antisense-C1

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagTGAGCTCATGC
TATTCCTATGACATAGATGAGCACTGGGTAGACCCCGTCCTGGTAACACTATTCATGCACT
AACCCCAGGCGGGGGCCGTGGAGGGCAGGCGGACTAGGAGCCAGCTTTGGGGACC

: *TMPRSS2* : *ERG* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Antisense-C2

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagTGAGCTCATGC
TATTCCTATGACATAGATGAGCACTGGGTAGACCCCGTCCTGGTAACACTATTCATGCACT
AACCAGCACTCCCAGTCCTCCTCCCCAAAGAGAAAAGGCGCACCGGTGCTCCCAG

: *TMPRSS2* : *ERG* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Antisense-C3

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagGTCCTGGTAAC
ACTATTCATGCACTAACAAGTTGGTTGCCAGTGAGACTTGATTATTATGACTCTGGGAGTG
CTGCCCAGGCGGGGGCCGTGGAGGGCAGGCGGACTAGGAGCCAGCTTTGGGGACC

: *TMPRSS2* : *ERG* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Antisense-D1

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagGAGAGACAGAG
AGAGAGAGGCTGGTAGAGGGAAGAGACAGAAGAAAGATGAAGGGATAAGTGTCCAGAATCC
CTGAGCGCTCGACCCTCGGGCGCACTCACCTGCCGCGCCGCGCTCCTCACACCCG

: *TMPRSS2* : *ERG* **+1**: Transcription start ctgcag: PstI

Antisense-D2

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagGAGAGACAGAG
AGAGAGAGGCTGGTAGAGGGAAGAGACAGAAGAAAGATGAAGGGATAAGTGTCCAGAATCC
CTGCCCAGCACTCTCCCAGCACCCCGGGAGGCGCCCTGCCCGGCTGGCCCCAGCG

: *TMPRSS2* : *ERG* **+1**: Transcription start ctgcag: PstI

Antisense-D3

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagTGAAGGGATAA
GTGTCCAGAATCCCTGGATCTGGGATGGAATAAAGGATCTGGATGGTAAACGGAGAGTGCT
GGGAGCGCTCGACCCTCGGGCGCACTCACCTGCCGCGCCGCGCTCCTCACACCCG

: *TMPRSS2* : *ERG* **+1**: Transcription start ctgcag: PstI

Antisense-E1

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagAGGGTATTCAG
TATTACTATTTGGCTTAGATAAGCTGGTAGTTACTTGCTAAAATTAATCTTTATTATAAAG
CAGAAAGCCGCATTCTGACATCACTCTCCATGGACAAAGATTCTTCGCTTGATCA

: *TMPRSS2* : *ERG* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Antisense-F1

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagAATGGATGAAT
ACATAAAATAAATTGTGGTGGATATATACAACGGAATATCATTTAGCCTTTATTATTATTG
AGAACTGTGCCGAGCCGGGCAGGACAGGATGAGGTGGACCGAAGCGCCCAGGTGC

: *TMPRSS2* : *ERG* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Antisense-G1

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagTTTATTCATGA
GAAAATTCTGGCACAATGGAAAACCCTGGCAAGCAAAAGATAGGGGCAGCAGATGTCCTGG
CCTACAAAGAACTCCAAGCCCATCGTCCCTAGAAAGCATGGTCTCCCATGACCCC

: *TMPRSS2* : *ERG* **+1**: Transcription start ctgcag: PstI

Sense-3

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcaggagtaggcgcg
agcttccatctctcactcctgtctGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCG
CGGCAGGTGAGCGGCGCCGGTACCAGGGTCCCGGCTCGGGGTCCGGGATTGGCTGTCTCAA
CCCCCAGCCAGGGTCACCATCAAAATGGAATGTAACCCTAGCCAGGTGAATGGCTCAAG

: *TMPRSS2* : *ERG* : Mutated region of *TMPRSS2* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Sense-4

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcaggagtaggcgcg
agcttccatctctcactcctgtctGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCG
CGGCAGGTGAGCGGCGCCGGTACCAGGGTCGGATTGGCTGTCTCAACCCCCAGCCAGGGTC
ACCATCAATATGGAATGTAACCCTAGCCAGGTGAATGGCTCAAG

: *TMPRSS2* : *ERG* : Mutated region of *TMPRSS2* **+1**: Transcription start ctgcag: PstI

Sense-5

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcaggagtaggcgcg
agcttccatctctcactcctgtctGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGAGCG
CGGCAGGTGAGCGGCGGATTGGCTGTCTCAACCCCCAGCCAGGGTCACCATCAAAATGGAA
TGTAACCCTAGCCAGGTGAATGGCTCAAG

: *TMPRSS2* : *ERG* : Mutated region of *TMPRSS2* **+1**: Transcription start ctgcag: PstI

Sense-6

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcaggagtaggcgcg
agcttccatctctcactcctgtctGGCGGAGGGCGAGGGGCGGGGAGCGCCGCCTGGGGAT
TGGCTGTCTCAACCCCCAGCCAGGGTCACCATCAAAATGGAATGTAACCCTAGCCAGGTGA
ATGGCTCAAG

: *TMPRSS2* : *ERG* : Mutated region of *TMPRSS2* **+1**: Transcription start ctgcag: PstI

Sense-B1

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCGGAGGCGGAG
GGCGAGGGGCGGGGAGCGCCGCCTGGAGCGCGGCAGGTGAGGCAGTTTCTCCATCTGGAAA
TGCTCCCCACCATATTGATTTACCCAGTACACCCCATTCTTCAAGGCTCAGGCCA

: *TMPRSS2* : *ERG* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Sense-C1

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagGGTCCCCAAAG
CTGGCTCCTAGTCCGCCTGCCCTCCACGGCCCCCGCCTGGGGTTAGTGCATGAAGAGTGTT
ACCAGGACGGGGTCTACCCAGTGCTCATCTATGTCATAGGAAGAGCATGAGCTCA

: *TMPRSS2* : *ERG* **+1**: Transcription start ctgcag: PstI

Sense-D1

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCGGGTGTGAGG
AGCGCGGCGCGGCAGGTGAGTGCGCCCGAGGGTCGAGCGCTCAGGGATTCTGGACACTTAT
CCCTTCATCTTTCTTCTGTCTCTTCCCTCTACCAGCCTCTCTCTCTCTGTCTCTC

: *TMPRSS2* : *ERG* **+1**: Transcription start promoter ctgcag: PstI

Antisense-TETV-1

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagTGCCGCATTAT
GTAAATCGTTCCAAGTTAAAGTCTTAGTTAGATTCAGTAGACTAGGAGCCAGCTTTGGGGA
CCCCGGGGGACTCTCTTCCACCAACTGG

: *TMPRSS2* : *ETV1* **+1**: Transcription start ctgcag: PstI

Antisense- TETV-2

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagGATTCAGTAAT
TTCAAAGTTTATTATATTTAAGATAAGACTGAAGTGCTCAACAAACTTAGTCTCACTTTAG
GTATTCCAAATGCCTTGTAACTGGGCTG

: *TMPRSS2* : *ETV1* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Antisense- TETV-3

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCTCTGAATAGA
AAAATAGAAGTCCATAGTATCAACTCTAATATTCATATTTGGCTGCATCCCCACTTCCTGG
AGTACCTTCCCAGATCTCCTGGGACAGG

: *TMPRSS2* : *ETV1* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Antisense- TETV-4

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagATATAAACAAA
AAGTGTCAGCATTTGTCTCAACTTCATTCTATTCAATGTAAGGCCCTTTGCGCTGGTAAAC
TCTCCCTGCCACACTCCCAACCCCCATC

: *TMPRSS2* : *ETV1* **+1**: Transcription start ctgcag: PstI

Antisense- TETV-5

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCTCTGAATAGA
AAAATAGAAGTCCATAGTATCAACTCTAATATTCATATTCTTCAGCAACCAAAACTGAACA
AGCACTCCATTGACCATTCACCTTTCCT

: *TMPRSS2* : *ETV1* **+1**: Transcription start A: T to A change to inactivate the cryptic transcription termination by U6 promoter ctgcag: PstI

Antisense- TETV-6

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagAGGAGGCCTTA
AGTATGATTCAGTGAACACATTATGGTCGATAAACAAGGTGGGCCCCATTCTCAGAGTCTG
ATGTAATAATTGGGACCAAGGCAATGA

: *TMPRSS2* : *ETV1* **+1**: Transcription start ctgcag: PstI

Antisense- TETV-7

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagTGTAATTGACT
TAGATCTTGAAAGAGTTCTAAAAAACAAGTCAAAGACATCTAGAAGAATCTCTAGATGAAG
GTTACCTACAACAAAGACCAGTGTTGCC

: *TMPRSS2* : *ETV1* **+1**: Transcription start ctgcag: PstI

Antisense- TETV-8

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagTTAACAAACAG
CTTAATAAATAAGCTCAGGGATACCAGAATTCACAAAAAGAAGCACTCTCCTCTGGGATCA
GAGTGGGTAGGAGGATGGGGTGCAATTG

: *TMPRSS2* : *ETV1* **+1**: Transcription start ctgcag: PstI

Sense- TETV-1

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCCAGTTGGTGG
AAGAGAGTCCCCCGGGGTCCCCAAAGCTGGCTCCTAGTCTACTGAATCTAACTAAGACTTT
AACTTGGAACGATTTACATAATGCGGCA

: *TMPRSS2* : *ETV1* **+1**: Transcription start ctgcag: PstI

Sense *AZI1* ex16-17

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCTGATTGAGGAC
AAGAAGGTCCTGAGTGAAAAGTGCGAGGCTGTGGTGGCCGAGCTGAAGCAGGAGGACCAGA
GATGCACCGAGCGTGTGGCCCAGGCACAGGCGCAGCACGAGCTGGAGATTAAAAAACTCAA
AGAATTAATGAGCGCCACCGAGAAAGCCCGCCGGGAGAAGTGGATCAGTGAGAAAACCAAG
AAGATCAAGGAGGTCACTGTCCGAG

: *AZI1 ex16* : *AZI1 ex17* **+1**: Transcription start ctgcag: PstI

Antiparallel *AZI1* ex16-17

+1tgtgctcgcttcggcagcacatatactaacattggaacgatcctgcagCTCGGACAGTGA
CCTCCTTGATCTTCTTGGTTTTCTCACTGATCCACTTCTCCCGGCGGGCTTTCTCGGTGGC
GCTCATTAATTCTTTGAGTTTTTTAATCTCCAGCTCGTGCTGCGCCTGTGCCTGGGCCACA
CGCTCGGTGCATCTCTGGTCCTCCTGCTTCAGCTCGGCCACCACAGCCTCGCACTTTTCAC
TCAGGACCTTCTTGTCCTCAATCAG

: *AZI1 ex16* : *AZI1 ex17* **+1**: Transcription start ctgcag: PstI

List of primers used

RT-PCR primers for amplifying induced fusion RNAs:

Induced fusion RNA (*TMPRSS2* ex1-*ERG* ex7)	
TMPRSS2 ex-1 F1	5'- TAGGCGCGAGCTAAGCAGGAG-3'
ERG ex-7 R1	5'- TAGCATGCATTAACCGTGGAGA-3'
Induced fusion RNA (*TMPRSS2* ex1-*ERG* ex4)	
TMPRSS2 ex-1 F1	5'- TAGGCGCGAGCTAAGCAGGAG -3'
ERG ex-4 R1	5'- CTTGAGCCATTCACCTGGCTAG-3'
Induced fusion RNA (*TMPRSS2* ex1-*ETV* ex5)	
TMPRSS2 ex-1 F1	5'- TAGGCGCGAGCTAAGCAGGAG-3'
ETV1 ex-6 R1	5'- TTCTTGACTGCAGGCAGAGCT -3'
TMPRSS2 ex-1 F2	5'- CAGGAGGCGGAGGCGGA-3'
ETV1 ex-5 R1	5'- CTTTCAGCCTGATAGTCTGGT-3'
TMPRSS2 ex-1 F3	5'- CGGAGGGCGAGGGGCGGGGA-3'
ETV1 ex-5 R2	5'- AACTGCTCATCATTGTCAGGT-3'

Primers used in three-round PCR for amplifying induced fusion RNAs:

1st round PCR	
TMPRSS2 ex-1 F1	5'- TAGGCGCGAGCTAAGCAGGAG -3'
ERG ex-4 R1	5'- CTTGAGCCATTCACCTGGCTAG -3'
2nd round PCR	
TMPRSS2 ex-1 F2	5'- CAGGAGGCGGAGGCGGA -3'
ERG ex-4 R2	5'- TGACCCTGGCTGGGGGTTGAGA -3'
3rd round PCR	

TMPRSS2 ex-1 F3	5'- CGGAGGGCGAGGGGCGGGGA -3'
ERG ex-4 R3	5'- TCCTGCTGAGGGACGCGTGG -3'

RT-PCR primers for amplifying endogenous parental mRNAs:

***TMPRSS2* parental mRNA (*TMPRSS2* ex2-*TMPRSS2* ex4)**	
TMPRSS2 ex-2 F1	5'- GTCATATTGAACATTCCAGA-3'
TMPRSS2 ex-4 R1	5'- GCGCAGCTCCCACGAGGAAGGT-3'
***ERG* parental mRNA (*ERG* ex3-*ERG* ex7): one-round PCR**	
ERG ex-3 F1	5'- CAGGTTCTGAACAGCTGGTA-3'
ERG ex-7 R1	5'- TAGCATGCATTAACCGTGGAGA-3'
ERG ex-1 F1	5'- CCCCCGAGGGACATGAGAGAA-3'
ERG ex-4 R1	5'- TGGGGGTTGAGACAGCCAAT-3'
***ERG* parental mRNA (*ERG* ex3-*ERG* ex7): three-round PCR**	
ERG ex-3 F1	5'- CAGGTTCTGAACAGCTGGTA-3'
ERG ex-7 R1	5'- TAGCATGCATTAACCGTGGAGA-3'
ERG ex-3 F2	5'- TGGGCTGGCTTACTGAAGGA-3'
ERG ex-7 R2	5'- TTGTAAGGCTTTATCAACAT-3'
ERG ex-3 F3	5'- ATGATTCAGACTGTCCCGGA-3'
ERG ex-7 R3	5'- CATCTGAAGTCAAATGTGGA-3'
***ERG* parental mRNA (*ERG* ex1-*ERG* ex4): three-round PCR**	
ERG ex-1 F1	5'- CCCCCGAGGGACATGAGAGAA-3'
ERG ex-4 R1	5'- TGGGGGTTGAGACAGCCAAT-3'
ERG ex-1 F2	5'- AGGGACATGAGAGAAGAGGA-3'
ERG ex-4 R2	5'- TGAGGGACGCGTGGGCTCAT-3'
ERG ex-1 F3	5'- GAGAGAAGAGGAGCGGCGCT-3'
ERG ex-4 R3	5'- CTTGGAAGTCTGTCCATAGT-3'

RT-PCR primers for amplifying input RNAs:

Sense-1	
Sense 1 F1	5’- ACGGAATTCGAGTAGGCGCGAGCTAAGCA-3’
Sense 1 R1	5’- TAGAAGCTTCTTGAGCCATTCACCTGGCT-3’
Sense-2	
Sense 2 F1	5’- ACGCTGCAGGAGTAGGCGCGAGCTAAGCA-3’
Sense 2 R1	5’- TAGAAGCTTAAAAAACTTGAGCCATTCACCTGGCT-3’
Sense-2 long	
Sense 2 long F1	5’- ACGCTGCAGGAGTAGGCGCGAGCTAAGCA-3’
Sense 2 long R1	5’- TAGAAGCTTAAAAAATAAGGCTTTATCAACATCAT-3’
Antisense-1	
Antisense 1 F1	5’- ACGGAATTCCTTGAGCCATTCACCTGGCT-3’
Antisense 1 R1	5’- TAGAAGCTTGAGTAGGCGCGAGCTAAGCA-3’
Antisense-2	
Antisense 2 F1	5’- ACGCTGCAGCTTGAGCCATTCACCTGGCTAGGGTT-3’
Antisense 2 R1	5’- TAGAAGCTTAAAAAAGAGTAGGCGCGAGCTAAGCA-3’
Antisense-3	
Antisense 3 F1	5’- ACGCTGCAGCTTGAGCCATTCACCTGGCTAGGGTT-3’
Antisense 3 R1	5’- TAGAAGCTTAAAAAAGAGTAGGCGCGAGCTTCCAT-3’
Antisense-4	
Antisense 4 F1	5’- ACGCTGCAGCTTGAGCCATTCACCTGGCTAGGGTT-3’
Antisense 4 R1	5’- TAGAAGCTTAAAAAAGAGTAGGCGCGAGCTTCCAT-3’
Antisense-5	
Antisense 5 F1	5’- ACGCTGCAGCTTGAGCCATTCACCTGGCTAGGGTT-3’

Antisense 5 R1	5’- TAGAAGCTTAAAAAAGAGTAGGCGCGAGCTTCCAT-3’
Antisense-6	
Antisense 6 F1	5’- ACGCTGCAGCTTGAGCCATTCACCTGGCTAGGGTT-3’
Antisense 6 R1	5’- TAGAAGCTTAAAAAAGAGTAGGCGCGAGCTTCCAT-3’
Antisense-7	
Antisense 7 F1	5’- ACGCTGCAGCTTGAGCCATTCACCTGGCTAGGGTT-3’
Antisense 7 R1	5’- TAGAAGCTTAAAAAAGAGTAGGCGCGAGCTTCCAT-3’
Antisense-8	
Antisense 8 F1	5’- ACGCTGCAGCTTGAGCCATTCACCTGGCTAGGGTT-3’
Antisense 8 R1	5’- TAGAAGCTTAAAAAAGAGTAGGCGCGAGCTTCCAT-3’
Antisense-9	
Antisense 9 F1	5’- ACGCTGCAGGAGACAGCCAATCCTGCTGAGGGACGCGTGGGC-3’
Antisense 9 R1	5’- TAGAAGCTTAAAAAAGAGTAGGCGCGAGCTTCCAT-3’
Antisense-5A	
Antisense 5A F1	5’- TAGAAGCTTAAAAAAGAGTAGGCGCGAGCTTCCAT-3’
Antisense 5A R1	5’- TAGCTGCAGGCCGCTCACCTGCCGCGCTCC-3’
Antisense-5B	
Antisense 5B F1	5’- ACGAAGCTTAAAAAAGGATTGGCTGTCTCAACCCCCA-3’
Antisense 5B R1	5’- ACGCTGCAGCTTGAGCCATTCACCTGGCTAGGGTT-3’
Antisense-B1	
Antisense B1 F1	5’-TAGCTGCAGTGGCCTGAGCCTTGAAGAAT-3’
Antisense B1 R1	5’-ACGAAGCTTAAAAAACGGAGGCGGAGGGCGAGGGG-3’
Antisense-B2	
Antisense B2 F1	5’- TAGCTGCAGTGGCCTGAGCCTTGAAGAAT-3’
Antisense B2 R1	5’- ACGAAGCTTAAAAAAGGTGAGCGGCGCCGGTAC-3’
Antisense-B3	

Antisense B3 F1	5’- TAGCTGCAGTGGGGAGCATTTCCAGATGGAGAAACTGCAAGG-3’
Antisense B3 R1	5’- ACGAAGCTTAAAAAACGGAGGCGGAGGGCGAGGG-3’
Antisense-C1	
Antisense C1 F1	5’- TAGCTGCAGTGAGCTCATGCTATTCCTATGA-3’
Antisense C1 R1	5’- ACGAAGCTTAAAAAAGGTCCCCAAAGCTGGCTCCT-3’
Antisense-C2	
Antisense C2 F1	5’- TAGCTGCAGTGAGCTCATGCTATTCCTATGA -3’
Antisense C2 R1	5’- ACGAAGCTTAAAAAACTGGGAGCACCGGTGCGC-3’
Antisense-C3	
Antisense C3 F1	5’- TAGCTGCAGGTCCTGGTAACACTATTCATGCACTAACAAGTT-3’
Antisense C3 R1	5’- ACGAAGCTTAAAAAAGGTCCCCAAAGCTGGCTCCT-3’
Antisense-D1	
Antisense D1 F1	5’-TAGCTGCAGGAGAGACAGAGAGAGAGAGG-3’
Antisense D1 R1	5’- ACGAAGCTTAAAAAACGGGTGTGAGGAGCGCGGCG-3’
Antisense-D2	
Antisense D2 F1	5’- TAGCTGCAGGAGAGACAGAGAGAGAGAGG-3’
Antisense D2 R1	5’- ACGAAGCTTAAAAAACGCTGGGGCCAGCCGGGC-3’
Antisense-D3	
Antisense D3 F1	5’- TAGCTGCAGTGAAGGGATAAGTGTCCAGAATCCCTGGATCTG-3’
Antisense D3 R1	5’- ACGAAGCTTAAAAAACGGGTGTGAGGAGCGCGGCG-3’
Antisense-E1	
Antisense E1 F1	5’- TAGCTGCAGAGGGTATTCAGTATTACTATTT-3’
Antisense E1 R1	5’- ACGAAGCTTAAAAAATGATCAAGCGAAGAATCTTTGTCCATG-3’
Antisense-F1	
Antisense F1 F1	5’- TAGCTGCAGAATGGATGAATACATAAAATAA-3’
Antisense F1 R1	5’- ACGAAGCTTAAAAAAGCACCTGGGCGCTTCGGTCCACCTCAT-3’

Antisense-G1	
Antisense G1 F1	5’- TAGCTGCAGTTTATTCATGAGAAAATTCTGG-3’
Antisense G1 R1	5’- ACGAAGCTTAAAAAAGGGGTCATGGGAGACCATGCTTTCTAG-3’
Sense-3	
Sense 3 F1	5’- ACGCTGCAGGAGTAGGCGCGAGCTTCCAT-3’
Sense 3 R1	5’- TAGAAGCTTAAAAAACTTGAGCCATTCACCTGGCT-3’
Sense-4	
Sense 4 F1	5’- ACGCTGCAGGAGTAGGCGCGAGCTTCCAT-3’
Sense 4 R1	5’- TAGAAGCTTAAAAAACTTGAGCCATTCACCTGGCT-3’
Sense-5	
Sense 5 F1	5’- ACGCTGCAGGAGTAGGCGCGAGCTTCCAT-3’
Sense 5 R1	5’- TAGAAGCTTAAAAAACTTGAGCCATTCACCTGGCT-3’
Sense-6	
Sense 6 F1	5’- ACGCTGCAGGAGTAGGCGCGAGCTTCCAT-3’
Sense 6 R1	5’- TAGAAGCTTAAAAAACTTGAGCCATTCACCTGGCT-3’
Sense-B1	
Sense B1 F1	5’- TAGCTGCAGCGGAGGCGGAGGGCGAGGGG-3’
Sense B1 R1	5’- ACGAAGCTTAAAAAATGGCCTGAGCCTTGAAGAAT-3’
Sense-C1	
Sense C1 F1	5’- TAGCTGCAGGGTCCCCAAAGCTGGCTCCT-3’
Sense C1 R1	5’- ACGAAGCTTAAAAAATGAGCTCATGCTCTTCCTAT-3’
Sense-D1	
Sense D1 F1	5’- TAGCTGCAGCGGGTGTGAGGAGCGCGGCG-3’
Sense D1 R1	5’- ACGAAGCTTAAAAAAGAGAGACAGAGAGAGAGAGG-3’
Antisense *TETV*-1	
Antisense *TETV*-	5’- TAGCTGCAGTGCCGCATTATGTAAATCGTTCCAAGTTAAAGTC-3’

1 F1	
Antisense *TETV*-1 R1	5’- ACGAAGCTTAAAAAACCAGTTGGTGGAAGAGAGTCCCCCGGG-3’
Antisense *TETV*-2	
Antisense *TETV*-2 F1	5’- TAGCTGCAGGATTCAGTAATTTCAAAGTTTATTATATTTAAGA-3’
Antisense *TETV*-2 R1	5’- ACGAAGCTTAAAAAACAGCCCAGTTACAAGGCATTTGGAATA-3’
Antisense *TETV*-3	
Antisense *TETV*-3 F1	5’- TAGCTGCAGCTCTGAATAGAAAAATAGAAGTCCATAGTATCAA-3’
Antisense *TETV*-3 R1	5’- ACGAAGCTTAAAAAACCTGTCCCAGGAGATCTGGGAAGGTAC-3’
Antisense *TETV*-4	
Antisense *TETV*-4 F1	5’- TAGCTGCAGATATAAACAAAAAGTGTCAGCATTTGTCTCAACT-3’
Antisense *TETV*-4 R1	5’- ACGAAGCTTAAAAAAGATGGGGGTTGGGAGTGTGGCAGGGAG-3’
Antisense *TETV*-5	
Antisense *TETV*-5 F1	5’- TAGCTGCAGCTCTGAATAGAAAAATAGAAGTCCATAGTATCAA-3’
Antisense *TETV*-5 R1	5’- ACGAAGCTTAAAAAAAGGAAAGGTGAATGGTCAATGGAGTGC-3’
Antisense *TETV*-6	
Antisense *TETV*-6 F1	5’- TAGCTGCAGAGGAGGCCTTAAGTATGATTCAGTGAACACATTA-3’
Antisense *TETV*-6 R1	5’- ACGAAGCTTAAAAAAATCATTGCCTTGGTCCCAATTATTACA-3’
Antisense *TETV*-7	
Antisense *TETV*-	5’- TAGCTGCAGTGTAATTGACTTAGATCTTGAAAGAGTTCTAAAA-

7 F1	3’
Antisense *TETV*-7 R1	5’- ACGAAGCTTAAAAAAGGCAACACTGGTCTTTGTTGTAGGTAA-3’
Antisense *TETV*-8	
Antisense *TETV*-8 F1	5’- TAGCTGCAGTTAACAAACAGCTTAATAAATAAGCTCAGGGATA-3’
Antisense *TETV*-8 R1	5’- ACGAAGCTTAAAAAACAATTGCACCCCATCCTCCTACCCACT-3’
Sense *TETV*-1	
Sense *TETV*-1 F1	5’- TAGCTGCAGCCAGTTGGTGGAAGAGAGT-3’
Sense *TETV*-1 R1	5’- ACGAAGCTTAAAAAATGCCGCATTATGTAAATC-3’

PCR primers used for amplifying the identified TMPRSS2-ERG genomic DNA breakpoint:

***TMPRSS2* intron 1**	
TMPRSS2 genomic bk F1	5’- ATGTGATATTAGTGCGGTTA -3’
TMPRSS2 genomic bk F2	5’- GGCTGGGATGTGTCCGTGGA -3’
***ERG* intron 3**	
ERG genomic bk R1	5’- CCATAAGTTTTACTGCGTCT-3’
ERG genomic bk R2	5’- GATACTGAGTGGTAAATTCT-3’

The rest of PCR primers used for genomic breakpoint analyses are not listed, but their locations are shown in Fig. S15A.

PCR primers used for amplifying the identified TMPRSS2-ETV1 genomic DNA breakpoint:

TMPRSS2	
TMPRSS2 genomic bk pt A1	5'- TCTGCTCGAGCACGGGTCCA-3'
TMPRSS2 genomic bk pt A2	5'- AAAACTGCCCCATGTCCAG-3'
TMPRSS2 genomic bk pt B1	5'- CAACCTGGGAGGCCCTGCCT-3'
TMPRSS2 genomic bk pt B2	5'- CAGCAACAGCACAAGCTTGT-3'
TMPRSS2 genomic bk pt C1	5'- GGCTGGGATGTGTCCGTGGA-3'
TMPRSS2 genomic bk pt C2	5'- TGGTGGTGGTGCTGTCTGGA-3'
TMPRSS2 genomic bk pt D1	5'- CAGGAGAATCACTTGAACCT-3'
TMPRSS2 genomic bk pt D2	5'- AGGCCACTGCACTCCAGCCT-3'
***ERG* intron 3**	
ETV1 genomic bk pt M1	5'-GCAAGTCTCGTTGATCGCCA-3'
ETV1 genomic bk pt M2	5'-TTGCACACGTTTGCGAATCA-3'
ETV1 genomic bk pt N1	5'-AGGGAGAGTTGCTTCCCAGT-3'
ETV1 genomic bk pt N2	5'-GCCGATCTTAGCACATTACT-3'

PCR primers used for amplifying the AZI1 RNA

AZI1 ex15-ex18	
AZI1 ex15 F1	5'- AAGGAGACAGAGAAGGCGCT -3'
AZI1 ex18 R1	5'- CTCCTGCTGGCCCAGCGCCT -3'

PCR primers used for amplifying the GAPDH RNA

GAPDH	
GAPDH F1	5'- GCGTCTTCACCACCATGGAGA -3'
GAPDH R1	5'- AGCCTTGGCAGCGCCAGTAGA -3'

References for Supplementary Information

Kimsey, I., and Al-Hashimi, H.M. (2014). Increasing occurrences and functional roles for high energy purine-pyrimidine base-pairs in nucleic acids. Curr Opin Struct Biol *24*, 72-80.

Mertz, K.D., Setlur, S.R., Dhanasekaran, S.M., Demichelis, F., Perner, S., Tomlins, S., Tchinda, J., Laxman, B., Vessella, R.L., Beroukhim, R., *et al.* (2007). Molecular characterization of TMPRSS2-ERG gene fusion in the NCI-H660 prostate cancer cell line: a new perspective for an old model. Neoplasia *9*, 200-206.

Rychlik, W., and Rhoads, R.E. (1989). A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acids Res *17*, 8543-8551.

Teles Alves, I., Hiltemann, S., Hartjes, T., van der Spek, P., Stubbs, A., Trapman, J., and Jenster, G. (2013). Gene fusions by chromothripsis of chromosome 5q in the VCaP prostate cancer cell line. Hum Genet *132*, 709-713.

Watson, J.D., and Crick, F.H. (1953). Genetical implications of the structure of deoxyribonucleic acid. Nature *171*, 964-967.

Weier, C., Haffner, M.C., Mosbruger, T., Esopi, D.M., Hicks, J., Zheng, Q., Fedor, H., Isaacs, W.B., De Marzo, A.M., Nelson, W.G., *et al.* Nucleotide resolution analysis of TMPRSS2 and ERG rearrangements in prostate cancer. J Pathol *230*, 174-183.