



OPEN

Prostate cancer subtyping and differential methylation analysis based on the ETS family of transcription factors fusion genes

Wenkang Niu, Guifang Li, Tingting Zhang[✉] & Lei Ma[✉]

Prostate cancer (PCa) is a highly heterogeneous disease, encompassing various molecular and clinical pathological subtypes. Fusion genes play a facilitating role in the occurrence and progression of PCa. We categorized PCa samples into the ETS family of transcription factors fusion positive and fusion negative subtypes based on fusion genes. This subtyping method is closely related to the epigenomic DNA methylation profiles of PCa, with each sample cluster including more than 85% of the patients. We conducted an analysis of the distribution of the ETS family fusion genes on chromosomes, fusion modes within reading frames, and predictions of structural domains. Among these, the highest frequency of the ETS family related fusion genes occurred on chromosome 21. Compared to the parental genes, fusion genes exhibited new structural domains, such as IG_like, and the most common fusion mode was out-of-frame fusion. The correlation between the methylation levels of hypermethylated CpG sites and the expression levels of their corresponding mRNAs indicates that CD8A and B3GNT5 (with correlations of -0.388 and -0.253 , respectively) could serve as potential prognostic markers for PCa.

Keywords The ETS family of transcription factors, Fusion genes, DNA methylation, Prostate cancer

Prostate cancer (PCa) is a highly heterogeneous disease, encompassing various molecular and clinical pathological subtypes¹. In PCa, approximately 400 fusion genes have been identified, playing a promoting role in the occurrence and development of cancer. Among them, several serve as potential biomarkers or targets for personalized therapy². Increasing evidence suggests that cancer samples can be classified into fusion positive and fusion negative subtypes through the formation of fusion genes³. The most common fusion genes occur between the TMPRSS2 gene and members of the ETS family⁴. TMPRSS2-ERG fusion positive and fusion negative tumors represent specific molecular subtypes of PCa. DNA methylation profiling can accurately distinguish these major PCa subtypes, thus offering new therapeutic approaches for fusion positive PCa⁴.

The human ETS transcription factor family consists of 27 members⁵. Among the ETS family members, ETV1, ETV4, ERG, and FLI1 are the most frequently involved in fusion gene events⁶. These fusion genes, frequently caused by translocation events in PCa, may regulate the functions of other the ETS family members⁷. ETS family members play crucial roles in tumor cells, including involvement in processes such as development, differentiation, proliferation, apoptosis, migration, etc⁸.

DNA methylation is an important epigenetic modification mechanism that can influence the development of tumors by altering the expression of tumor suppressor genes and oncogenes⁹. Specifically, methylation of CpG islands can lead to gene silencing, rendering important genes such as DNA repair genes and tumor suppressor genes inactive, closely associated with the formation of cancer¹⁰. In PCa samples containing the fusion gene TMPRSS2-ERG, high methylation may result in the inhibition of miR-26a, leading to reduced EZH2 enzyme inhibition, further disrupting the overall DNA methylation profile¹¹. However, there is limited research on the subdivision of PCa DNA methylation patterns based on whether fusion genes involving the ETS family are present or not.

College of Life Science, Shihezi University, Shihezi City, Xinjiang, China. ✉email: zting@shzu.edu.cn; malei1979@hotmail.com

The present study aims to identify a novel subtyping of PCa based on the ETS family fusion positive and fusion negative categories. We conducted an analysis of the distribution of the ETS family fusion genes on chromosomes, fusion modes within reading frames, and predictions of structural domains for both fusion genes and their parental genes. Subsequently, we investigated the epigenomic DNA methylation patterns of PCa in different subgroups, including recurrent fusion positive and fusion negative, TMPRSS2-ERG fusion positive and fusion negative, and the ETS family fusion positive and fusion negative, to identify differentially methylated CpG sites. Finally, we integrated the methylation results with gene expression data from the same patient samples to explore the potential impact of distinct DNA methylation patterns on mRNA expression levels in PCa. These findings may provide new approaches for the treatment of PCa patients and offer new insights into personalized treatment.

Materials and methods

Publicly available data

The fusion gene data used in the present study were sourced from the ChimerDB4.0¹² database (<https://www.kobic.re.kr/chimerdb/download>). We obtained nucleotide and protein sequences for a total of 15,632 fusion genes associated with PCa from this database, and the data underwent deduplication. Additionally, we acquired data containing breakpoint information for 5,653 fusion genes. Methylation data related to PCa were obtained from the UCSC Xena¹³ database (<https://xenabrowser.net/datapages/>). Gene names and their corresponding methylation site ID/Gene Mapping data also originated from the UCSC Xena database.

Differentially expressed parental fusion genes

The RNA-seq dataset for PCa patients was downloaded from the TCGA database (<https://www.cancer.gov/>), which includes 481 PCa samples and 51 adjacent non-tumor tissue samples. The samples in this dataset correspond one-to-one with the fusion gene dataset samples from the ChimerDB4.0 database. Using our laboratory's previous research, we identified 764 genes differentially expressed between PCa and normal tissues¹⁴. We used the R package 'DESeq2' to screen for differentially expressed genes between PCa samples and normal tissue samples, with thresholds of $|\log_{2}FC| > 1$, $FDR < 0.01$, and $P < 0.05$. By intersecting with 5,653 fusion genes specific to PCa from the ChimerDB4.0 database (reduced to 1,399 after deduplication), we identified 165 differentially expressed fusion precursors (hereafter referred to as parental genes). Among them, 23 parental genes exhibited repetitive fusions (occurring 2 times or more). Additionally, we identified 106 fusion genes, including 45 with repetitive fusions. Additionally, we identified 106 fusion genes, of which 104 are Single parental differentially expressed fusion genes (SPDEFGs) and 2 are Dual parental differentially expressed fusion genes (DPDEFGs). Subsequently, we extracted chromosomal location data for these 106 fusion genes, obtaining a total of 1,334 loci information.

The distribution and fusion patterns of fusion genes on chromosomes

We utilized Circos (version 0.69–8) (<https://circos.ca/>) to illustrate the distribution of the previously obtained 106 fusion genes on chromosomes. Additionally, we conducted a statistical analysis of the reading frame fusion patterns for these 106 fusion genes and employed the 'ggplot2' R package to generate a circular plot for visualization.

Domain analysis of fusion genes

We used the SMART¹⁵ website (<http://smart.embl-heidelberg.de/>) for domain prediction. For parental genes, we initially obtained protein sequences for 165 parental genes. During this process, four genes (LPHN2, C4orf22, C16orf62, ARHGAP23) did not have corresponding protein sequences, so we ended up with protein sequences for 161 parental genes. For fusion genes, due to data limitations, we only queried genes with repetitive fusions. Among the 106 fusion genes, we found 59 pairs (comprising a total of 72 parental genes) with repetitive fusions. Among the 72 parental genes, the queried protein sequences include those of 22 recurrent fusion parental genes and 50 non-recurrent fusion parental genes. By retrieving protein sequences for these 59 pairs of genes with repetitive fusions, we ultimately obtained sequences for 23 pairs of fusion genes and predicted their structural domains.

Survival analysis of fusion positive and fusion negative PCa

We grouped all tumor samples containing fusion genes into recurrent fusion positive ($n = 82$) and recurrent fusion negative ($n = 399$) groups, then analyzed the correlation between overall survival (OS) and the recurrent fusion positive and recurrent fusion negative groups. Additionally, considering TMPRSS2-ERG as the most common fusion gene in PCa, we further categorized tumor into TMPRSS2-ERG fusion positive ($n = 186$) and TMPRSS2-ERG fusion negative ($n = 295$) groups based on its presence, and then analyzed the correlation between OS and the TMPRSS2-ERG fusion positive and fusion negative groups. Furthermore, based on whether tumor samples contained the ETS family fusion genes, we subdivided the samples into the ETS family fusion positive ($n = 227$) and the ETS family fusion negative ($n = 254$) groups, and subsequently analyzed the correlation between OS and the ETS family fusion positive and fusion negative groups.

Differential methylation expression and pathway enrichment analysis

We employed the above three survival analysis grouping methods to investigate DNA methylation differences in PCa. Differential expression analyses for these three grouping methods were performed using the 'limma' R package. In the differential expression analysis, we selected $p < 0.05$ & $|\log_{2}FC| > 0.5$ as the filtering criteria to obtain significantly differentially expressed CpG sites. Subsequently, we used the 'pheatmap' R package to create heatmaps separately, illustrating the results of differential expression. We performed KEGG pathway enrichment analysis on the mRNAs corresponding to differentially methylated CpG sites using the DAVID¹⁶ website

(<https://david.ncicrf.gov/>) to identify the important pathways these genes are involved in. Subsequently, using the R package 'ggplot2', we visualized the significantly enriched ($p < 0.05$) KEGG pathways with bubble charts.

Results

Distribution of fusion genes on and between chromosomes

In order to obtain information about the fusion patterns of fusion genes on chromosomes, we queried the fusion relationships of 106 fusion genes on and between chromosomes (Fig. 1A). Fusion events most commonly involve chromosomes 1, 4, 3, and 21. Overall, the frequency of interchromosomal fusion events was higher than that of intrachromosomal fusion events. Among them, fusion genes related to the ETS family had the highest occurrence frequency on chromosome 21, and the frequency of interchromosomal fusion events exceeded that of intrachromosomal fusion events (Fig. 1B).

Reading frame fusion patterns of fusion genes

We conducted a statistical analysis of the reading frame fusion patterns of the 106 fusion genes (Fig. 2A, C). Approximately one-third of fusion events occurred within the reading frame (36.79%, $n = 39$). This suggests the preservation of the original reading frame of the parental genes, indicating potential functional retention when translation occurs. Additionally, 26.40% of fusion events ($n = 28$) involved UTR (Untranslated Regions), which often contain regulatory elements that can alter the expression levels of adjacent genes. Finally, one-third of fusion events occurred outside the reading frame (36.79%, $n = 39$), potentially leading to gene functional degradation. Among fusion genes involving the ETS family, out-of-frame fusion was the most common reading frame fusion pattern (Fig. 2B).

Structural domains of fusion genes

To study the differences between the domains of recurrent fusion genes and their corresponding parental genes, we separately analyzed the domains of recurrent fusion genes and their parental genes. The DPDEFGs ERG-ABCC4 and OR51E2-AMACR are shown (Fig. 3A, B). For SPDEFGs, we identified a total of 23 pairs (Fig. 3C–G). Domain prediction was conducted for these 23 pairs of recurrent fusion genes and their parental genes, resulting in a total of 5 fusion types (with one fusion gene unable to find corresponding domains) (Supplementary Table S1):

- Retained partial domains of one parental gene ($n = 12$);
- Combined domains of both parental genes, resulting in new domains ($n = 2$);
- Retained the domains of both parental genes separately ($n = 3$);
- Neither of the parental genes' domains were retained ($n = 4$);
- Retained all domains of one parental gene ($n = 1$).

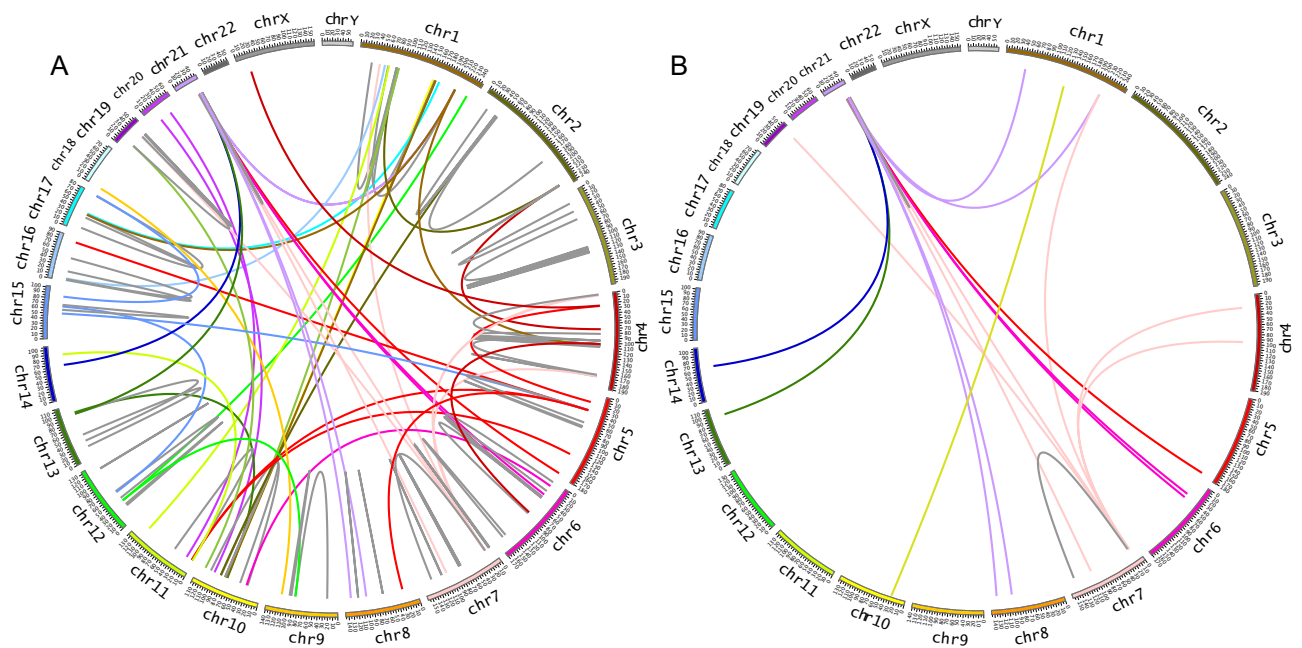


Fig. 1. Distribution of Fusion Genes on and between Chromosomes. **(A)** Distribution of 106 fusion genes on and between chromosomes; **(B)** Distribution of fusion genes involving the ETS family on and between chromosomes. Colored lines represent fusion events occurring between different chromosomes, while gray lines represent fusion events occurring within the same chromosome.

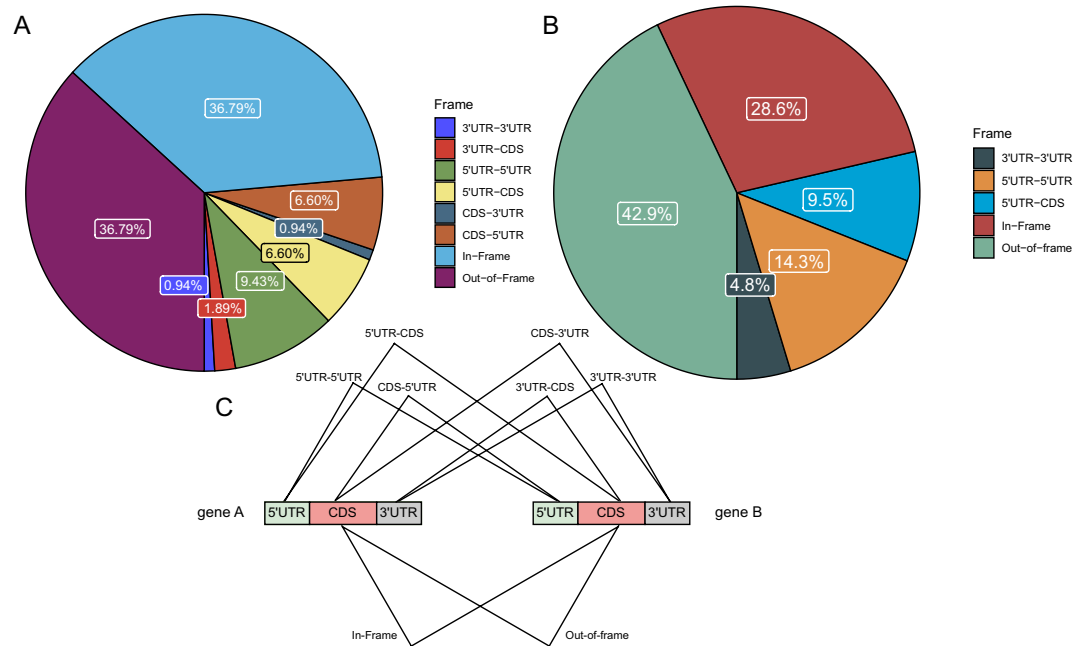


Fig. 2. Reading Frame Fusion Patterns of Fusion Genes. **(A)** Reading frame fusion patterns of 106 fusion genes; **(B)** Reading frame fusion patterns of fusion genes involving the ETS family; **(C)** Schematic representation of reading frame fusion patterns of fusion genes, where "In-Frame" indicates that the coding regions of gene A and gene B are connected within the same reading frame, avoiding frameshift mutations, while "Out-of-Frame" indicates that the coding regions of gene A and gene B are not within the same reading frame, potentially causing frameshift mutations.

In summary, the structural domains of fusion genes exhibited several scenarios: a portion involved the direct overlay of the structural domains of both parental genes, another portion entailed fusion genes retaining the domains of only one parental gene, and a further portion involved partial retention of the structural domains of both parental genes. Only a very small number of fusion genes introduced new domains on top of inheriting the original parental gene domains. Specifically, fusion genes with the ETS family genes as parental genes mostly retained partial domains of only one parental gene.

Survival analysis of fusion positive and fusion negative groups

To explore the relationship between fusion positive and fusion negative and overall survival (OS), we divided PCa samples into three groups for evaluation. Survival analysis of the recurrent fusion positive group ($n = 82$) and recurrent fusion negative group ($n = 399$) (Fig. 4A) revealed that, within the first 9 years of cancer diagnosis, the survival rate of the recurrent fusion negative group was higher than that of the recurrent fusion positive group ($p = 0.032$). Survival analysis conducted between the TMPRSS2-ERG fusion positive group ($n = 168$) and the TMPRSS2-ERG fusion negative group ($n = 295$) (Fig. 4B) revealed that the survival rate of the TMPRSS2-ERG fusion negative group was lower than that of the TMPRSS2-ERG fusion positive group ($p = 0.41$). For the ETS family fusion positive group ($n = 227$) and ETS family fusion negative group ($n = 254$), survival analysis results (Fig. 4C) revealed that within the first 9 years, the survival rate of the ETS family fusion positive group was higher than that of the ETS family fusion negative group ($p = 0.99$).

DNA methylation in fusion positive and fusion negative PCa tumor

To explore the DNA methylation status of parental fusion genes, we conducted differential expression analysis of methylated CpG sites in recurrent fusion positive and recurrent fusion negative PCa tumor based on differential groupings. Between the fusion positive and fusion negative groups, we identified a total of 183 differentially methylated CpG sites, with 12 upregulated and 171 downregulated (Fig. 5A) (Supplementary Table S2). Based on the methylation levels of these 183 differentially expressed CpG sites, the samples were clustered into two main clusters, with one cluster contained 37% of fusion positive tumor and the other cluster contained 86% of fusion negative tumor (Fig. 5B). These data suggested that the epigenetic characteristics of these 183 CpG sites could be used to distinguish between fusion positive and fusion negative PCa.

Similarly, based on the methylation differences between the TMPRSS2-ERG fusion positive group and TMPRSS2-ERG fusion negative group, we identified a total of 141 differentially methylated CpG sites, with 99 upregulated and 42 downregulated (Fig. 5C) (Supplementary Table S3). Based on the methylation levels of these 141 differentially expressed CpG sites, the samples were clustered into two main clusters, with one cluster contained 83% of TMPRSS2-ERG fusion positive tumor and the other cluster contained 89% of TMPRSS2-ERG

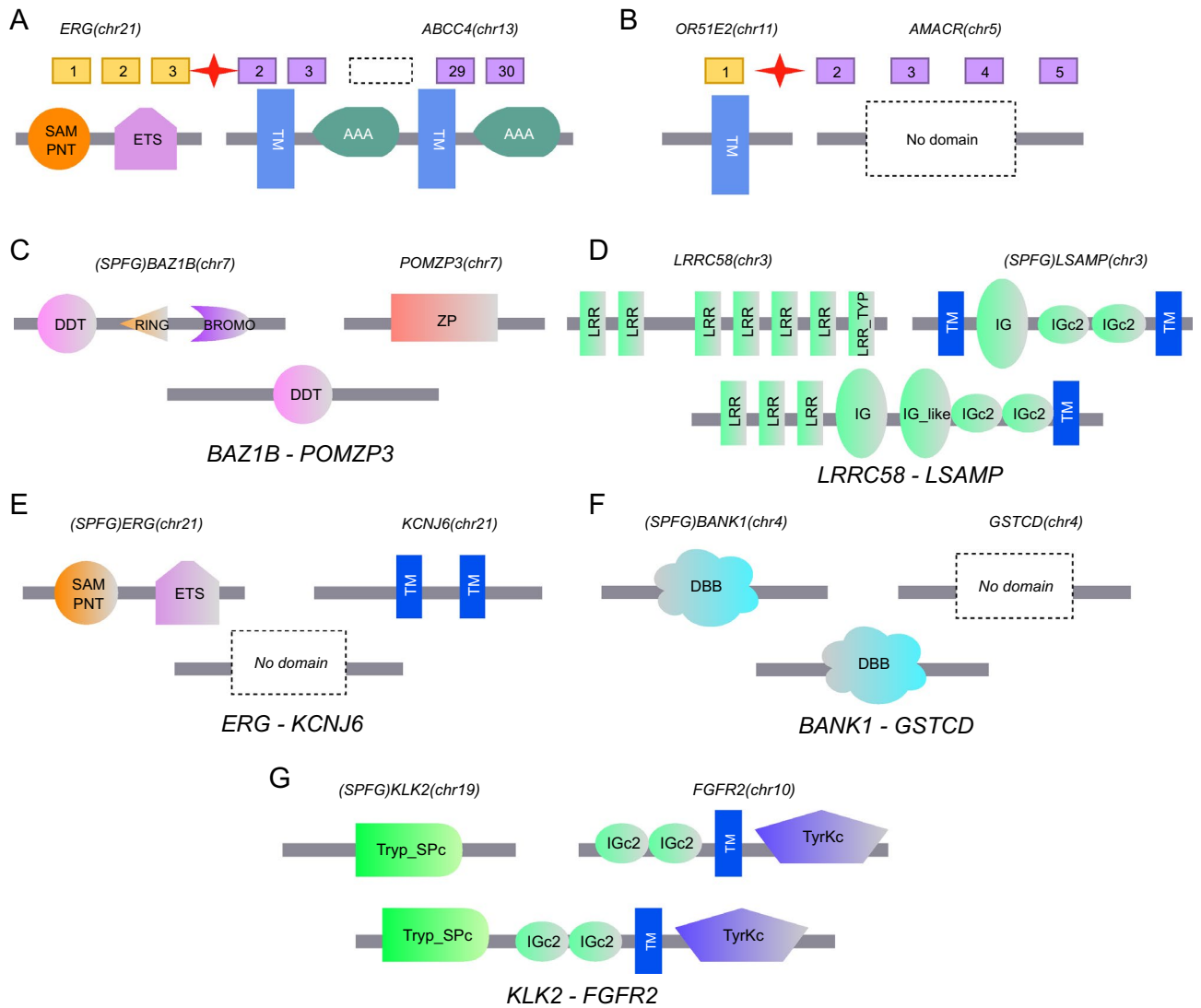


Fig. 3. Structural Domains of Fusion Genes. Each fusion gene structure is displayed with the gene name and corresponding chromosome above it (in parentheses). (A, B) Structural domains of DPDEFGs. Gray lines represent genes; boxes with numbers represent exons (5' gene: orange, 3' gene: purple); dashed boxes indicate exons not shown in the figure; red stars represent fusion points; protein domains are displayed below the exon boxes, with TM in the blue box indicating transmembrane domains. (C–G) Structural domains of SPDEFGs. Various shapes on the lines represent different domains. The SPDEFGs are: BAZ1B, LSAMP, ERG, BANK1, and KLK2.

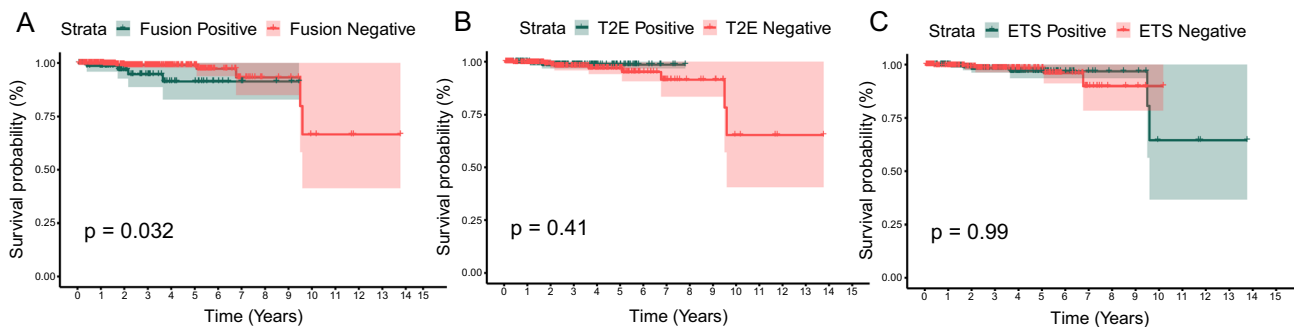


Fig. 4. Survival analysis of fusion positive and fusion negative groups in different subgroups. (A) Survival analysis of the recurrent fusion positive and recurrent fusion negative groups. (B) Survival analysis of T2E positive and T2E negative groups, where T2E represents the fusion gene T2E-ERG. (C) Survival analysis of the ETS family fusion positive and fusion negative groups.

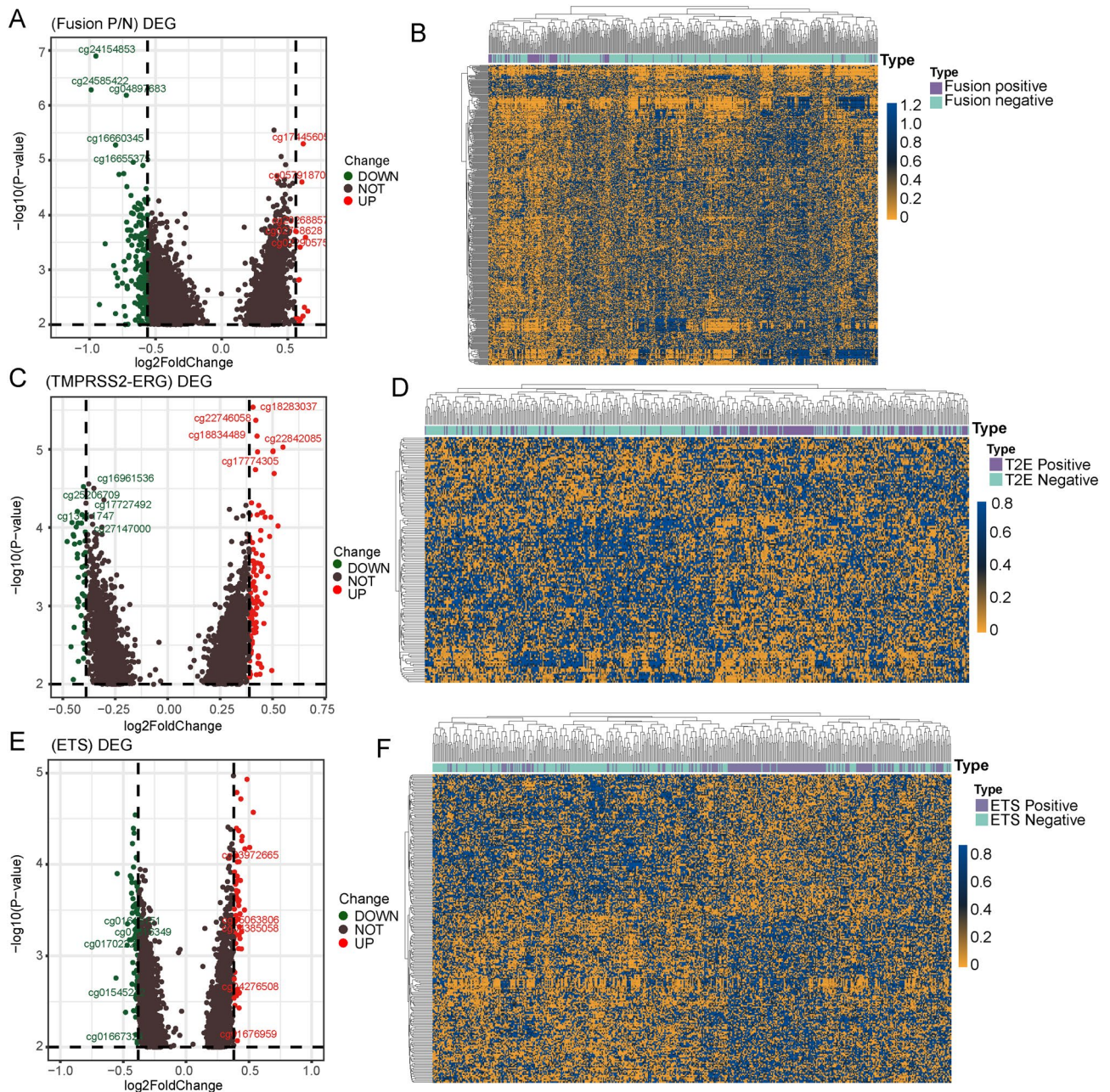


Fig. 5. DNA methylation in PCA tumor with different fusion status. **(A)** Volcano plot showing differential expression of CpG sites between recurrent fusion positive and recurrent fusion negative groups. **(B)** Hierarchical clustering heatmap of CpG methylation levels in recurrent fusion positive and recurrent fusion negative groups. **(C)** Volcano plot depicting differential expression of CpG sites between TMPRSS2-ERG fusion positive and fusion negative groups. **(D)** Hierarchical clustering heatmap of CpG methylation levels in TMPRSS2-ERG fusion positive and fusion negative groups. **(E)** Volcano plot illustrating differential expression of CpG sites between the ETS family fusion positive and fusion negative groups. **(F)** Hierarchical clustering heatmap of CpG methylation levels in the ETS family fusion positive and fusion negative groups.

fusion negative tumor (Fig. 5D). These data suggested that the epigenetic characteristics based on these 141 CpG sites could distinguish between TMPRSS2-ERG fusion positive and TMPRSS2-ERG fusion negative PCA.

In addition, based on the differential methylation analysis between the ETS family fusion positive group and the ETS family fusion negative group, we identified a total of 134 differentially methylated CpG sites, with 60 upregulated and 74 downregulated (Fig. 5E) (Supplementary Table S4). Based on the methylation levels of these 134 differentially expressed CpG sites, the samples were clustered into two main clusters, with one cluster containing 88% of the ETS family fusion positive tumor and the other cluster containing 89% of the ETS family fusion negative tumor (Fig. 5F). These data suggested that the epigenetic characteristics based on these 134 CpG sites could distinguish between the ETS family fusion positive and the ETS family fusion negative PCA.

Impact of DNA methylation on mRNA expression levels

We integrated DNA methylation results with mRNA expression data from the same PCa patient samples to investigate the potential impact of different DNA methylation on mRNA expression levels. Since we focused on ETS family fusion positive and fusion negative subtypes, we only analyzed the correlation between DNA methylation and mRNA expression levels in the ETS family fusion positive and fusion negative groups. Among the ETS family fusion negative and positive groups, we identified 134 CpG sites with significant methylation differences, which corresponded to the correlation between the expression levels of 93 mRNAs (Supplementary Fig. S1). We selected the top 10 CpG sites with the most significant differential expression based on p-values and demonstrated the correlation between the methylation levels of these sites and the expression levels of the corresponding 8 mRNAs (Fig. 6A–H). The methylation levels of most CpG sites showed a weak positive correlation with the corresponding mRNA expression levels. However, two hypermethylated CpG sites showed a strong negative correlation with their corresponding mRNA expression levels, namely, cg24345747 with CD8A showed a strong negative correlation ($\text{cor} = -0.388$, $p = 2.041e-18$) (Fig. 6B), and cg17701886 with B3GNT5 showed a strong negative correlation ($\text{cor} = -0.253$, $p = 1.946e-08$) (Fig. 6G).

Then, we performed KEGG pathway enrichment analysis on the 93 mRNAs corresponding to the 134 differentially methylated CpG sites (Supplementary Fig. S2). These genes were involved in several KEGG pathways including fatty acid metabolism, Notch signaling pathway, adrenergic signaling in cardiomyocytes, and the synthesis and degradation of ketone bodies, primarily related to metabolic processes and regulation of signal transduction.

Discussion

In this study, we proposed considering the ETS family gene fusions in PCa as two distinct subtypes, positive and negative, and found substantial differences in the DNA methylation profiles between these subtypes. We analyzed the distribution of fusion genes in PCa and ETS family fusion positive genes on chromosomes, the fusion modes of reading frames, and the structural domain predictions of the fusion genes and their parental genes, in order to reveal the role of ETS family fusion genes in gene structure and functional regulation. Subsequently, we investigated the epigenomic DNA methylation patterns in different subgroups, including PCa recurrent fusion positive and negative, TMPRSS2-ERG fusion positive and negative, and the ETS family fusion positive and negative subtypes. This exploration aimed to identify differentially methylated CpG sites, shedding light on the relationships between various subgroups and overall survival rates. The findings revealed an increasing trend in mortality rates for PCa tumor with recurrent fusion genes, those without TMPRSS2-ERG fusion genes, and those with the ETS family fusion genes. Finally, we integrated the methylation results with gene expression data from the same patient samples to explore the potential impact of different DNA methylation patterns on mRNA expression levels in PCa.

In order to validate the ETS family fusion positive and negative subtypes of PCa, we conducted survival analysis and clustering of DNA methylation profiles using a uniform analytical approach for both the recurrent fusion positive and negative groups, as well as the TMPRSS2-ERG fusion positive and negative groups. The

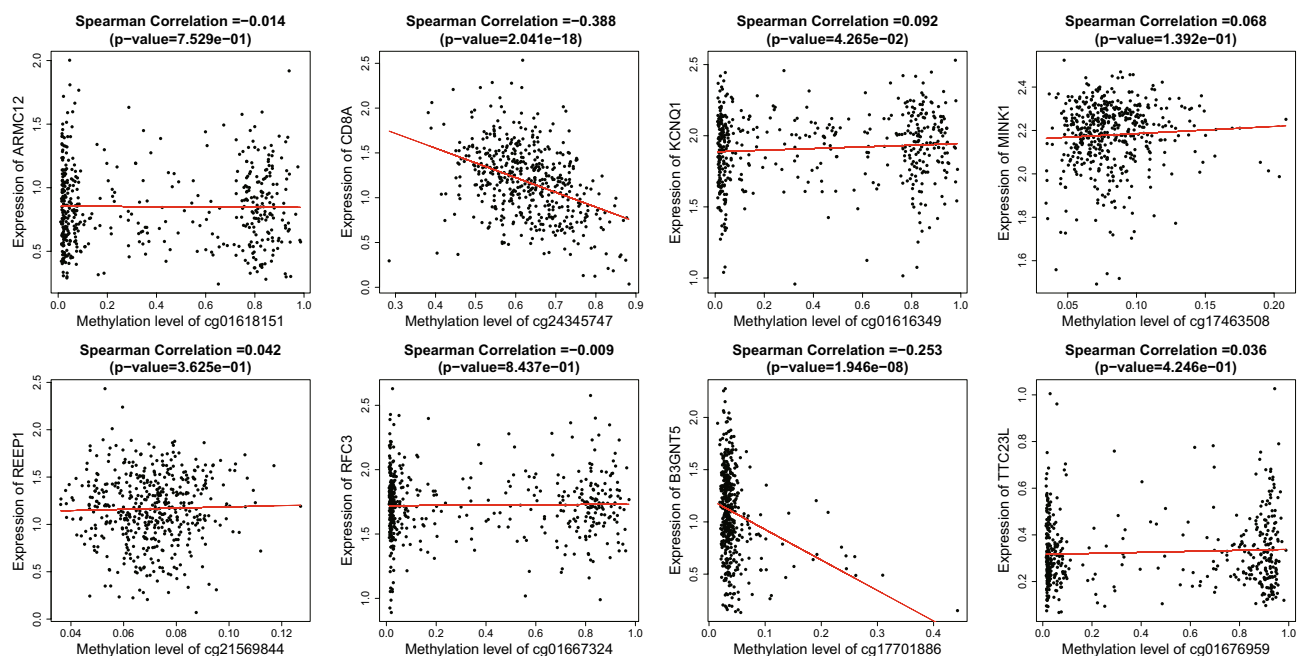


Fig. 6. The correlation between the methylation levels of the top 10 CpG sites with the most significantly different p-values in the ETS family fusion positive and fusion negative tumor groups and the expression levels of the corresponding 8 mRNAs. Among them, the low methylation sites cg06741367 and cg07059052 did not have corresponding mRNAs identified.

study results showed that ETS family fusion positive and negative PCa tumor could be distinctly differentiated by their DNA methylation profiles. In PCa, the subdivision of subtypes based on the DNA methylation landscape reveals significant differences between TMPRSS2-ERG gene fusion positive and TMPRSS2-ERG gene fusion negative tumors, elucidating distinct potential carcinogenic pathways between these molecular subtypes¹¹. Furthermore, for rhabdomyosarcoma, the study of methylation characteristics of the PAX3-FOXO1 and PAX7-FOXO1 fusion genes, which were formed by linking the DNA binding domain of PAX3 or PAX7 to the transactivation domain of FOXO1, provided a new method for distinguishing between fusion positive and fusion negative rhabdomyosarcoma¹⁷.

We found that fusion events of PCa fusion genes were mainly concentrated on chromosomes 1, 4, and 21, with fusion events within chromosomes 3 and 4 being the most common. The frequency of the ETS family fusion events on chromosome 21 was the highest, and the frequency of interchromosomal fusion was higher than intrachromosomal fusion. These fusion events on chromosomes were closely related to cancer and diseases. For example, the FIP1L1-PDGFR α fusion gene associated with Hypereosinophilic Syndrome (HES) was located on chromosome 4. Cell molecular cytogenetic analysis techniques have shown that due to interstitial deletion of chromosome 4, the FIP1L1 gene fused and constitutively activated PDGFR α gene, leading to the production of a protein with tyrosine kinase activity, thereby stimulating sustained proliferation of eosinophils¹⁸. Additionally, the fusion gene AML1-ETO was associated with Acute Myeloid Leukemia (AML), leading to decreased survival rates and increased recurrence rates¹⁹. The fusion gene AML1-ETO, resulting from translocation between chromosomes 8 and 21, was one of the most common chromosomal abnormalities in AML²⁰. For the ETS family of transcription factors, ETS Proto-Oncogene 2 (ETS2) was an important member of its transcription factor family, located on chromosome 21. The protein encoded by ETS2 was a Ca²⁺ dependent phosphorylated protein involved in regulating physiological and pathological processes such as cell proliferation, differentiation, and apoptosis²¹.

The distribution of fusion gene reading frame fusion modes in PCa was roughly equal between in-frame, out-of-frame, and 3'-UTR and 5'-UTR. The reading frame fusion ratio of the ETS family fusion genes mainly occurred out-of-frame. In Acute Myeloid Leukemia, t(1;21) could lead to out-of-frame fusion of RUNX1-CLCA2, and these out-of-frame fusions could generate hypothetical truncated RUNX1 isoforms²². Out-of-frame fusions retained the DNA-binding Runt domain but lacked the transcriptional regulatory domains of RUNX1. Truncated RUNX1 gene could promote the development of leukemia in patients²³. In Philadelphia chromosome (Ph) positive leukemia, the COOH terminal portion of the transcription product of the tumor specific antigen Bcr-ABL contained an out-of-frame coding amino acid sequence from the ABL gene. These variants were expressed in Ph positive Chronic Myeloid Leukemia (CML) and Acute Lymphoblastic Leukemia (ALL) patients²⁴. If a frameshift fusion occurs at the gene fusion point, it can convert a functional in-frame fusion into a dysfunctional out-of-frame fusion, thereby affecting the structure and function of the fusion protein²⁵. Future research should precisely identify fusion breakpoints and analyze the impact of frameshift fusions on the reading frame to understand disease mechanisms and guide therapeutic strategies.

We conducted a statistical analysis of the domains of parental genes and their corresponding fusion genes. The domains of fusion genes could be categorized into several situations: some fusion gene domains were an overlap of the domains of both parental genes, while another portion of fusion gene domains retained only the domains of one parental gene. There was also a subset that retained only a portion of the domains of one parental gene. Only a very small number of fusion genes introduced new domains on the basis of the original parental genes. These preserved fusion gene domains played crucial roles in transcription, cell signaling, and the immune system. For example, the DDT (the DNA-binding homeobox containing proteins and the different transcription) domain included proteins and various transcription and chromatin remodeling factors. It collaborated with other protein domains to regulate biological processes such as transcription, replication, and repair²⁶. The DBB domain referred to a domain shared by Dof (DNA binding with one finger), BANK1 (B-cell scaffold protein with ankyrin repeats 1), and BCAP (B-cell adapter for PI3K) proteins. This domain typically contained functional regions related to DNA binding or cell signaling²⁷. The NLRP3 protein inflammasome was a crucial component of the innate immune system, and its aberrant activation could lead to inflammatory diseases. The LRR (Leucine-Rich Repeat) domain controlled inflammasome activation by mediating NLRP3 protein self association, oligomerization, and interaction with the essential regulator NEK7²⁸. Additionally, new domains like IG_like had emerged. The polycystic kidney disease gene PKD1 encoded polycystin-1, which included 16 IG_like domains (or PKD domains), indicating its significant role in cell-cell or cell-matrix interactions²⁹.

Survival analysis results revealed that after 9 years follow up, the survival rate of the recurrent fusion negative group declined, whereas due to data limitations, the recurrent fusion positive group could not show subsequent survival rates. A significant statistical trend indicated that tumors in the recurrent fusion positive group had a higher risk of death compared to the recurrent fusion negative group. Although the statistical trend for the TMPRSS2-ERG fusion positive and negative groups was not very significant, compared to the TMPRSS2-ERG fusion positive group, the risk of disease progression or death in the TMPRSS2-ERG fusion negative group increased over time. Similarly, for the ETS family fusion positive and negative groups, after 9 years, the survival rate of the ETS family fusion positive group significantly decreased, while the fusion negative group could not provide subsequent survival rates due to data limitations. Although the statistical trend was not pronounced, the trend estimate suggested that compared to the ETS family fusion negative group, the risk of disease progression or death in the ETS family fusion positive group increased. This was also a limitation of the study; we only used PCa samples from TCGA, which represented a small dataset for finding differences. By analyzing the methylation levels of differentially expressed CpG sites between the ETS family fusion positive and negative groups, we conducted a cluster analysis of the samples and identified two main clusters. One cluster contained 88% of fusion positive tumors, and the other cluster contained 89% of fusion negative tumors. This suggested that the differentially methylated CpG sites were closely related to the ETS family fusion positive and negative status.

We identified hypermethylated CpG sites with significant differential expression between the ETS family fusion positive and fusion negative groups, including cg24345747 and cg17701886. We found a strong negative correlation between the methylation levels of these two CpG sites and the corresponding mRNA expression of CD8A and B3GNT5. These genes played crucial roles in the pathogenesis of cancer and diseases. CD8A (Cluster of Differentiation 8A) was a member of the T cell cytotoxic pathway related genes, encoding the CD8 antigen that collaborated with the T cell receptor on T cells to recognize and present antigens³⁰. Additionally, in childhood asthma samples, higher methylation of CpG sites in the CD8A promoter region significantly downregulated CD8A expression, affecting the TCR (T-cell receptor) signaling pathway, thereby regulating the progression of childhood atopic asthma³¹. Radiogenomic features indicated that predicting the expression of CD8A in bladder cancer patients preoperatively contributed to predicting patient prognosis and sensitivity to immunotherapy³². Copy number amplification and hypomethylation of the promoter region of B3GNT5 (β -1,3-N-acetylglucosaminyltransferase 5) gene contributed to its overexpression in the most invasive subtype of breast cancer, basal-like breast cancer (BLBC). It served as a prognostic marker and therapeutic target for BLBC³³. Dysregulation of sphingolipid metabolism was the major pathway in non-small cell lung cancer patients³⁴, and B3GNT5 gene, along with GAL3ST1 (β -1,4-galactosyltransferase 1) gene, altered the levels of metabolites such as lactate, sphingolipids, and sulfides in the serum of non-small cell lung cancer patients³⁵. This differential regulation affected the proliferation, migration, and invasion of tumor cells³⁵.

The correlation between the top ranked differentially expressed CpG sites with significant p-values for hypermethylation and the corresponding mRNA expression levels could be either positive or negative, depending on the location of aberrantly methylated CpG sites within the gene³⁶. For instance, the downregulation of the DNMT3B gene could result in upregulation through DNA remethylation, depending on its local chromatin structure³⁷.

Conclusions

In conclusion, there were substantial differences in the DNA methylation profiles between PCa tumors that are positive and negative for ETS family gene fusions. We analyzed the distribution of fusion genes in PCa and ETS family fusion positive genes on chromosomes, the fusion modes of reading frames, and the structural domain predictions of the fusion genes and their parental genes, in order to reveal the role of ETS family fusion genes in gene structure and functional regulation. Furthermore, hypermethylated CpG sites showed a strong negative correlation with the corresponding mRNA of CD8A and B3GNT5, indicating potential as prognostic markers for PCa. Therefore, by categorizing PCa into ETS family fusion positive and fusion negative subtypes based on epigenomic DNA methylation profiles, we have provided new perspectives on the treatment of PCa.

Data availability

The datasets generated and/or analysed during the current study can be obtained from the following sources: Nucleotide sequences and protein sequences of fusion genes are provided by the ChimerDB4.0 database (https://www.kobic.re.kr/chimerdb/downloads?name=ChimerDB4.0_sequence.txt.gz); breakpoint information data for fusion genes are also provided by the ChimerDB4.0 database (<https://www.kobic.re.kr/chimerdb/downloads?name=ChimerPub4.xlsx>); prostate cancer methylation data are provided by the UCSC Xena database (<https://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-PRAD.methylation450.tsv.gz>); gene mapping data correlating gene names with their corresponding methylation sites ID are provided by the UCSC Xena database (https://gdc-hub.s3.us-east-1.amazonaws.com/download/illuminaMethyl450_hg38_GDC); data for predicting gene structural domains are provided by the SMART database (http://smart.embl-heidelberg.de/smart/set_mode.cgi?NORMAL=1). The KEGG pathway enrichment data are provided by the DAVID database (<https://david.ncifcrf.gov/>).

Received: 10 April 2024; Accepted: 27 August 2024

Published online: 06 September 2024

References

- Suzuki, K. et al. Clinical features of treatment-related neuroendocrine prostate cancer: A case series. *Anticancer Res.* **40**, 3519–3526. <https://doi.org/10.21873/anticancer.14340> (2020).
- Song, C. J. & Chen, H. Overview of research on fusion genes in prostate cancer. *Transl. Cancer Res.* **9**, 1998. <https://doi.org/10.21037/tcr.2020.01.34> (2020).
- Alholle, A. et al. Genetic analyses of undifferentiated small round cell sarcoma identifies a novel sarcoma subtype with a recurrent CRTCL-SS18 gene fusion. *J. Pathol.* **245**, 186–196. <https://doi.org/10.1002/path.5071> (2018).
- Farooqi, A. A., Hou, M. F., Chen, C. C., Wang, C. L. & Chang, H. W. Androgen receptor and gene network: Micromechanics reassemble the signaling machinery of TMPRSS2-ERG positive prostate cancer cells. *Cancer Cell Int.* **14**, 12. <https://doi.org/10.1186/1475-2867-14-34> (2014).
- Peter, C. et al. Expression profiles frame the promoter specificity dilemma of the ETS family of transcription factors. *J. Biopharm. Stat.* **32**, 5693–5702. <https://doi.org/10.1093/nar/gkh906> (2004).
- Hermans, K. G. et al. TMPRSS2: ERG fusion by translocation or interstitial deletion is highly relevant in androgen-dependent prostate cancer, but is bypassed in late-stage androgen receptor-negative prostate cancer. *Cancer Res.* **66**, 10658–10663. <https://doi.org/10.1158/0008-5472.Can-06-1871> (2006).
- Shaikhbrahim, Z., Lindstrot, A., Langer, B., Buettner, R. & Wernert, N. Differential expression of ETS family members in prostate cancer tissues and androgen-sensitive and insensitive prostate cancer cell lines. *Int. J. Mol. Med.* **28**, 89–93. <https://doi.org/10.3892/ijmm.2011.672> (2011).
- Oikawa, T. J. C. S. ETS transcription factors: Possible targets for cancer therapy. *Cancer Sci.* **95**, 626–633. <https://doi.org/10.1111/j.1349-7006.2004.tb03320.x> (2010).
- Kunst, A. E., Leinsalu, M., Kasmel, A. & Habicht, J. Epigenomic profiling of prostate cancer identifies differentially methylated genes in: Fusion-positive versus fusion-negative tumors. *Clin. Epigenet.* **7**, 12. <https://doi.org/10.1186/s13148-015-0161-6> (2015).

10. Grady, W. M. & Ulrich, C. M. DNA alkylation and DNA methylation: Cooperating mechanisms driving the formation of colorectal adenomas and adenocarcinomas?. *Gut* **56**, 318–320. <https://doi.org/10.1136/gut.2006.106849> (2007).
11. Borno, S. T. et al. Genome-wide DNA methylation events in TMPRSS2-ERG fusion-negative prostate cancers implicate an EZH2-dependent mechanism with miR-26a hypermethylation. *Cancer Discov.* **2**, 1024–1035. <https://doi.org/10.1158/2159-8290.CD-12-0041> (2012).
12. Jang, Y. E. et al. ChimerDB 4.0: An updated and expanded database of fusion genes. *Nucleic Acids Res.* **48**, D817–D824. <https://doi.org/10.1093/nar/gkz1013> (2020).
13. Goldman, M. J. et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **38**, 675–678. <https://doi.org/10.1038/s41587-020-0546-8> (2020).
14. Niu, W., Zhang, T. & Ma, L. J. S. R. Correlation analysis between immune-related genes and cell infiltration revealed prostate cancer immunotherapy biomarkers linked to T cells gamma delta. *Sci. Rep.* **13**, 11. <https://doi.org/10.1038/s41598-023-28475-6> (2023).
15. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496. <https://doi.org/10.1093/nar/gkx922> (2018).
16. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57. <https://doi.org/10.1038/nprot.2008.211> (2009).
17. Meltzer, P. S. et al. Distinct methylation profiles characterize fusion-positive and fusion-negative rhabdomyosarcoma. *Modern Pathol.* **28**, 1214–1224. <https://doi.org/10.1038/modpathol.2015.82> (2015).
18. Kutok, J. et al. A tyrosine kinase created by fusion of the PDGFRA and FIP1L1 genes as a therapeutic target of imatinib in idiopathic hypereosinophilic syndrome. *N. Engl. J. Med.* **348**, 1201–1214. <https://doi.org/10.1056/NEJM0A025217> (2003).
19. Chou, F. S., Wunderlich, M., Griesinger, A. & Mulloy, J. C. N-RasG12D induces features of stepwise transformation in preleukemic human umbilical cord blood cultures expressing the AML1-ETO fusion gene. *Blood J. Am. Soc. Hematol.* **117**, 2237–2240. <https://doi.org/10.1182/blood-2010-01-264119> (2011).
20. Miyamoto, T., Weissman, I. L. & Akashi, K. AML1/ETO-expressing nonleukemic stem cells in acute myelogenous leukemia with 8;21 chromosomal translocation. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 7521–7526. <https://doi.org/10.1073/pnas.97.13.7521> (2000).
21. Wasyluk, B., Hagman, J. & Gutierrez-Hartmann, A. Ets transcription factors: Nuclear effectors of the Ras-MAP-kinase signaling pathway. *Trends Biochem. Sci.* **23**, 213–216. [https://doi.org/10.1016/s0968-0004\(98\)01211-0](https://doi.org/10.1016/s0968-0004(98)01211-0) (1998).
22. Ullah, H. et al. Generation of RUNX1-null reporter human embryonic stem cell line GIBHe008-A. *Stem Cell Res.* **45**, 5. <https://doi.org/10.1016/j.scr.2020.101800> (2020).
23. Giguère, A. & Hébert, J. CLCA2, anovel RUNX1 partner gene in a therapy-related leukemia with t(1;21)(p22;q22). *Cancer Genet. Cytogenet.* **202**, 94–100. <https://doi.org/10.1016/j.cancergencyto.2010.07.116> (2010).
24. Casnici, C. et al. Characterization of a monoclonal antibody specific for novel Bcr/Abl out-of-frame fusion proteins. *Hybridoma* **30**, 261–269. <https://doi.org/10.1089/hyb.2010.0122> (2011).
25. Nicolas, A. et al. Assessment of the structural and functional impact of in-frame mutations of the DMD gene, using the tools included in the eDystrophin online database. *Orphanet J. Rare Dis.* **7**, 45. <https://doi.org/10.1186/1750-1172-7-45> (2012).
26. Dong, J. Q. et al. SLIDE, the protein interacting domain of imitation switch remodelers, binds DDT-domain proteins of different subfamilies in chromatin remodeling complexes. *J. Integr. Plant Biol.* **55**, 928–937. <https://doi.org/10.1111/jipb.12069> (2013).
27. Lauenstein, J. U. et al. Negative regulation of TLR signaling by BCAP requires dimerization of its DBB domain. *J. Immunol.* **204**, 2269–2276. <https://doi.org/10.4049/jimmunol.1901210> (2020).
28. Duan, Y. H., Wang, J. H., Cai, J., Kelley, N. & He, Y. The leucine-rich repeat (LRR) domain of NLRP3 is required for NLRP3 inflammasome activation in macrophages. *J. Biol. Chem.* **298**, 11. <https://doi.org/10.1016/j.jbc.2022.102717> (2022).
29. Oxana, I. B. et al. Strong homophilic interactions of the Ig-like domains of polycystin-1, the protein product of an autosomal dominant polycystic kidney disease gene, PKD1. *Hum. Mol. Genet.* **9**, 1641–1649. <https://doi.org/10.1093/hmg/9.11.1641> (2000).
30. Li, H. H. et al. Recombinant jurkat cells (HMGN2-T cells) secrete cytokines and inhibit the growth of tumor cells. *J. Mol. Histol.* **53**, 741–751. <https://doi.org/10.1007/s10735-022-10084-8> (2022).
31. Liu, F. et al. Identification of methylated gene markers in childhood atopic asthma by integrating gene expression and methylation profiles based on bioinformatic analysis. *Pediatr. Allergy Immunol. Pulmonol.* **31**, 97–106. <https://doi.org/10.1089/ped.2017.0793> (2018).
32. Zheng, Z. T. et al. CD8A as a prognostic and immunotherapy predictive biomarker can be evaluated by MRI radiomics features in bladder cancer. *Cancers* **14**, 19. <https://doi.org/10.3390/cancers14194866> (2022).
33. Miao, Z. R. et al. Elevated transcription and glycosylation of B3GNT5 promotes breast cancer aggressiveness. *J. Exp. Clin. Cancer Res.* **41**, 15. <https://doi.org/10.1186/s13046-022-02375-5> (2022).
34. Alberg, A. J. et al. Plasma sphingolipids and lung cancer: A population-based, nested case-control study. *Cancer Epidemiol. Biomarkers Prev.* **22**, 1374–1382. <https://doi.org/10.1158/1055-9965.Epi-12-1424> (2013).
35. Meng, Q. et al. A circular network of coregulated sphingolipids dictates lung cancer growth and progression. *EBioMedicine* **66**, 18. <https://doi.org/10.1016/j.ebiom.2021.103301> (2021).
36. Lou, S. K. et al. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biol.* **15**, 21. <https://doi.org/10.1186/s13059-014-0408-0> (2014).
37. Yang, X. J. et al. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**, 577–590. <https://doi.org/10.1016/j.ccr.2014.07.028> (2014).

Acknowledgements

We thank the editor, reviewers, and colleagues for improving the manuscript.

Author contributions

WN performed the data analyses and drafted the manuscript. GL contributed to experiment analysis. ZT and LM revised the manuscript and administrated the project. The authors read and approved the final manuscript.

Funding

The National Natural Science Foundation of China (32060300, 32060145 and 31860308), the China Scholarship Council, and the Science Foundation of Shihezi University (RCZK201953) supported the research. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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/s41598-024-71285-7>.

Correspondence and requests for materials should be addressed to T.Z. or L.M.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024