


# Re-Evaluate Fusion Genes in Prostate Cancer

Ting Wei<sup>1\*</sup>, Ji Lu<sup>2\*</sup>, Tao Ma<sup>1</sup>, Haojie Huang<sup>3</sup>,  
Jean-Pierre Kocher<sup>1</sup> and Ligu Wang<sup>1,3,4</sup> 

<sup>1</sup>Division of Computational Biology, College of Medicine and Science, Mayo Clinic, Rochester, MN, USA. <sup>2</sup>Department of Urology, The First Hospital of Jilin University, Changchun, People's Republic of China. <sup>3</sup>Department of Biochemistry and Molecular Biology, College of Medicine and Science, Mayo Clinic, Rochester, MN, USA. <sup>4</sup>Bioinformatics and Computational Biology Graduate Program, University of Minnesota Rochester, Rochester, MN, USA.

Cancer Informatics  
Volume 20: 1–13  
© The Author(s) 2021  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/11769351211027592



## ABSTRACT

**BACKGROUND:** Thousands of gene fusions have been reported in prostate cancer, but their authenticity, incidence, and tumor specificity have not been thoroughly evaluated, nor have their genomic characteristics been carefully explored.

**METHODS:** We developed FusionVet to dedicatedly validate known fusion genes using RNA-seq alignments. Using FusionVet, we re-assessed 2727 gene fusions reported from 36 studies using the RNA-seq data generated by The Cancer Genome Atlas (TCGA). We also explored their genomic characteristics and interrogated the transcriptomic and DNA methylomic consequences of the E26 transformation-specific (ETS) fusions.

**RESULTS:** We found that nearly two-thirds of reported fusions are intra-chromosomal, and 80% of them were formed between 2 protein-coding genes. Although most (76%) genes were fused to only 1 partner, we observed many fusion hub genes that have multiple fusion partners, including ETS family genes, androgen receptor signaling pathway genes, tumor suppressor genes, and proto-oncogenes. More than 90% of the reported fusions cannot be validated by TCGA RNA-seq data. For those fusions that can be validated, 5% were detected from tumor and normal samples with similar frequencies, and only 4% (120 fusions) were tumor-specific. The occurrences of *ERG*, *ETV1*, and *ETV4* fusions were mutually exclusive, and their fusion statuses were tightly associated with overexpressions. Besides, we found *ERG* fusions were significantly co-occurred with *PTEN* deletion but mutually exclusive with common genomic alterations such as *SPOP* mutation and *FOXA1* mutation.

**CONCLUSIONS:** Most of the reported fusion genes cannot be validated by TCGA samples. The ETS family and androgen response genes were significantly enriched in prostate cancer-specific fusion genes. Transcription activity was significantly repressed, and the DNA methylation was significantly increased in samples carrying *ERG* fusion.

**KEYWORDS:** Prostate cancer, fusion gene, gene fusion, *TMPRSS2-ERG*, meta-analysis

**RECEIVED:** January 24, 2021. **ACCEPTED:** June 6, 2021.

**TYPE:** Original Research

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the Center for Individualized Medicine, Mayo Clinic; National Natural Science Foundation of China (31501052); and Natural Science Foundation of Jilin Province of China (20200201315JC).

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHORS:** Jean-Pierre Kocher, Division of Computational Biology, Mayo College of Medicine and Science, Mayo Clinic, Rochester, MN 55905, USA. Email: kocher.jeanpierre@mayo.edu

Ligu Wang, Division of Computational Biology, Mayo College of Medicine and Science, Mayo Clinic, Rochester, MN 55905, USA. Email: Wang.Ligu@mayo.edu

## Background

Fusion genes are made from the aberrant linkage of 2 physically separated genes. They generally originate from balanced chromosome rearrangements, including translocations, insertions, and inversions. Unbalanced chromosome rearrangements, such as the deletion of an interstitial chromosomal segment, could also lead to fusion genes.<sup>1</sup> The first discovered fusion gene *BCR-ABL1*, caused by the chromosome translocation between chromosomes 9 and 22 in chronic myeloid leukemia (CML) cells, was discovered in the early 1980s.<sup>2</sup> No specific initiating factor has been identified for most fusion genes<sup>2</sup> and only a few agents, including DNA topoisomerase II poisons and radiation, have been shown to increase the risk of fusion genes. Many identified fusion genes are cancer-specific, such as the *PAX3-FOXO1* fusion in Rhabdomyosarcoma,<sup>3,4</sup> the

*EML4-ALK* fusion in lung cancer,<sup>5,6</sup> the *MLL-AF4* fusion in acute lymphoblastic leukemia,<sup>7</sup> and the *TMPRSS2-ERG* fusion in prostate cancer.<sup>7</sup> As fusion genes could produce chimeric proteins with new or altered activities, they have been recognized as playing important roles in tumorigenesis. Until now, tens of thousands of cancer-associated fusion genes have been reported, including 20,731 fusion genes identified from ~10,000 The Cancer Genome Atlas (TCGA) tumor samples across 33 cancer types.<sup>8</sup>

Prostate cancer is the second most commonly occurring cancer in men and the second leading cause of cancer death in the United States.<sup>9</sup> The androgen and androgen receptor (AR) signaling axis plays a central role in prostate cancer development and progression. Androgens such as testosterone are synthesized primarily by the Leydig cells in the testes. After entering prostate cells, testosterone is converted into a more potent 5 $\alpha$ -dihydrotestosterone (DHT). The DHT binds to

\*Equal contributions.



the AR with high affinity and promotes the translocation of AR into the nucleus. In the nucleus, AR dimers bind to androgen response elements (AREs) to regulate the target gene transcription. While some AREs are located in genes' promoter regions (such as *KLK3* and *TMPRSS2*), many AREs are located in distal enhancer regions, and the enhancer-bound AR regulates its target genes (such as *UBE2C*) through chromatin loopings.<sup>10</sup> Understanding the AR-regulated downstream transcriptional programs might open new avenues for prostate cancer treatment.

Recent whole-genome sequencing studies suggest that the genomic landscape of prostate cancer differs significantly from that of other solid tumors, such as breast and colon cancer. One of the key characteristics of the prostate cancer genome is fusion genes.<sup>11</sup> The *TMPRSS2-ERG* fusion, first discovered in 2005,<sup>12</sup> is by far the most common one and presents in 40% to 80% of prostate cancer genomes.<sup>13</sup> Both *TMPRSS2* and *ERG* genes locate on chromosome 21 with a genomic distance of 2.8 Mb, and the fusion is made by joining the promoter and 5' exons of the *TMPRSS2* gene with the coding sequences of the *ERG* oncogene. The *TMPRSS2-ERG* fusion gene could be driven by 2 distinct mechanisms: the genomic region between the 2 genes was interstitially deleted, which is the case in approximately 60% of the fusion-positive tumors, or it can be the result of more complex genomic translocations.<sup>14,15</sup> As *TMPRSS2* is a prostate-specific and androgen-regulated gene, the *ERG* expression is inappropriately upregulated by AR (mediated through *TMPRSS2*) in almost all the fusion-positive tumors. Due to the high prevalence, *TMPRSS2-ERG* fusion becomes one of the important biomarkers for prostate cancer screening, diagnosis, and target therapy.<sup>16</sup> The *TMPRSS2-ERG* fusion has been associated with more aggressive prostate cancers, the higher pathologic stages, the higher rate of recurrence, and therefore poorer prognosis in independent cohorts of surgically treated localized prostate cancer cases.<sup>17-19</sup> In addition to *ERG* fusion, other members of the E26 transformation-specific (ETS) family, including *ETV1*, *ETV4*, *ETV5*, and *FLI1*, were also frequently detected in prostate cancers with fusion partners, including *TMPRSS2*, *SLC45A3*, and *KLK2*.<sup>12,20-23</sup>

With the advent of genome and transcriptome sequencing, thousands of candidate fusion genes have been reported in prostate cancer by various groups. However, these candidate fusions are detected from different cohorts of patients and have not been thoroughly evaluated independently. In this study, we interrogated the genomic characteristics of 2727 fusion genes reported by 36 previous studies and used the prostate adenocarcinoma (PRAD) cohort from the TCGA to re-assess their authenticity, incidence rate, and tumor specificity.

## Methods

### Collection of fusion genes

Candidate fusion genes were collected from 36 studies listed in Supplementary Table 1. Redundant fusion genes were removed.

Their genomic coordinates were converted into hg19/GRCh37. All gene identifiers were converted into HUGO (Human Genome Organization) gene symbols when possible. A fusion is called ETS fusion when one or both the fusion partners are ETS family members, including *E1AF*, *EHF*, *ELF1*, *ELF2*, *ELF3*, *ELF4*, *ELF5*, *ELK1*, *ELK3*, *ELK4*, *ER71*, *ER81*, *ERF*, *ERG*, *ERM*, *ESE1*, *ESE2*, *ESE3*, *ESX*, *ETS1*, *ETS2*, *ETV1*, *ETV2*, *ETV3*, *ETV4*, *ETV5*, *ETV6*, *ETV7*, *FEV*, *FLI1*, *GABPA*, *MEF*, *NERF*, *NET*, *PDEF*, *PE1*, *PE2*, *PEA3*, *PSE*, *SAP1*, *SAP2*, *SPDEF*, *SPI1*, *SPIB*, *SPIC*, *TEL*, and *TEL2*. Similarly, a fusion is called AR target fusion when one or both fusion partners are AR target genes as defined in the androgen-responsive gene database.<sup>24</sup> The fusion will be tagged as "immediate neighbor" if the 2 fusion partners were adjacent to each other on the chromosome. GENCODE annotation (v22) was used to assign each gene to "protein-coding," "lincRNA," "antisense," "transcribed\_unprocessed\_pseudogene," and "unprocessed\_pseudogene" categories.

### Validate candidate fusion genes using TCGA prostate cancer samples

To evaluate the authenticity and tumor specificity of collected fusion genes, we used the TCGA PRAD cohort, one of the largest primary prostate cancer cohorts with both RNA-seq and clinical data available, for the cross-examination.

The RNA-seq data of 496 tumor samples and 52 normal samples from the TCGA PRAD cohort was downloaded from dbGAP<sup>25</sup> with the accession number *phs000178.v11.p8*. Six normal samples (TCGA-CH-5769-11A, TCGA-EJ-7115-11A, TCGA-G9-6351-11A, TCGA-G9-6362-11A, TCGA-G9-6363-11A, TCGA-G9-6499-11A) were removed due to contamination of tumor content.<sup>26</sup> All the samples (496 tumors and 46 normals) used to validate candidate fusions are listed in Supplementary Table 5. We used FusionVet (Fusion Visualization and Evaluation Tool) to search for chimeric "split reads" (ie, a single read was split and mapped to 2 different genes) or "read pairs" (ie, 2 reads of a read pair were mapped to 2 different genes) that support the fusion. Fusion was determined as validated if 2 or more split reads or read pairs were detected.

### FusionVet

We developed FusionVet to quickly and accurately examine whether a given fusion gene existed in a particular sample. FusionVet requires an alignment file (in BAM [Binary Alignment Map] format) and a file describing the genomic coordinates of chimeras. It generates a BAM format file containing the supporting chimeric reads, a BED (Browser Extensible Data) format file for intra-chromosomal fusion visualization, an Interact format file for both intra- and inter-chromosomal fusion visualization and a summary file. FusionVet is implemented in Python programming language, and the source code of FusionVet has been deposited into both

GitHub (<https://github.com/liguowang/FusionVet>) and the Python package index (<https://pypi.org/project/FusionVet/>) to facilitate downloading and installation, online documentation is available from <https://fusionvet.readthedocs.io/en/latest/>.

To evaluate the performance of FusionVet, we ran this algorithm to detect *TMPRSS2-ERG* fusions from the 333 prostate cancer samples published in the original TCGA paper<sup>27</sup> and compared the results with those of 4 other algorithms, including FusionSeq,<sup>28</sup> MapSplice,<sup>29</sup> PRADA,<sup>30</sup> and DEEPEST.<sup>31</sup> *TMPRSS2-ERG* fusion called by FusionSeq and MaSplice were downloaded from the TCGA publication,<sup>27</sup> *TMPRSS2-ERG* fusion called by PRADA were downloaded from the tumor fusion gene data portal,<sup>8</sup> and *TMPRSS2-ERG* fusion called by DEEPEST were downloaded from the original publication.<sup>31</sup> As the ground truth was unknown, we cannot directly compare the performance of these tools. Instead, we compared the *ERG* expression of samples that were specifically detected by these tools; the rationale is that *TMPRSS2-ERG* fusion usually leads to *ERG* overexpression in prostate cancer.<sup>12,32</sup>

### RNA-seq data analysis

We downloaded the HTseq gene count, FPKM (fragments per kilobase of transcript per million mapped reads), FPKM-UQ files (version 07-22-2019) of all the TCGA prostate cancer samples from the Genomic Data Commons (GDC) Data Portal.<sup>33</sup> The original  $\log_2(x + 1)$  transformed values were converted back into the original count values to facilitate differential gene expression analysis. The FPKM values were used to perform the correlation analysis with beta values from DNA methylation. To remove the potential confounding effects of other common genomic alterations, we chose a total of 133 TCGA samples for the differential gene expression analyses. In particular, we selected 52 samples harboring only *ERG* fusion but no other fusions and common genomic alterations (here, common alterations referred to mutations or copy number alterations occurred in *AR*, *FOXA1*, *SPOB*, *TP53*, *RB1*, *PTEN*, *LRP1B*, *KMT2C*, *KMT2D*). Similarly, we selected 5 samples that are *ETV1* fusion only and another 5 samples that are *ETV4* fusion only. We also selected 71 samples with no ETS fusions and other common genomic alterations as the control group. Differentially expressed genes (DEGs) were identified using edgeR with the Trimmed Mean of M-values normalization method to estimate scaling factors.<sup>34</sup> false discovery rate (FDR)  $\leq 0.05$  and  $|\log_2(\text{FC})| \geq 1$  were used to call differential expressed genes in *ETV1* and *ETV4* fusion samples, and FDR  $\leq 0.01$  and  $|\log_2(\text{FC})| \geq 1$  were used to call differential expressed genes in *ERG* fusion samples.

### The DNA methylation analysis

We downloaded the DNA methylation beta values (Illumina HumanMethylation450 BeadChip array, level 3, version 2012-02-24) from the GDC website.<sup>33</sup> As described above, 6 normal

samples with contaminated tumor content were removed. An offset value of 0.5 was added to all beta values to convert them between 0 (ie, unmethylated) and 1 (ie, methylated). Principal components analysis (PCA) and differential analysis were performed by CpGTools developed by our team.<sup>35</sup> In particular, the top 10 000 most variable CpG sites (ranked by the standard deviation) were used to perform the PCA analysis. The “glm” model was used to detect differentially methylated CpGs between *ERG*-fusion positive samples and control samples as described above. The Benjamini-Hochberg procedure was used for the multiple test correction, and adjusted  $P \leq 0.01$  was used as the cutoff. For gene enrichment analysis, we picked the top 1000 most significant CpGs, retrieved the associated gene symbols from the MethylationEpic Manifest file, and then performed the enrichment analysis using the “curated gene sets” in mSigDB.<sup>36</sup>

### Co-occurrence and mutual exclusivity analyses

Suppose we have 2 genes A and B and  $n$  patients, then the  $n$  patients can be divided into 4 groups according to the alteration status of genes A and B:  $N_A$  (patients only having gene A altered but not gene B),  $N_B$  (patients only having gene B altered but not gene A),  $N_{AB}$  (patients having both A and B altered simultaneously),  $N_{none}$  (patients having neither gene A nor gene B altered). The log odds ratio (OR) is calculated as 
$$\text{OR} = \log_2((N_{none} \times N_{AB}) / (N_A \times N_B))$$

where  $\text{OR} > 0$  indicates co-occurrence and  $\text{OR} < 0$  indicates mutual exclusivity. The  $P$  value is calculated as 1-sided Fisher's exact test.

## Results

### Overview of fusion genes identified from prostate cancer

Among the 2727 fusions (3863 unique genes) we collected from 36 previous prostate cancer studies, 80% were formed between 2 protein-coding genes, and 18% were formed between 1 protein-coding gene and 1 lincRNA, pseudogene, or anti-sense RNA (Supplementary Tables 1 and 2). These fusions comprise 1758 (65%) intra-chromosomal and 969 (36%) inter-chromosomal fusions. Intra-chromosomal fusions could be further divided into fusions whose partner genes were immediate neighbors on the genome (9%), and fusions whose partner genes were separated by other genes (55%). In contrast to inter-chromosomal fusions, intra-chromosomal fusions that occurred between partner genes locating on the same DNA strand were significantly more frequent than that occurred between genes locating on different strands ( $P = 2.56 \times 10^{-6}$ , Fisher's exact test) (Figure 1A). This finding suggested that these fusion genes may derive from the transcriptional read-through or impaired post-transcriptional processing (eg, trans-splicing). As genes that can fuse to multiple partners are likely to be oncogenic drivers, we ranked the 3863 unique genes by the number of fusion partners (Supplementary Table 3). While

most (76%) genes fuse to only 1 partner, we found 34 genes fused to at least 6 different partners (Figure 1B). The top candidates include ETS gene family members: *ERG* (26 partners), *ETV1* (22 partners), and *ETV4* (14 partners); AR signaling pathway genes: *TMPRSS2* (23 partners), *SLC45A3* (9 partners), *AR* (7 partners), and *FOXA1* (7 partners); tumor suppressors: *TP53* (14 partners) and *PTEN* (11 partners); and proto-oncogenes: *BRAF* (9 partners) and *PIK3C2A* (8 partners) (Figures 1B and 2A). Despite most analyzed fusions were intra-chromosomal (Figure 1A), we found that most fusion partners of *ERG* (77%), *TMPRSS2* (65%), and *ETV1* (77%) are located on different chromosomes (Figures 1B and 2A). In contrast, most fusion partners of *ETV4* are located on chromosome 17, which has the highest density of fusion genes (Supplementary Figure 1).

The AR-mediated transcription plays critical role in prostate cancer development and progression. Most AR binding sites are located in distal enhancer regions far from promoters.<sup>37-39</sup> When AR exerts its transcriptional regulation through long-range chromatin looping, distant AR-regulated genes can be brought into proximity in 3-dimensional space, and therefore promotes gene fusions when double-strand DNA breaks occur.<sup>10</sup> Therefore, we sought to investigate whether AR-regulated genes are predisposed to fusions. Out of 2727 collected fusion genes, 701 fusions (26%) with one or both fusion partners being AR-regulated genes (Supplementary Figure 2). We first divided all the collected fusion genes into 3 classes: the “1-to-1” class that includes fusions in which 1 gene has only 1 fusion partner, the “1-to-2” class that includes fusions that 1 gene has 2 fusion partners, and the “1-to-multiple” class that contains fusions that 1 gene has been detected to fuse to 3 or more partners. Then, we further divided each class into 3 groups: fusions that involve AR target genes (termed as “AR fusions”), fusions that involve ETS family genes (termed as “ETS fusions”), and fusions that involve other genes (termed as “other fusions”). By comparison, we found that while “other fusions” were evenly distributed in all 3 classes, the vast majority (91%) of “ETS fusions” were found in the “1-to-multiple” class, suggesting ETS fusions are fusion hubs and probably the oncogenic drivers in prostate cancer. Likewise, “AR fusions” were also significantly over-represented in the “1-to-multiple” class, compared with the “1-to-1” class ( $P = 2.2 \times 10^{-16}$ ) and the “1-to-2” class ( $P = 2.57 \times 10^{-16}$ ) (Figure 2B and Supplementary Figure 3). The enrichment of “AR fusions” in the “1-to-multiple” class suggests that AR-mediated chromatin looping may be another mechanism driving gene fusions in prostate cancer.

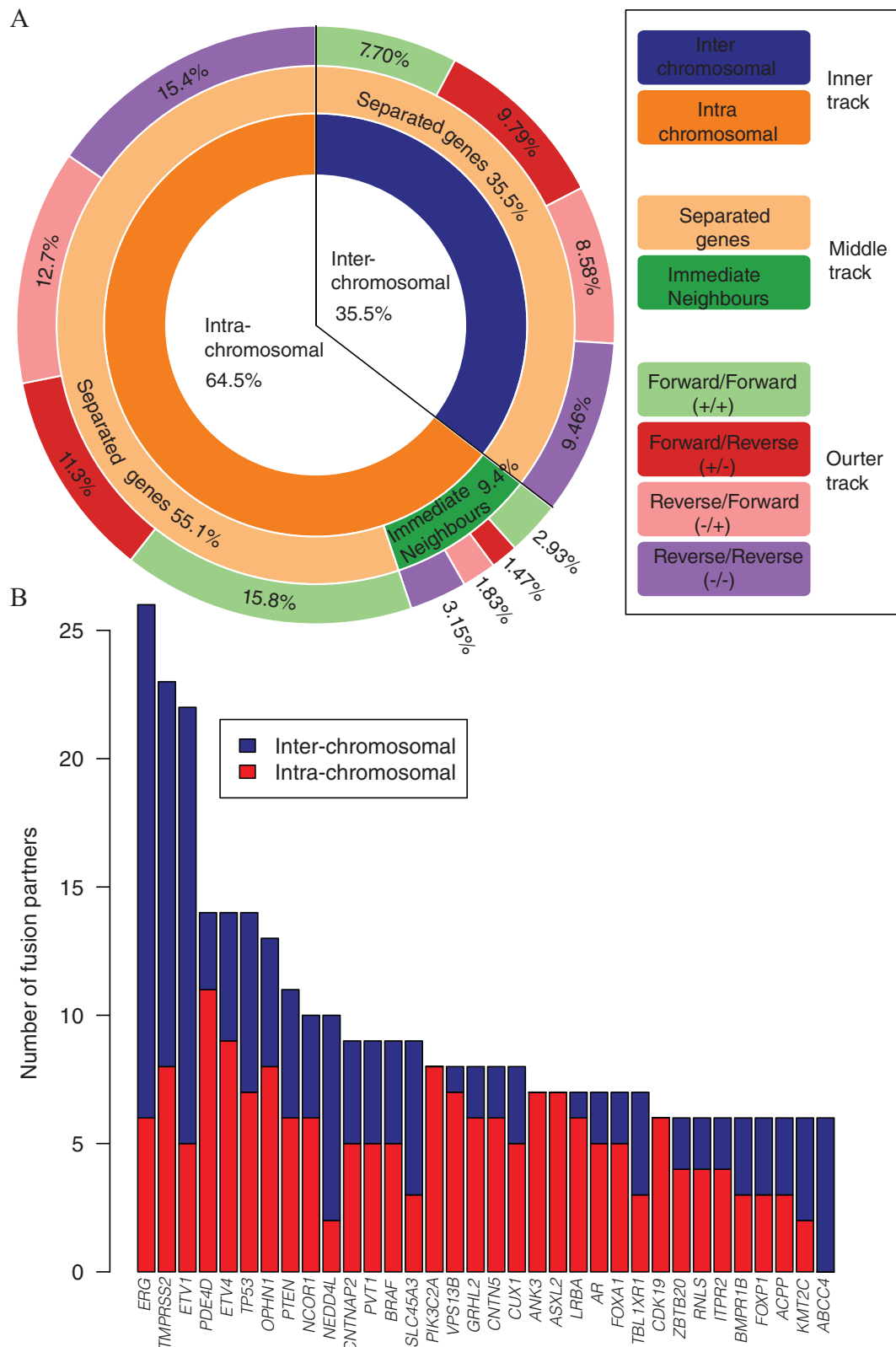
#### Validate fusion genes using TCGA prostate cancer cohort

Thousands of fusion events have been reported in prostate cancer. However, they were identified in various tumor stages by

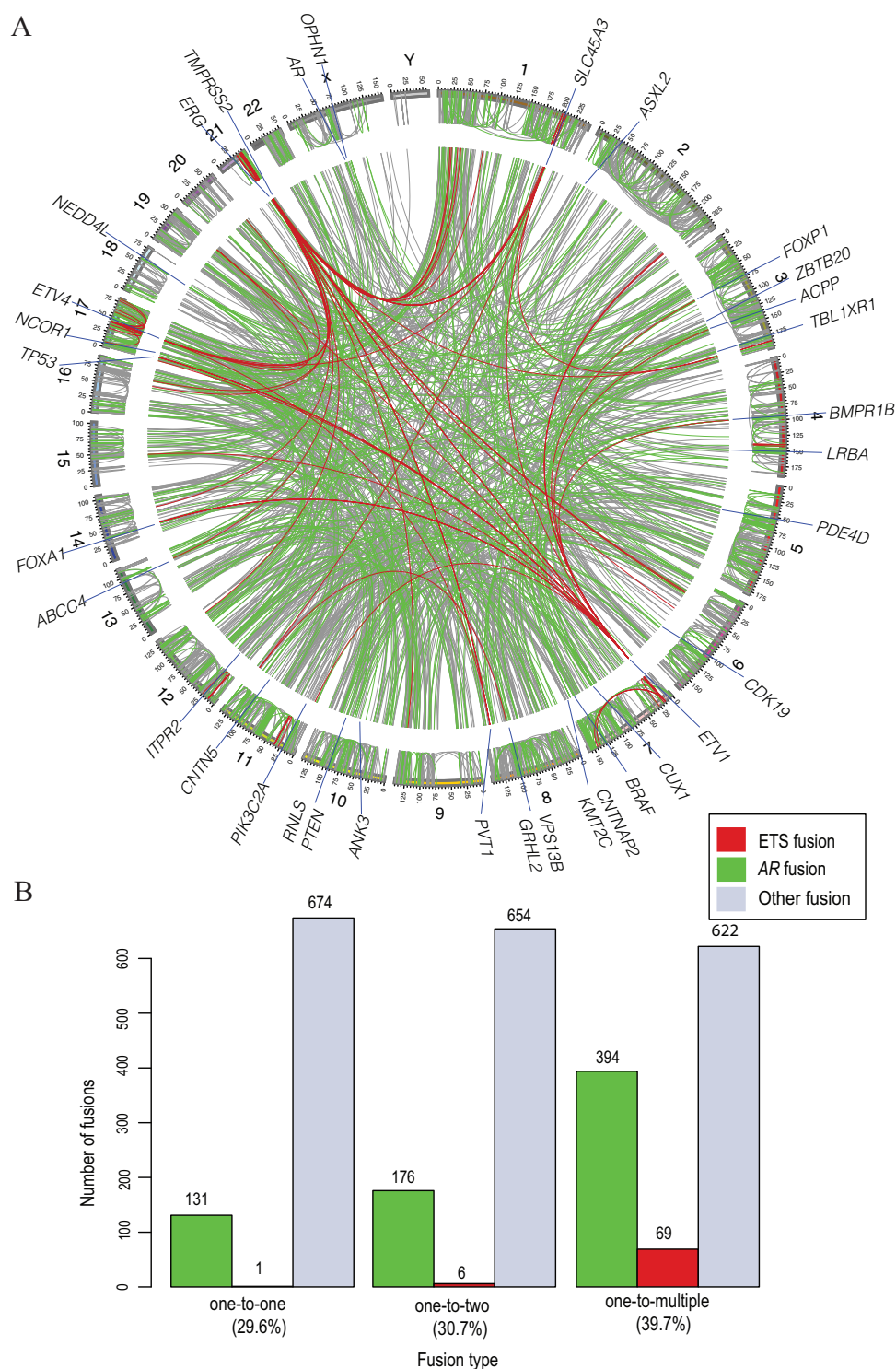
different groups using different protocols and analysis pipelines. Thus, except for a few recurrent fusions such as *TMPRSS2-ERG* and *TMPRSS2-ETV1*, most of these reported fusion genes may be study-specific or tumor stage-specific. Evaluating the authenticity and tumor specificity of these fusion genes using the same dataset and algorithm has not been done. We chose the TCGA PRAD cohort because it is one of the largest cohorts with RNA-seq and other genomic and clinical data available. As a primary prostate cancer cohort, it is also a valuable resource to cross-examine if previously identified fusion genes occurred early in prostate cancer.

To verify fusion genes using RNA-seq data, we developed FusionVet—a bioinformatic tool designed to validate given fusions using RNA-seq data (see “Methods” section). To demonstrate its performance, we applied FusionVet, along with other state-of-the-art tools, including FusionSeq,<sup>28</sup> MapSplice,<sup>29</sup> PRADA,<sup>30</sup> and DEEPEST,<sup>31</sup> to 333 prostate cancer samples published by TCGA<sup>27</sup> to examine the existence of *TMPRSS2-ERG* fusion. In summary, FusionVet, FusionSeq, MapSplice, PRADA, and DEEPEST detected *TMPRSS2-ERG* fusion in 128, 138, 121, 132, and 136 samples, respectively. There was a shared list of 112 samples among the results of all 5 algorithms, showing a high concordance rate (Supplementary Figure 4 and Supplementary Table 4). Besides, we found that the *TMPRSS2-ERG* fusion-positive samples detected exclusively by our FusionVet have significantly increased *ERG* expression, which is consistent with the fact that *TMPRSS2-ERG* fusion usually leads to *ERG* overexpression in most prostate cancers,<sup>12,32</sup> thus suggesting FusionVet has a superior performance in validating genuine fusions (Supplementary Figure 5).

Then, we used FusionVet to assess the collected 2727 gene fusions with TCGA PRAD RNA-seq dataset consisting of 496 tumor samples and 46 matched normal samples. Overall, 2462 (90%) collected gene fusions cannot be confirmed by any of the TCGA samples, suggesting that most reported fusions could be patient-specific, tumor stage-specific (eg, Castrate-resistant prostate cancer or CRPC), cohort-specific sporadic events, or false positives (Figure 3A and Supplementary Table 5). For the remaining 265 (10%) fusions that can be validated by TCGA samples, 145 fusions exhibited similar validation rates in both tumor and normal samples (Pearson’s  $r = 0.94$ ,  $P < 2.2E-16$ ), showing that they were not cancer-specific (Figure 3B). Among 120 fusions detected only in tumor samples, 27 are involved with ETS family genes, including *ERG*, *ETV1*, *ETV4*, and *ETV5* (Figure 3B and C and Supplementary Table 6). Statistical analysis by Fisher’s exact test showed that only 4 fusions were significantly enriched in tumor samples, including *TMPRSS2-ERG* ( $P = 1.04 \times 10^{-9}$ ), *AMACR-SLC45A2* ( $P = 3.4 \times 10^{-8}$ ), *MLLT11-GABPB2* ( $P = 5.6 \times 10^{-4}$ ), and *ARHGEF38-INTS12* ( $P = 0.049$ ) (Figure 3C). Both fusion partners of *AMACR-SLC45A2*, *MLLT11-GABPB2*, and *ARHGEF38-INTS12* are immediate neighbors on the same chromosome,



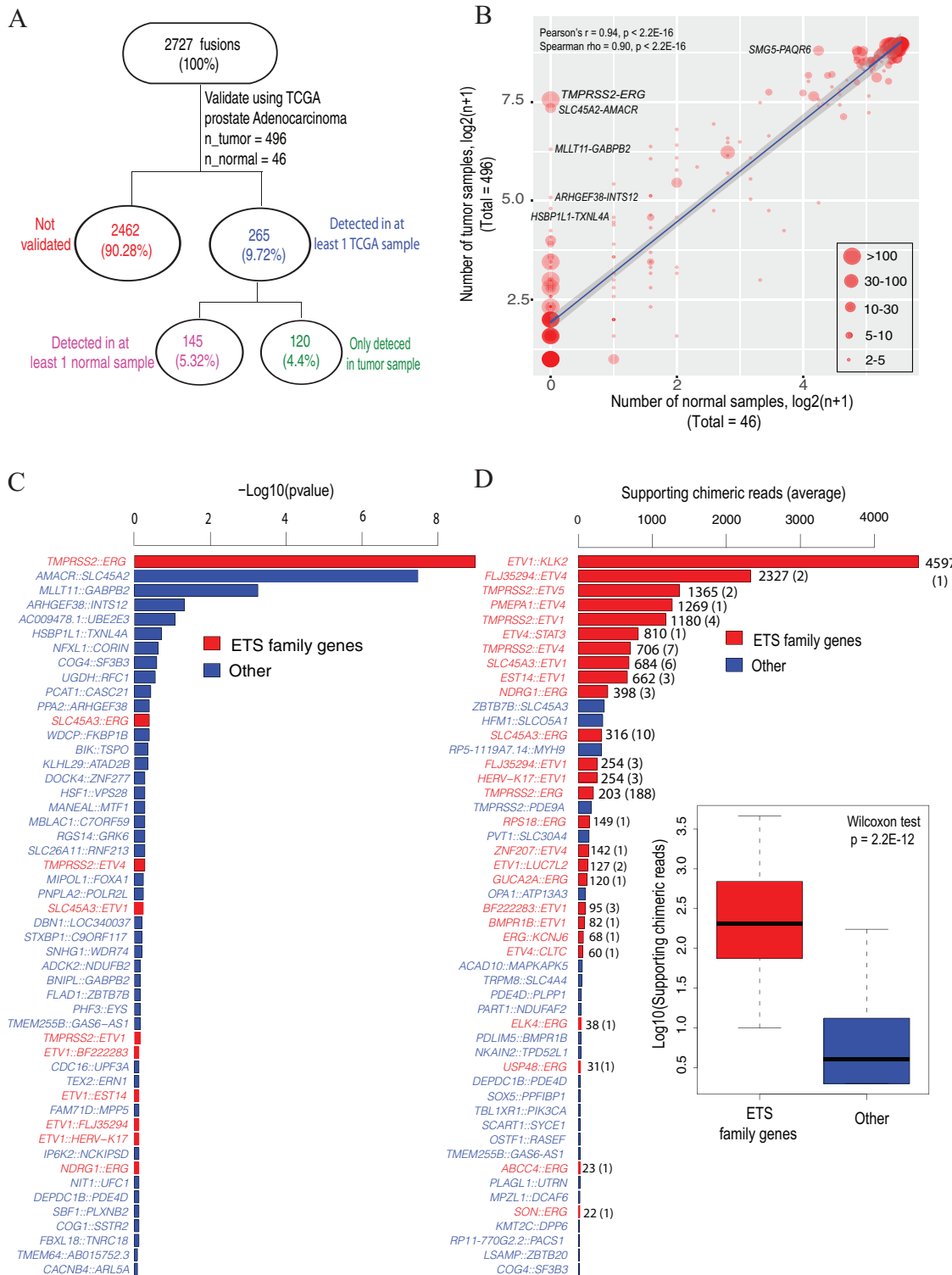
**Figure 1.** Genomic characteristics of collected fusion genes: (A) A 3-layered sunburst diagram showing the genomic features of the collected fusion genes in prostate cancer. The inner layer indicated the breakout of intra- (orange) and inter-chromosomal (blue) fusions. In the middle layer, intra-chromosomal fusions were further divided into fusions whose fusion partner genes were immediate neighbors on the chromosome (green), and those fusions whose fusion partner genes were separated by other genes (yellow). The outer layer showed the fractions of fusions by deoxyribonucleic acid strands: forward/forward (lime), forward/reverse (red), reverse/forward (pink), and reverse/reverse (purple). (B) Bar plot showing the frequency of fusion partners of the top 34 potential oncogenic driver genes that have at least 6 different fusion partners. Inter-chromosomal and intra-chromosomal genes are indicated in blue and red, respectively.



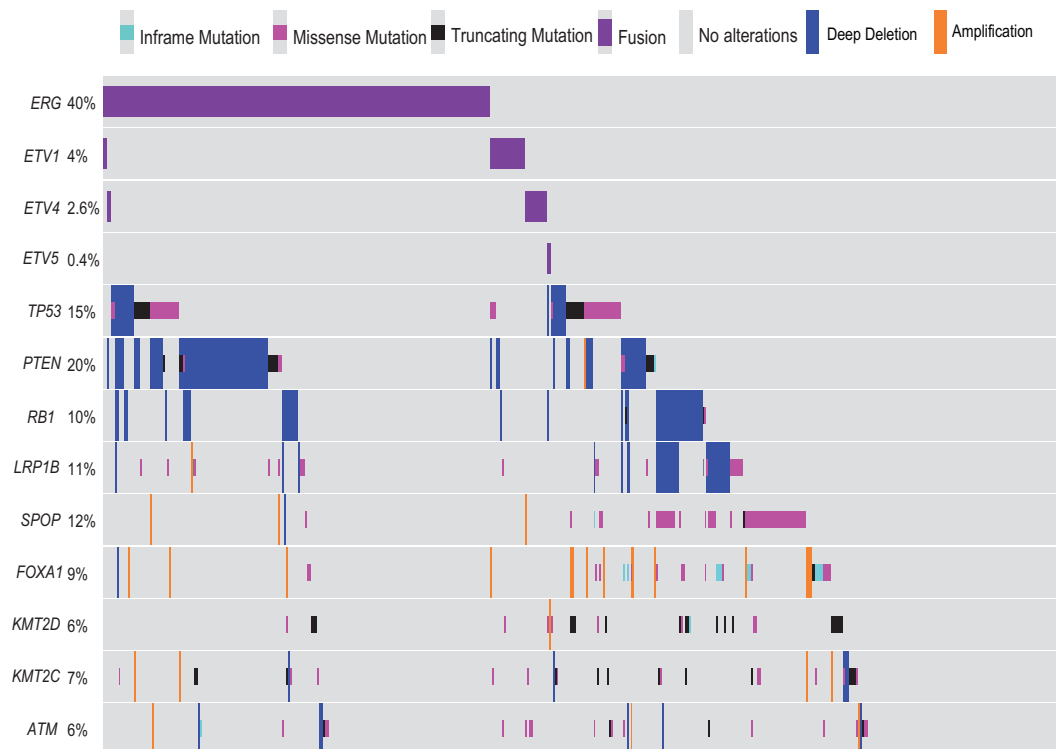
**Figure 2.** Genomic position and frequency of “ETS fusions” (ie, fusions involving ETS family genes), “AR fusions” (ie, fusions involving AR-regulated genes), and “Other fusions” (ie, all the other fusions except ETS and AR fusions). (A) Circos plot showing the positions “ETS fusions” (red arches), “AR fusions” (green arches), and “Other fusions” (gray arches). (B) All fusions were divided into 3 classes including “1-to-1” (1 gene only fused to another gene), “1-to-2” (1 gene has been detected to fuse to 2 different genes), and “1-to-multiple” (1 gene has been detected to fuse to 3 or more genes). The bar plot comparing the fractions of “ETS fusion” (red bars), “AR fusion” (green bars), and “Other fusion” (gray bars) among the 3 classes. AR indicates androgen receptor; ETS, E26 transformation-specific.

suggesting they were transcription-derived fusions (see “Discussion” section). Most of the other ETS fusions, such as *SLC45A3-ERG* and *TMPRSS2-ETV4*, that did not reach statistical significance, were likely due to smaller sample sizes

(Figure 3C). It is worth noting that the genes involved in ETS gene fusion exhibited much higher expressions than those of other gene fusions, highlighting their functional significance (Figure 3D and Supplementary Table 6). Interestingly, we



**Figure 3.** Evaluation of the collected fusion genes using RNA-seq data of 496 tumor samples and 46 matched normal samples from the TCGA prostate cancer cohort. (A) The analysis procedure of validating collected fusions using TCGA RNA-seq data. (B) Correlation of the detection frequencies of 265 fusions in TCGA normal, x-axis, measured by  $\log_2(n+1)$  and tumor, y-axis, measured by  $\log_2(n+1)$  samples, where  $n$  is the number of TCGA samples carrying the fusion. The size of circles indicates the average number of supporting chimeric reads (ie, expression) of the fusions. The solid blue line was the linear regression line fitted to 265 dots and the gray band around the line represented the standard error of the regression line. (C) Fisher's exact test  $P$  values of top 50 fusions. The ETS fusions were indicated in red bars, other gene fusions were indicated in blue bars. (D) Bar plot showing the average number of supporting chimeric reads (of all TCGA samples) for the top 50 fusions. Numbers in parentheses indicate the number of TCGA samples carrying the fusion. The ETS fusions were indicated in red bars, other gene fusions were indicated in blue bars. Boxplot showing the difference in abundance between ETS fusions and other fusions. ETS indicates E26 transformation-specific; TCGA, The Cancer Genome Atlas.



**Figure 4.** Oncoprint plot showing the mutual exclusivity and co-occurrence relationships among E26 transformation-specific fusions (including *ERG*, *ETV1*, *ETV4*, and *ETV5*) and other common genomic alterations (including *ATM*, *FOXA1*, *KMT2C*, *KMT2D*, *LRP1B*, *PTEN*, *RB1*, *SPOP*, and *TP53*) detected in The Cancer Genome Atlas prostate cancer cohort.

found *ERG* was almost exclusively fused to *TMPRSS2* in prostate cancers, with 90% of *ERG* fusion-positive samples harboring *TMPRSS2-ERG*. *ETV4* was less selective than *ERG*, with 54% of *ETV4* fusions being *TMPRSS2-ETV4*. In contrast, *ETV1* fused to its partner genes at comparable frequency levels. For example, the frequencies of *SLC45A3-ETV1*, *TMPRSS2-ETV1*, and *EST14-ETV1* were 23%, 15%, and 12%, respectively (Supplementary Figure 6).

#### *Co-occurrence and mutual exclusivity between ETS fusions and other common genomic alterations*

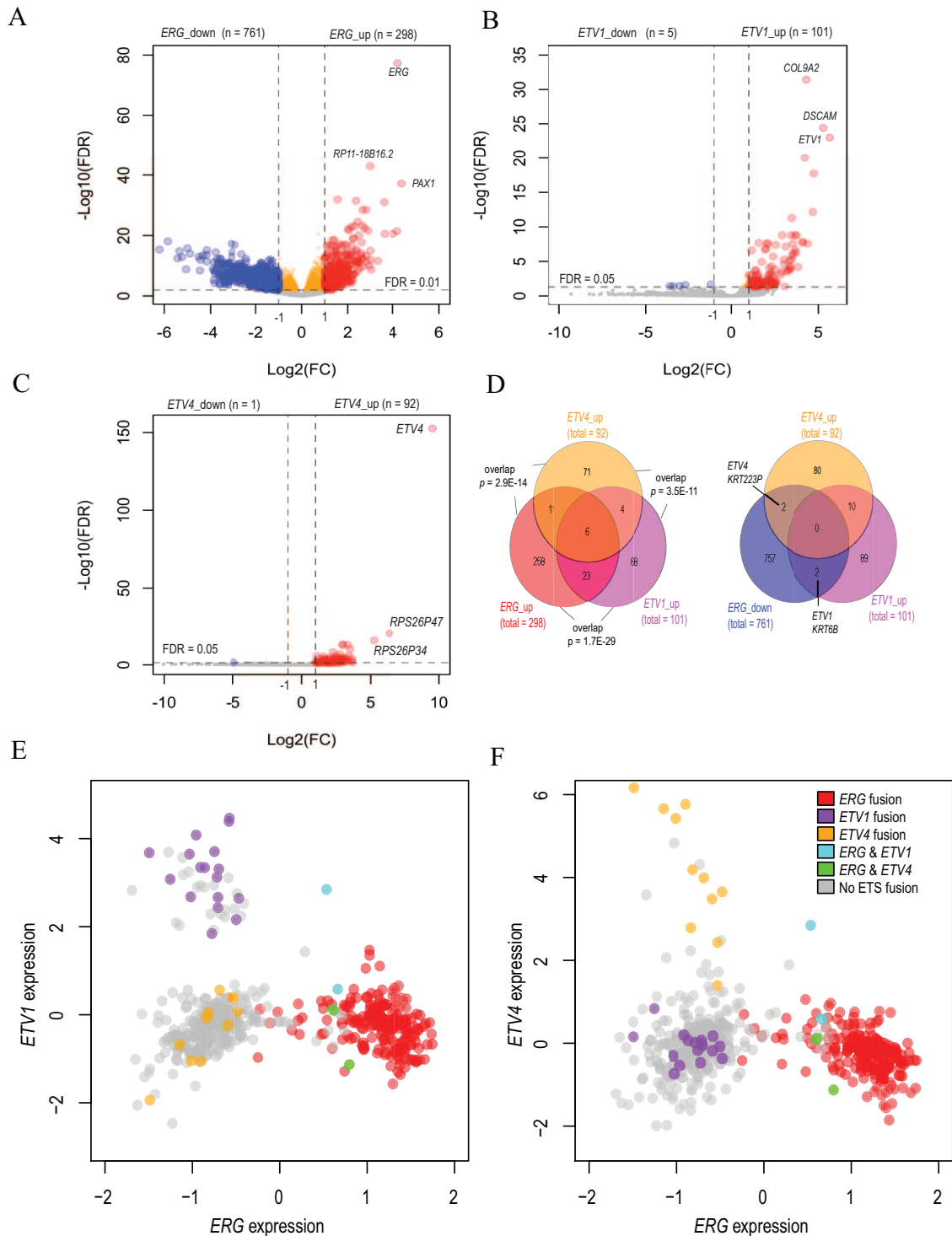
Consistent with previous studies,<sup>40</sup> our analyses indicated that the occurrences of ETS family gene fusions are mutually exclusive, suggesting ETS fusions are functionally redundant. For example, our analysis revealed the mutual exclusivities between *ERG* fusions and *ETV1* fusions ( $P = 0.001$ ,  $q$  value = 0.012), and between *ERG* fusions and *ETV4* fusions ( $P = 0.036$ ,  $q$  value = 0.19) (Figure 4). Besides gene fusions, many other common genomic alterations have been identified in prostate cancer, including somatic mutations in *SPOP*, *TP53*, *FOXA1*, *KMT2C*, *KMT2D*, and *ATM*; and deletions in *RB1*, *PTEN*, and *LRP1B*. We confirmed the findings from a previous TCGA study that the *ERG* fusions co-occurred with *PTEN* deletion ( $q$  value < 0.001), but mutually exclusive with *SPOP* mutation ( $q$  value < 0.001) and *FOXA1* mutation ( $q$  value < 0.001).<sup>27</sup> Besides, we also uncovered the mutual exclusivities between *ERG* fusions and *KMT2D* mutation ( $q$  value = 0.007)

and between *ERG* fusions and *LRP1B* deletion ( $q$  value = 0.021) (Figure 4 and Supplementary Table 7). These results suggest that *ERG* fusion and *PTEN* deletion may work together during prostate tumorigenesis, which was supported by a previous study showing that transgenic overexpression of *ERG* in mouse prostate tissue promotes acceleration and progression of high-grade prostatic intraepithelial neoplasia (HGPIN) to prostatic adenocarcinoma in the heterozygous *PTEN* background.<sup>41</sup> The molecular mechanism underpinning the mutual exclusivity between *ERG* fusion and most somatic alterations is not fully understood. One potential explanation could be that the coexistence of *ERG* fusion and other somatic alterations provide, if not deleterious, no beneficial effects on tumor cell survival.

#### *Transcriptional consequences of ETS fusions*

Due to the prevalence of ETS fusions in prostate cancer (Figures 1B, 2B, and 3C and D), we sought to investigate the downstream transcriptional consequence of ETS fusions. To this end, we compared TCGA tumor samples carrying *ERG*, *ETV1*, or *ETV4* fusion with the TCGA samples without any gene fusions or other common genomic alterations (Supplementary Table 8). Overall, we identified 1059, 106, and 93 DEGs in *ERG*, *ETV1*, and *ETV4* fusion samples, respectively (Supplementary Tables 9–11). The numbers of DEGs identified in *ERG* fusion samples are significantly larger than those in the other 2 groups, most likely due to its larger sample size and higher statistical power





**Figure 5.** Evaluations of the transcriptional consequences of ETS fusions. (A), (B), and (C) volcano plots of differentially expressed genes regulated by *ERG* fusion, *ETV1* fusion, and *TEV4* fusion, respectively.  $\text{FDR} \leq 0.01$  and  $|\log_2(\text{FC})| \geq 1$  were used to determine differentially expressed genes in *ERG* fusion and  $\text{FDR} \leq 0.05$  and  $|\log_2(\text{FC})| \geq 1$  were used to determine differentially expressed genes in *ETV1* and *ETV4* fusion. (D) Venn diagrams denoting the overlaps between differentially expressed genes detected from the 3 fusion groups. (E) The mutually exclusive relationships between *ERG* and *ETV1* overexpression in TCGA samples. All TCGA samples were divided into 6 groups: *ERG* fusion (red), *ETV1* fusion (purple), *ETV4* fusion (orange), *ERG* and *ETV1* fusion (cyan), *ERG* and *ETV4* fusion (green), and no ETS fusion (gray). (F) The mutually exclusive relationships between *ERG* and *ETV4* overexpression in TCGA samples. ETS indicates E26 transformation-specific; FC, fold change; FDR, false discovery rate; TCGA, The Cancer Genome Atlas.

(Supplementary Table 8). There were 2.5 times more down-regulated genes ( $n = 761$ ) than upregulated genes ( $n = 298$ ) in *ERG* fusion samples (Figure 5A), and these down-regulated genes significantly enriched in “H3K27me3,” EED, and SUZ12

target genes (Supplementary Table 12). This result is consistent with the previous finding that *ERG* disrupts AR signaling and induces a Polycomb-mediated repressive epigenetic program.<sup>13</sup> On the contrary, almost all the DEGs identified in *ETV1* fusion

samples and *ETV4* fusion samples are upregulated, with only 5 and 1 gene identified as down-regulated, respectively (Figure 5B and C). These data suggest *ERG* might play different molecular functions than *ETV1* and *ETV4* in prostate cancer. As we expected, the expressions of *ERG*, *ETV1*, and *ETV4* themselves were elevated significantly in the corresponding fusion samples, putting them on the top of the identified DEG list (Figure 5A-C). The upregulated gene lists in *ERG*, *ETV1*, and *ETV4* fusion samples are significantly overlapped (Figure 5D), and the genes upregulated in *ERG* and *ETV1* fusions are also significantly overlapped with the *TMPRSS2-ERG* gene signature identified by a previous study<sup>42</sup> (Supplementary Table 12). We detected 6 genes (*AC004947.2*, *COL9A2*, *GDF11*, *GRPR*, *RP11-431J24.2*, and *SLC22A16*) that were commonly upregulated in all 3 fusion groups. Intriguingly, *ETV1* and *ETV4* expressions were significantly decreased in *ERG* fusion-positive samples (Figure 5D), confirming that overexpression of *ERG*, *ETV1*, and *ETV4* are mutually exclusive in prostate cancers (Figure 5E and F and Supplementary Figure 7). Except for *ETV1* and *ETV4*, *KRT6B*, and *KRT223P* (pseudogene) were the other 2 genes we identified as overexpressed in *ETV1/ETV4* fusions samples but under-expressed in *ERG* fusion samples.

#### The DNA methylome of ETS fusions

We analyzed the DNA methylation profiles of *ERG* fusion samples (n = 52), *ETV1* fusion samples (n = 5), *ETV4* fusion samples (n = 5), control samples (ie, samples without ETS fusion and other common genomic alterations, n = 71), and normal samples (n = 45). A previous TCGA study found that only one-third of *ERG* fusion-positive samples were hypermethylated.<sup>27</sup> We found that 56% of *ERG* fusion samples were hypermethylated with the averaged beta value of 0.5 or higher, and the *ERG* fusion samples had significantly higher methylation levels than those of control samples, normal samples, and *ETV1* fusion samples. However, the difference in methylation level was not statistically significant between *ERG* and *ETV4* fusion samples (Figure 6A). Principal components analysis using the top 10 000 most variable CpGs revealed distinct clusters for *ERG* fusion samples, control samples, and normal samples. In particular, the first principal component (PC1) separated normal from tumor samples, and the second principal component (PC2) separated *ERG* fusion samples from control samples (Figure 6B). When comparing *ERG* fusion samples with control samples, we detected 31 066 differentially methylated CpGs, of which 81% had increased DNA methylation in *ERG* fusion samples (Supplementary Figure 8 and Supplementary Table 13). This is consistent with the result of our gene expression analysis that 72% of DEGs were down-regulated in *ERG* fusion samples (Figure 5A). Most of the DEGs had inverse correlations between gene expression and DNA methylation (Supplementary Figure 9). Gene set enrichment analyses suggest that hypermethylated CpGs in *ERG* fusion samples were significantly associated with the PRC2

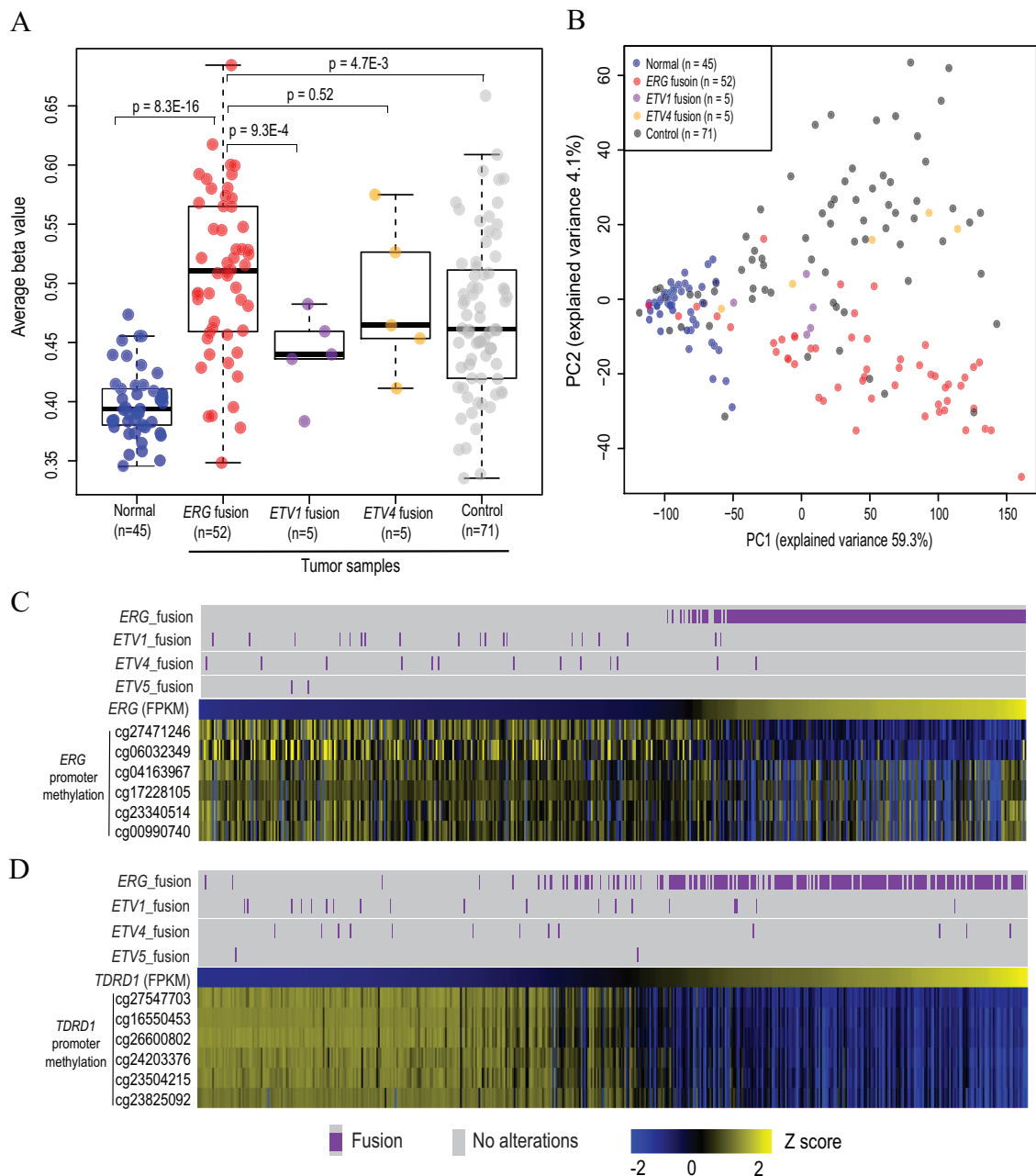
complex target genes, which were significantly down-regulated in prostate cancer samples. On the contrary, hypomethylated CpGs in *ERG* fusion samples were significantly associated with genes upregulated in prostate cancer (Supplementary Table 14).

The methylation of the *ERG* promoter was significantly decreased in fusion-positive samples compared with non-fusion samples, which is consistent with our finding that the *ERG* expression level is significantly increased in fusion-positive samples (Figure 5A). Importantly, we detected significant inverse correlations between *ERG* expression and the DNA methylations of 6 CpGs located in the promoter region (Figure 6C and Supplementary Table 15). These data suggest that the 6 CpGs could be used as surrogate biomarkers for *ERG* overexpression in prostate cancers.

The *TDRD1*, a direct target of *ERG*,<sup>43,44</sup> plays a critical role during spermatogenesis and has been identified as an important urinary biomarker for the early diagnosis of prostate cancer.<sup>45</sup> We detected a strong negative correlation between the promoter methylation and the gene expression of *TDRD1*, with Pearson's correlation coefficients ranging from -0.83 to -0.91 (Figure 6D and Supplementary Table 15). These results suggest that the promoter hypomethylation of *TDRD1* is an excellent surrogate of its overexpression.

#### Discussion

Prostate cancer is a heterogeneous disease with variable cause and androgen deprivation therapy (ADT) response, as a result of its genetic heterogeneity.<sup>46</sup> Although thousands of fusion candidates have been reported in prostate cancer, most of them have not been cross-validated by independent cohorts. In this study, we re-evaluated the 2727 gene fusions with the TCGA PRAD cohort. We found that about 10% of reported fusions can be confirmed by TCGA RNA-seq data, and only 4% are tumor-specific. Such a low validation rate suggests that most of those identified fusions are rare sporadic events, tumor stage-specific events (such as CRPC), or false positives. In addition, our analysis shows that only 4 fusions (*TMPRSS2-ERG*, *SLC45A2-AMACR*, *MLLT11-GABPB2*, and *ARHGEF38-INTS12*) were statistically significantly enriched in tumor samples compared with normal samples. Other tumor-specific fusions did not reach statistical significance, which is most likely due to low occurrences. For example, *TMPRSS2-ETV1* was only detected in 4 tumor samples. Our analysis points to 3 possible mechanisms driving gene fusions in prostate cancer. First, we found fusions with both genes located on the same DNA strand (such as *SLC45A2-AMACR*) occur more frequently than fusions with 2 genes located on different strands, suggesting transcriptional read-through may be the underpinning mechanism of these fusions. Second, our data indicate that ETS family members *ERG*, *ETV1*, and *ETV4* were fused to 26, 22, and 14 different gene partners, respectively, showing evidence that the ETS gene rearrangement may play a key role in driving ETS fusions in prostate cancer. Finally, we found



**Figure 6.** DNA methylation landscape of *ERG* fusion samples. (A) Comparison of the average DNA methylation levels between 5 groups including “normal” samples ( $n = 45$ , blue dots), “*ERG* fusion” samples ( $n = 52$ , red dots), *ETV1* fusion samples ( $n = 5$ , purple dots), *ETV4* fusion samples ( $n = 5$ , orange dots), and control samples ( $n = 71$ , gray dots). Statistical differences between different groups were evaluated using the 2-sample *t*-test. DNA methylation level is measured by beta value (y-axis) which is the ratio of methylated probe intensity and the overall intensity ( $0 \leq \text{beta value} \leq 1$ ). (B) PCA 2-dimensional map of the above 5 groups. (C) Correlation between ETS fusion status, *ERG* expression (measured by FPKM), and DNA methylation values of 6 promoter CpGs. Correlation between ETS fusion status, *TDRD1* expression (measured by FPKM), and DNA methylation values of 6 promoter CpGs. ETS indicates E26 transformation-specific; FPKM, fragments per kilobase of transcript per million mapped reads; PCA, principal components analysis.

AR-regulated genes *TMPRSS2*, *SCL45A3*, *AR*, and *FOXA1* have 23, 9, 7, and 7 gene fusion partners, respectively, indicating that AR binding mediated chromatin looping may facilitate the fusions of these genes.

The ETS fusion is the most well-known genomic aberration in prostate cancer. Among the 496 tumor samples in the TCGA PRAD cohort, we detected ETS fusions in 230 (46%) samples, of which 199, 20, 13, and 2 tumors harboring *ERG*, *ETV1*,

*ETV4*, and *ETV5* fusions, respectively. It is worth noting that none of these ETS fusions were detected in the normal samples, suggesting they are highly tumor-specific. It is not surprising to find that *ERG* and *ETV1* were predominantly fused to *TMPRSS2*, as reported previously.<sup>27</sup> However, we found *ETV4* was not as “selective” as *ERG* and *ETV1*, with similar fusion frequencies to *SLC45A3*, *TMPRSS2*, *EST14*, *FLJ35294*, and *HERV-K17* (Supplementary Figure 6). Besides, contrary to

*ERG* and *ETV1*, whose fusion partners are predominantly located on a different chromosome, *ETV4* is mainly fused to intra-chromosomal partner genes (Figure 1B). Finally, for all the genes that were upregulated by fusions involving *ERG*, *ETV1*, and *ETV4*, the number of overlapped genes between *ERG* fusion and *ETV1* fusion was much higher than that of *ERG* fusion and *ETV4* fusion (Figure 5D). All the above findings suggest a different mechanism for *ETV4* fusions.

We detected 4 fusions that were significantly enriched in tumor samples including *TMPRSS2-ERG*, *SLC45A2-AMACR*, *MLLT11-GABPB2*, and *ARHGEF38-INTS12*, all of which are intra-chromosomal. However, in contrast to *TMPRSS2-ERG*, whose fusion partners are located 2.8 Mb away, fusion partners of *SLC45A2-AMACR*, *MLLT11-GABPB2*, and *ARHGEF38-INTS12* are immediate neighbors, separated only by 2.3, 2.1, and 1.7 Kb, respectively. Therefore, they seem to be transcript fusions produced by transcriptional read-through or intergenic splicing of adjacent genes. The *SLC45A2-AMACR* fusion was first discovered in a prostate cancer study in 2014 and was reported as the most frequent fusion event observed among all prostate samples studied.<sup>47</sup> It has also been identified in bladder cancer,<sup>48</sup> lung cancer,<sup>49</sup> and liver cancer.<sup>50</sup> Few studies have been focused on *MLLT11-GABPB2* fusion and *ARHGEF38-INTS12* fusion, probably due to their low expression. Further investigations are needed to clarify their clinical significance in prostate cancer.

## Conclusions

In this study, we explored the genomic and transcriptomic characteristics of 2727 gene fusions reported by 36 prostate cancer studies and used the TCGA PRAD cohort to re-evaluate their authenticity, incidence rate, and tumor specificity. While most (76%) genes were fused to 1 partner, we found ETS family genes (*ERG*, *ETV1*, *ETV4*), AR signaling pathway genes (*TMPRSS2*, *SLC45A3*, *AR*, *FOXA1*), tumor suppressor genes (*TP53*, *PTEN*), and proto-oncogenes (*BRAF*, *PIK3C2A*) have at least 6 fusion partners, suggesting they are the drivers of gene fusions in prostate cancer. More than 90% of the 2727 gene fusions cannot be validated by any of the TCGA PRAD RNA-seq samples, and only 120 fusions (4%) showed tumor specificity. Further analyses revealed that ETS family genes and AR-regulated genes are significantly enriched in these prostate cancer-specific fusions. The *ERG*, *ETV1*, and *ETV4* fusions were mutually exclusive with each other, suggesting their functional redundancy. In addition, we found *ERG* fusions co-occurred with *PTEN* deletions, but mutually exclusive with *SPOP* mutations, *FOXA1* mutations, *KMT2D* mutations, and *LRP1B* deletions. There were 2.5 times more down-regulated genes than upregulated genes in *ERG* fusion samples, whereas almost all the DEGs identified from *ETV1/ETV4* fusion samples were upregulated, showing that *ERG* and *ETV1/ETV4* fusions may play different roles in prostate cancer. Promoter methylations of *ERG* and *TDRD1* exhibited excellent reverse correlations with their RNA expressions, suggesting the DNA

methylation can be used as a good surrogate for the overexpression of these 2 important biomarkers in prostate cancer.

## Acknowledgments

The results shown here are in part based on data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>. We thank the contribution of the TCGA Prostate Cancer Work Group.

## Author Contributions

LW, J-PK, and HH conceived and designed the study; TW, JL, and LW performed the statistical analyses of the data; LW, TW, and JL wrote the manuscript with input from all authors; TW and LW edited and revised the manuscript. All authors have read and approved the final manuscript.

## Availability of Data and Materials

The datasets supporting the conclusions of this article are included within the additional files. FusionVet is available from GitHub: <https://github.com/liguowang/FusionVet>

## Consent for Publication

All authors read and approved the manuscript.

## ORCID iD

Liguo Wang  <https://orcid.org/0000-0003-2072-4826>

## Supplemental Material

Supplemental material for this article is available online.

## REFERENCES

- Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. *Nat Rev Cancer*. 2015;15:371-381.
- Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer*. 2007;7:233-245.
- Linardic CM. PAX3-FOXO1 fusion gene in rhabdomyosarcoma. *Cancer Lett*. 2008;270:10-18.
- Barr FG. Gene fusions involving PAX and FOX family members in alveolar rhabdomyosarcoma. *Oncogene*. 2001;20:5736-5746.
- Martelli MP, Sozzi G, Hernandez L, et al. EML4-ALK rearrangement in non-small cell lung cancer and non-tumor lung tissues. *Am J Pathol*. 2009;174:661-670.
- Soda M, Choi YL, Enomoto M, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*. 2007;448:561-566.
- Djabali M, Selleri L, Parry P, Bower M, Young BD, Evans GA. A trithorax-like gene is interrupted by chromosome 11q23 translocations in acute leukaemias. *Nat Genet*. 1992;2:113-118.
- Tumor fusion gene data portal. <https://www.tumorfusions.org/>.
- Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin*. 2020;70:7-30.
- Wu D, Zhang C, Shen Y, Nephew KP, Wang Q. Androgen receptor-driven chromatin looping in prostate cancer. *Trends Endocrinol Metab*. 2011;22:474-480.
- Berger MF, Lawrence MS, Demichelis F, et al. The genomic complexity of primary human prostate cancer. *Nature*. 2011;470:214-220.
- Tomlins SA, Rhodes DR, Perner S, et al. Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer. *Science*. 2005;310:644-648.
- Yu J, Yu J, Mani RS, et al. An integrated network of androgen receptor, polycomb, and *TMPRSS2-ERG* gene fusions in prostate cancer progression. *Cancer Cell*. 2010;17:443-454.
- Hermans KG, van Marion R, van Dekken H, Jenster G, van Weerden WM, Trapman J. *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*. 2006;66:10658-10663.

15. Perner S, Demichelis F, Beroukhir R, et al. TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer. *Cancer Res.* 2006;66:8337-8341.
16. Hessels D, Schalken JA. Recurrent gene fusions in prostate cancer: their clinical implications and uses. *Curr Urol Rep.* 2013;14:214-222.
17. Mehra R, Tomlins SA, Shen R, et al. Comprehensive assessment of TMPRSS2 and ETS family gene aberrations in clinically localized prostate cancer. *Mod Pathol.* 2007;20:538-544.
18. Nam RK, Sugar L, Wang Z, et al. Expression of TMPRSS2:ERG gene fusion in prostate cancer cells is an important prognostic factor for cancer progression. *Cancer Biol Ther.* 2007;6:40-45.
19. Nam RK, Sugar L, Yang W, et al. Expression of the TMPRSS2:ERG fusion gene predicts cancer recurrence after surgery for localised prostate cancer. *Br J Cancer.* 2007;97:1690-1695.
20. Tomlins SA, Laxman B, Dhanasekaran SM, et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature.* 2007;448:595-599.
21. Han B, Mehra R, Dhanasekaran SM, et al. A fluorescence in situ hybridization screen for E26 transformation-specific aberrations: identification of DDX5-ETV4 fusion protein in prostate cancer. *Cancer Res.* 2008;68:7629-7637.
22. Helgeson BE, Tomlins SA, Shah N, et al. Characterization of TMPRSS2:ETV5 and SLC45A3:ETV5 gene fusions in prostate cancer. *Cancer Res.* 2008;68:73-80.
23. Paulo P, Barros-Silva JD, Ribeiro FR, et al. FLI1 is a novel ETS transcription factor involved in gene fusions in prostate cancer. *Genes Chromosomes Cancer.* 2012;51:240-249.
24. Jiang M, Ma Y, Chen C, et al. Androgen-responsive gene database: integrated knowledge on androgen-responsive genes. *Mol Endocrinol.* 2009;23:1927-1933.
25. The Cancer Genome Atlas (TCGA). dbGaP study accession: phs000178.v11.p8. [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000178.v11.p8](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v11.p8).
26. 2016\_01\_28 data snapshot. [http://gdac.broadinstitute.org/runs/stddata\\_\\_2016\\_01\\_28/samples\\_report/PRAD\\_Notifications.html](http://gdac.broadinstitute.org/runs/stddata__2016_01_28/samples_report/PRAD_Notifications.html).
27. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell.* 2015;163:1011-1025.
28. Sboner A, Habegger L, Pflueger D, et al. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol.* 2010;11:R104.
29. Wang K, Singh D, Zeng Z, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* 2010;38:e178.
30. Torres-Garcia W, Zheng S, Sivachenko A, et al. PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics.* 2014;30:2224-2226.
31. Dehghanasiri R, Freeman DE, Jordanski M, et al. Improved detection of gene fusions by applying statistical methods reveals oncogenic RNA cancer drivers. *Proc Natl Acad Sci U S A.* 2019;116:15524-15533.
32. Tomlins SA, Laxman B, Varambally S, et al. Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia.* 2008;10:177-188.
33. Genomic Data Commons Data Portal. <https://portal.gdc.cancer.gov/projects/TCGA-PRAD>.
34. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 2010;11:R25.
35. Wei T, Nie J, Larson NB, et al. CpGtools: a python package for DNA methylation analysis. *Bioinformatics.* 2019;btz916.
36. Gene Set Enrichment Analysis. <https://www.gsea-msigdb.org/gsea/msigdb/index.jsp>.
37. Pomerantz MM, Li F, Takeda DY, et al. The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat Genet.* 2015;47:1346-1351.
38. He Y, Lu J, Ye Z, et al. Androgen receptor splice variants bind to constitutively open chromatin and promote abiraterone-resistant growth of prostate cancer. *Nucleic Acids Res.* 2018;46:1895-1911.
39. He HH, Meyer CA, Shin H, et al. Nucleosome dynamics define transcriptional enhancers. *Nat Genet.* 2010;42:343-347.
40. Svensson MA, LaFargue CJ, MacDonald TY, et al. Testing mutual exclusivity of ETS rearranged prostate cancer. *Lab Invest.* 2011;91:404-412.
41. Carver BS, Tran J, Gopalan A, et al. Aberrant ERG expression cooperates with loss of PTEN to promote cancer progression in the prostate. *Nat Genet.* 2009;41:619-624.
42. Setlur SR, Mertz KD, Hoshida Y, et al. Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer. *J Natl Cancer Inst.* 2008;100:815-825.
43. Boormans JL, Korsten H, Ziel-van der Made AJ, et al. Identification of TDRD1 as a direct target gene of ERG in primary prostate cancer. *Int J Cancer.* 2013;133:335-345.
44. Kacprzyk LA, Laible M, Andrasiuk T, et al. ERG induces epigenetic activation of Tudor domain-containing protein 1 (TDRD1) in ERG rearrangement-positive prostate cancer. *PLoS ONE.* 2013;8:e59976.
45. Leyten GH, Hessels D, Smit FP, et al. Identification of a candidate gene panel for the early diagnosis of prostate cancer. *Clin Cancer Res.* 2015;21:3061-3070.
46. Lindberg J, Klevebring D, Liu W, et al. Exome sequencing of prostate cancer supports the hypothesis of independent tumour origins. *Eur Urol.* 2013;63:347-353.
47. Yu YP, Ding Y, Chen Z, et al. Novel fusion transcripts associate with progressive prostate cancer. *Am J Pathol.* 2014;184:2840-2849.
48. Yoshihara K, Wang Q, Torres-Garcia W, et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene.* 2015;34:4845-4854.
49. Klijn C, Durinck S, Stawiski EW, et al. A comprehensive transcriptional portrait of human cancer cell lines. *Nat Biotechnol.* 2015;33:306-312.
50. Yu YP, Tsung A, Liu S, et al. Detection of fusion transcripts in the serum samples of patients with hepatocellular carcinoma. *Oncotarget.* 2019;10:3352-3360.