Published in final edited form as:
J Pathol. 2013 June ; 230(2): 174–183. doi:10.1002/path.4186.

Nucleotide resolution analysis of TMPRSS2 and ERG rearrangements in prostate cancer

Christopher Weier[1], **Michael C. Haffner**[1], **Timothy Mosbruger**[1], **David M. Esopi**[1], **Jessica Hicks**[2], **Qizhi Zheng**[2], **Helen Fedor**[2], **William B. Isaacs**[1,2,3], **Angelo M. De Marzo**[1,2,3], **William G. Nelson**[1,2,3], and **Srinivasan Yegnasubramanian**[1,†]

[1]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, Maryland, 21231 USA

[2]Department of Pathology, Johns Hopkins University, Baltimore, Maryland, 21231 USA

[3]Brady Urological Institute, Johns Hopkins University, Baltimore, Maryland, 21231 USA

Abstract

TMPRSS2-ERG rearrangements occur in approximately 50% of prostate cancers and therefore represent one of the most frequently observed structural rearrangements in all cancers. However, little is known about the genomic architecture of such rearrangements. We therefore designed and optimized a pipeline involving target-capture of *TMPRSS2* and *ERG* genomic sequences coupled with paired-end next generation sequencing to resolve genomic rearrangement breakpoints in

†Correspondence and reprint requests to: Srinivasan Yegnasubramanian, 1650 Orleans Street, David H. Koch Cancer Research Building Room 145, Baltimore, MD 21128, USA; phone: (410)502-3425; fax: (410)502-9817; syegnasu@jhmi.edu.

Conflict of Interest Disclosures: A.M.D. is currently employed at Predictive Biosciences, Inc., Lexington MA, while being part-time adjunct Professor of Pathology, Oncology, and Urology at the Johns Hopkins University School of Medicine. No funding or other support was provided by Predictive Biosciences, Inc. for any of the work in this manuscript. The terms of the relationship between A.M.D. and Predictive Biosciences, Inc. are managed by the Johns Hopkins University in accordance with its conflict-of-interest policies.

AUTHOR CONTRIBUTIONS

CW, MCH, WGN, WBI, AMD, and SY conceived the study design and experiments. CW, MCH, DME, JH, QZ, HF carried out experiments. CW, MCH, TM, and SY performed data analysis. WBI and AMD contributed archived tissue samples. AMD performed pathological analyses. CW, MCH, and SY wrote the manuscript. All authors were involved in editing and approving the final manuscript.

LIST OF SUPPORTING INFORMATION

Supplementary Methods.
Supplementary Figure 1. Detection and rearrangement junctions in reference samples.
Supplementary Figure 2. Sensitivity and specificity determination of geBACS by FISH.
Supplementary Figure 3. *In situ* RNA detection of *ETV4* and *ETV5* in cases harboring *TMPRSS2-ETV4* and *TMPRSS2-ETV5* rearrangements.
Supplementary Figure 4. The most common TMPRSS2-ERG rearrangement involves the first intron of TMPRSS2 and the third intron of ERG resulting in an exon 1 – exon 4 TMPRSS2-ERG fusion transcript.
Supplementary Figure 5. Inter- and Intra- genic rearrangements in TMPRSS2 and ERG are only detected in tumor, and not in normal adjacent tissue.
Supplementary Figure 6. Detailed characterization of rearrangements in Case 995.
Supplementary Figure 7. Detailed characterization of rearrangements in VCaP cell line.
Supplementary Figure 8. Genomic rearrangements involving TMPRSS2, SLC45A3, ERG, ETV4 and ETV5.
Supplementary Figure 9. Cases showing complex genomic rearrangements involving TMPRSS2 and ERG.
Supplementary Figure 10. Strong enrichment of microhomologies at breakpoint junctions.
Supplementary Table 1. Clinical and pathological characteristics of study samples
Supplementary Table 2. Detailed clinical and pathological characteristics of the cohort
Supplementary Table 3. Junction specific primer sets
Supplementary Table 4, Characteristics of rearrangement junctions identified in patients, reference and cell line samples
Supplementary File. Synthesized RNA baits

TMPRSS2 and *ERG* at nucleotide resolution in a large series of primary prostate cancer specimens (n = 83). This strategy showed >90% sensitivity and specificity in identifying *TMPRSS2-ERG* rearrangements, and allowed identification of intra- and inter-chromosomal rearrangements involving *TMPRSS2* and *ERG* with known and novel fusion partners. Our results indicate that rearrangement breakpoints show strong clustering in specific intronic regions of *TMPRSS2* and *ERG*. The observed *TMPRSS2-ERG* rearrangements often exhibited complex chromosomal architecture associated with several intra- and inter-chromosomal rearrangements. Nucleotide resolution analysis of breakpoint junctions revealed that the majority of *TMPRSS2* and *ERG* rearrangements (~88%) occurred at or near regions of microhomology or involved insertions of one or more base pairs. This architecture implicates nonhomologous end joining (NHEJ) and microhomology mediated end joining (MMEJ) pathways in the generation of such rearrangements. These analyses have provided important insights into the molecular mechanisms involved in generating prostate cancer-specific recurrent rearrangements.

Keywords

TMPRSS2; *ERG*; rearrangement; hybrid capture; targeted next generation sequencing; prostate cancer; genomic breakpoint detection

INTRODUCTION

The genomic landscape of prostate cancer often features complex structural alterations [1]. Structural rearrangements involving androgen regulated genes and ETS family transcription factors are the most common recurrent genetic alteration in prostate cancers, occurring in 30–70% of cases [2]. The most common of these rearrangements involves fusion of the androgen regulated gene, *TMPRSS2*, with the ETS transcription factor, *ERG*, both located on chromosome 21, resulting in an androgen-regulated *TMPRSS2-ERG* fusion transcript [3]. Several hypotheses have suggested a key role for AR signaling events in the generation of these genomic rearrangements [2, 4]. Although the prevalence and biological relevance of the *TMPRSS2-ERG* fusion genes have been studied extensively [2, 5], little is known about the underlying genomic architecture of these rearrangements.

The complex genomic architecture of the *TMPRSS2* and *ERG* loci has made previous investigations of genomic breakpoints very challenging. Labor-intensive and costprohibitive methods such as long-range PCR followed by Sanger sequencing or whole genome sequencing have thus far yielded only a handful of *TMPRSS2-ERG* genomic breakpoints [1, 6, 7]. Large-scale analysis of *TMPRSS2-ERG* rearrangements has been restricted to the detection of fusion mRNA transcripts or methods such as visualization of gross chromosomal alterations using fluorescence in situ hybridization (FISH) and array CGH and SNP array methodologies which are incapable of resolving rearrangement breakpoints at nucleotide resolution [3, 8, 9].

Detailed knowledge on the genomic structure of *TMPRSS2-ERG* rearrangement breakpoints at nucleotide resolution could help to elucidate sequence characteristics associated with these rearrangements and could indirectly provide mechanistic insight into the processes involved in the generation of these genomic fusions [10]. Furthermore, since such rearrangements are prostate cancer-specific [11], the ability to detect rearrangement breakpoints efficiently could allow development of personalized biomarkers for prostate cancer detection and disease monitoring [12, 13].

We have developed an efficient pipeline for identifying nucleotide-resolution genomic breakpoints of rearrangements involving *TMPRSS2* and *ERG* using targeted hybrid-capture

coupled with paired-end next generation sequencing [14, 15]. We applied our pipeline to a large series of primary prostate cancers (n=83) and control samples, creating the most extensive catalog to date of *TMPRSS2* and *ERG* rearrangement breakpoint sequences. These analyses revealed several insights into the sequence characteristics of these recurrent rearrangements in prostate cancer.

MATERIALS AND METHODS

Prostate tissues and genomic DNA

Fresh frozen blocks from prostate tissues were obtained from 83 men undergoing radical prostatectomy for treatment of prostate adenocarcinoma. Clinicopathological characteristics of all 83 adenocarcinomas are summarized in Supplementary Table 1. Tissues were trimmed to yield sections containing >60% tumor nuclei and subjected to DNA isolation (Supplementary Table 2), as previously described [16]. Genomic DNA was isolated from the *TMPRSS2-ERG* rearrangement positive VCaP prostate cancer cell line as previously described [6, 16]. Reference specimens from 3 samples in which the *TMPRSS2-ERG* rearrangement breakpoints had been determined previously using Sanger sequencing were included as controls [6]. All studies were carried out in accordance with the Helsinki Declaration of 1975, as revised in 1983, and under approval by the Johns Hopkins Institutional Review Board.

Genomic breakpoint identification by targeted hybrid-capture coupled with next generation sequencing

Samples were divided into 6 groups, each containing 14 genomic DNA samples from primary prostate cancers or the VCaP cell line and one reference sample. The DNA samples within each group were pooled (600ng/sample), fragmented by sonication on a Covaris S2 Sonicator (Covaris, Inc, Woburn, MA), and size selected to a modal length of ~200 bp. Sequencing libraries were generated using the NEBNext DNA Sample Prep Reagent Set and Genome Analyzer sequencing adapters (NEB, Ipswich, MA), using the manufacturer's protocols. Libraries were then enriched for *TMPRSS2* and *ERG* sequences using the Agilent SureSelect Custom Target Enrichment System (Agilent, Santa Clara, CA). Briefly, a custom RNA bait library was designed using the Agilent eArray online tool and synthesized as part of the Agilent Custom SureSelect Kit. From the sense strands of *TMPRSS2* (hg18, chr21:41738351–41841779; 53,478 bp after repeat-masking) and *ERG* (chr21:38650671–38979461; 222,161 bp after repeat-masking) loci, 9,995 RNA baits were generated, excluding known repetitive sequences and tiled to achieve ~5x/base bait redundancy (see Supplementary File for chromosomal locations of baits). Pooled genomic DNA was hybridized with the RNA bait library for 24 hours at 60° C. Streptavidin-coated, magnetic beads (Dynabeads, Invitrogen, Carlsbad, CA) were used to isolate hybridized fragments and the enriched pools were amplified for 18 cycles using adapter-specific primers. Paired-end sequencing with 50 bp reads was carried out on an Illumina Genome Analyzer (Illumina, San Diego, CA) following the manufacturer's protocols.

Bioinformatics analysis

Paired-end reads, consisting of a P1 read and a P2 read, were aligned to the human reference (UCSC build hg18) using a tiered application of the Bowtie2 short read alignment tool (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) [17]. An initial end-to-end paired alignment (using the default settings and the "*--very-fast*" argument) was used to eliminate "concordant" pairs aligning within five hundred bp of each other, as these reads likely represent contiguous genomic segments that do not span a rearrangement. The remaining pairs were then subjected to a second local paired alignment strategy (using the default settings and the "*--very-sensitive-local*" argument). Discordant pairs and pairs characterized

as being neither concordant nor discordant (but otherwise having local alignment within the genome) were retained and filtered to those aligning on different chromosomes or greater than 8 kb apart as these likely represent candidate structural rearrangements. Pairs were subsequently filtered to retain those in which at least one read mapped within the baited regions of *TMPRSS2* or *ERG* (chr21:41738351–41841779; chr21:38650671–38979461). Duplicate pairs were removed to exclude those that may have arisen through clonal amplification. Pairs were sorted by genomic position of the P1 alignment and binned such that P1 reads clustered within a 200 base pair window. Bins with a minimum of 3 pairs were then sub-binned based on corresponding P2 reads that clustered into a 200 bp window. All such sub-divided bins containing at least three paired-end reads with unique start positions for both reads were considered as spanning a putative rearrangement.

Bins often contained a mix of read pairs that flanked the putative junction and reads that overlapped the putative junction manifesting as partial alignments. For such partially aligned reads, a manual gapped alignment was performed using BLAT [18] which allowed identification of nucleotide resolution rearrangement junctions. Single pairs with one read aligning within *TMPRSS2* and the other aligning within *ERG*, but otherwise lacking the additional pairs to be placed into a putative junction bin, were rescued for manual reanalysis in order to maximize sensitivity; 13 such single-read pairs were identified.

To test whether *TMPRSS2-ERG* breakpoints were enriched for microhomology, we created 10,000 simulated datasets of 22 randomly generated breakpoints, each joining a position in the introns of *TMPRSS2* with a position from the introns of *ERG*; for each of these 220,000 simulated rearrangements, we determined the presence and extent of microhomology directly at the junction. An empirical p-value for significance of enrichment of microhomologies in our observed rearrangement breakpoints was calculated as the number of simulated datasets with at least the number of microhomology-containing breakpoints observed in our experimental dataset divided by 10,000 (the total number of simulated datasets). Additionally, the statistical significance of the observed vs. simulated distribution of the size of microhomologies at rearrangement breakpoints was assessed using the Mann-Whitney U Test. The significance of clustering of *TMPRSS2-ERG* rearrangements within a 30 kb region in intron 3 of *ERG* was estimated by χ^2 test comparing the number of observed breakpoints in that region to the number that would be expected if the breakpoints were uniformly distributed across intron 3.

PCR confirmation and case-specific breakpoint assignment

Identified genomic breakpoint junctions were confirmed through a tiered PCR strategy. Initially, PCR primers designed for each putative junction were used to confirm the presence of that breakpoint in the pooled genomic DNA group from which the breakpoint was identified. Confirmed junctions were advanced to the second tier, in which the same PCR was carried out for each of the individual genomic DNA samples from the pool (see Supplementary Table 3 and Supplementary Methods for PCR primers and methodology). This tiered approach was used to confirm and assign each identified junction to a single subject in the original pooled group. PCR products from the second tier were gel purified after electrophoresis, sub-cloned into pCR®2.1-TOPO® vectors (Invitrogen) and analyzed by Sanger sequencing.

Fluorescence *In Situ* Hybridization (FISH)

FISH probes targeting *TMPRSS2* and *ERG* flanking sequences (TMP ERG del-TECT™, product # CYMO-21D-23–100) were obtained from CymoGenDx (Irvine, CA). Prior to hybridization, FFPE slides were deparaffinized and pretreated in 0.01 N HCl supplemented with 750 U/ml pepsin (Sigma-Aldrich, cat#P6887) for 10 min at 37°C. Probes were applied

on the slide, denatured for 5 min at 95°C, and then incubated at 37°C in a humid chamber (StatSpin ThermoBrite, IRISInc, MA). Slides were then washed in Wash buffer (2xSSC, 0.3 % NP-40) for 2 min at 71 °C, counterstained with DAPI and mounted with Prolong Gold (Invitrogen). In each case the rearrangement status was evaluated in a minimum of 100 cells similar to previous reports [19, 20].

Immunohistochemistry and *in situ* mRNA detection

Immunohistochemical detection of ERG protein expression in paraffin-embedded prostate cancer samples was performed as described previously [21], except that the ERG-specific monoclonal antibody [22] was from Biocare Medical clone 9FY (Concord, CA). *In situ* detection of *ETV4* and *ETV5* transcripts in paraffin-embedded-formalin-fixed tissues was performed using the Affymetrix QuantiGene ViewRNA ISH Tissue Assay kit (Affymetrix, Santa Clara, CA) following the manufacturer's instructions.

RESULTS

genomic Breakpoint Analysis by Capture Sequencing (geBACS) identifies genomic rearrangement sites in *TMPRSS2* and *ERG*

To identify genomic breakpoints in *TMPRSS2* and *ERG* in a large series of prostate cancer samples, we adapted a hybrid-capture strategy combined with targeted next generation sequencing [13, 23]. The geBACS workflow is outlined in Figure 1. Briefly, DNA samples were pooled, subjected to target-enrichment for *TMPRSS2* and *ERG* sequences, and sequenced on the Illumina Genome Analyzer. This strategy resulted in a target-specific enrichment of nearly 600-fold over the non-baited genomic background. An average coverage depth of 458x in the baited region was achieved, with 99% of baited regions exhibiting at least 30x coverage, and only 0.01% of baits having no coverage. Since there were 15 samples in each lane, we would expect an average coverage of 30x across all targeted base pairs for each sample. Computational analysis of these data yielded rearrangement breakpoints involving *TMPRSS2* and/or *ERG* sequences, which were then verified by targeted PCR and Sanger sequencing (Figure 1).

We included 3 reference samples (control cases 45, 66, and 77) for which we had previously determined the *TMPRSS2-ERG* genomic breakpoint [6]. We detected these junctions at nucleotide resolution in all 3 reference samples with sufficiently high read numbers (18, 24, and 21 paired-end reads overlapping or flanking the known breakpoint for control cases 45, 66, and 77 respectively; Supplementary Figure 1), thus validating the geBACS pipeline.

We identified 26 *TMPRSS2-ERG* rearrangements in 25 out of 83 primary prostate cancer specimens analyzed (Supplementary Table 4, Figure 2). To assess the sensitivity and specificity of the geBACS pipeline, we capitalized on the observation that the presence of any *TMPRSS2-ERG* rearrangement is highly correlated with positive ERG immunohistochemistry [21, 22, 24]. We performed ERG immunohistochemistry in 38 cases. Out of 13 cases that stained positive for ERG (representative case shown in Figure 2D), we identified *TMPRSS2-ERG* rearrangement breakpoints in 12 cases, indicating 92.3% sensitivity (Figure 2C). Among 25 cases that did not show ERG staining, we detected a *TMPRSS2-ERG* rearrangement in a single case by geBACS. This case may represent a false positive by the geBACS approach (indicating 96% specificity). However, as this junction was corroborated by PCR and Sanger sequencing, the lack of ERG staining may represent an IHC false negative due to disparities between the section taken for staining and the material used for DNA extraction. Furthermore, in two cases in which we detected *TMPRSS2-ETV4* and *TMPRSS2-ETV5* rearrangements, overexpression of ETV4 and ETV5 respectively was confirmed by RNA *in situ* hybridization (Supplementary Figure 3).

Because it is possible that ERG overexpression in prostate cancer may be induced by factors other than rearrangement, we also verified the sensitivity and specificity of our geBACS approach by performing FISH on a subset of samples (n = 20) using a 4 color probe system to detect *TMPRSS2-ERG* rearrangements. Similar to previous reports, we observed a high concordance between ERG immunohistochemistry and *TMPRSS2-ERG* FISH (Supplementary Figure 2) [21], [24]. Using FISH as the gold standard, the geBACS approach showed 100% sensitivity and 87.5% specificity.

Clustering and chromosomal architecture of rearrangements in *TMPRSS2* and *ERG*

While each *TMPRSS2-ERG* breakpoint was unique at the nucleotide level, we observed a pronounced clustering of rearrangement events within *TMPRSS2* and *ERG*. Out of the 26 observed *TMPRSS2-ERG* rearrangements, 15 (57%) joined the first intron of *TMPRSS2* with the third intron of *ERG*; 5 (19%) joined the second intron of *TMPRSS2* with the third intron of *ERG*; 2 (8%) joined the fifth intron of *TMPRSS2* with the third intron of *ERG*; 2 (8%) joined the third intron of TMPRSS2 and the third intron of ERG; 1 (4%) joined the second intron of *TMPRSS2* and the fourth intron of *ERG*; and 1 (4%) joined the fourth exon of *TMPRSS2* and the third intron of *ERG* (Figure 2A, Supplementary Table 4). The predominance of genomic breakpoints occurring within intron 1 of *TMPRSS2* and intron 3 of *ERG* is consistent with previous studies showing that the most common *TMPRSS2-ERG* mRNA fusion transcript juxtaposes exon 1 of *TMPRSS2* with exon 4 of *ERG* (Supplementary Figure 4) [2, 3, 5]. Furthermore, we observed a significant clustering of breakpoints within a 30 kbp region within intron 3 of *ERG* (chr21:38776561–38804491; p-value < 0.0001). Analyses of GC content, transcription factor binding, and homology within this region failed to reveal unique sequence characteristics that might account for such a hotspot (data not shown).

The majority of *TMPRSS2-ERG* genomic fusion breakpoints (21/26, 80%) showed an expected orientation in which the sense strands of TMPRSS2 and ERG were fused as 5' and 3' partners respectively. The remaining five rearrangements displayed atypical orientations: three cases with the sense strand of *TMPRSS2* fused to the antisense strand of *ERG* and two cases with the 5’ portion of *ERG* fused to the 3’ portion of *TMPRSS2*. A single case contained both a typical and atypical rearrangement junction. Interestingly four of five cases with atypical *TMPRSS2-ERG* junctions exhibited *ERG* overexpression by IHC, indicating that the atypical rearrangement junctions may be part of a complex *TMPRSS2-ERG* rearrangement that ultimately resulted in *ERG* overexpression (Supplementary Table 4). These cases will require further investigation to understand the source of ERG overexpression.

Complex intra- and inter-genic rearrangements involving *TMPRSS2* and *ERG*

In addition to the rearrangements between *TMPRSS2* and *ERG*, we also observed 9 intragenic rearrangements in prostate cancer cases of which 7 fell in *TMPRSS2* and 2 in *ERG* (Figure 2A). Such intragenic *TMPRSS2* and *ERG* rearrangements were only observed in tumor tissue (Supplementary Figure 5) and were typically found in cases harboring other intergenic rearrangements at the *TMPRSS2* and *ERG* loci, suggesting that such recurrent rearrangements can often show complex architecture, as recently reported in a whole genome sequencing effort [1]. A particularly interesting case (995) involved an inversion in *ERG*, a fusion to *TMPRSS2*, and multiple intragenic rearrangements in *TMPRSS2*, likely affecting only one allele as determined by FISH, and resulting in an in-frame fusion gene producing ERG protein over expression (see Figure 2, Supplementary Figure 6). The *TMPRSS2-ERG* rearrangement in the VCaP cell line involved a *TMPRSS2* intragenic rearrangement and a *TMPRSS2-ERG* intergenic rearrangement (Supplementary Figure 7).

In addition to rearrangements within and between *TMPRSS2* and *ERG*, we also detected several rearrangements involving fusion of *TMPRSS2* or *ERG* with other genomic regions (Figure 2B). We characterized the genomic architecture of three previously described recurrent fusions: *SLC45A3-ERG, TMPRSS2-ETV4* and *TMPRSS2-ETV5* (Supplementary Figure 8) [2, 25, 26]. In the two cases that harbored *TMPRSS2-ETV4* and *TMPRSS2-ETV5* fusions, we observed *ETV4* and *ETV5* overexpression respectively using RNA *in situ* detection (Figure 2E, F). This *in situ* transcript detection method is highly specific since the *ETV5* rearranged case did not show *ETV4* over-expression and vice versa (Supplementary Figure 3). Furthermore, we found a number of previously uncharacterized rearrangements involving *TMPRSS2* or *ERG* with a novel fusion partner (Figure 2B). In total, 17 junctions were resolved in 11 different cases, 15 involving *TMPRSS2* and 2 involving *ERG* (Table 1, Figure 2B). Neither of the 2 novel *ERG* rearrangements was predicted to generate productive transcripts. Sequence alignment revealed that the majority of novel rearrangements fused coding and non-coding strands. All novel rearrangement partners were observed once except for *MX1*, a gene downstream of *TMPRSS2*, which was found to be rearranged with *TMPRSS2* in two cases (Supplementary Table 4). Several cases showed multiple complex rearrangements involving different inter- and intrachromosomal fusion partners (Supplementary Figure 9). Taken together, our data show that complex intra- and inter-chromosomal rearrangements in *TMPRSS2* and *ERG* are a common feature of prostate cancer [1].

Nucleotide architecture of *TMPRSS2-ERG* rearrangements

geBACS allowed resololution of *TMPRSS2-ERG* rearrangements at nucleotide level and subsequently categorize rearrangements into distinct patterns according to sequence characteristics occurring at the junction. First, 4 of 26 rearrangements (15.4%) involved insertion of one to two bp at the breakpoint (referred to as “non-templated insertions”) that did not align to either *TMPRSS2* or *ERG* (Figure 3A, Supplementary Table 4). Of the remaining 22 rearrangements, 11 (50%) showed a distinct transition from *TMPRSS2* to *ERG* sequences (referred to as “blunt fusions”) without any evidence of microhomology or insertions (Figure 3B, Table 1), and another 11 (50%) displayed short sequences of microhomology (1 to 4 bp) directly spanning the rearrangement junction (Figure 3C, Table 1). To determine whether this high frequency of microhomology could have arisen by chance given the sequence characteristics of *TMPRSS2* and *ERG* introns, we created 10,000 randomly generated, simulated datasets, each containing the same number of rearrangements identified in our current dataset (22), producing a total of 220,000 simulated *TMPRSS2-ERG* breakpoints. Compared to the distribution of microhomology-containing breakpoints in these simulated datasets, our observed dataset was significantly enriched for presence of microhomology ($p = 0.0087$; Supplementary Figure 10). Additionally, the extent of microhomology (number of bp showing microhomology) in our observed dataset was significantly greater than that in the simulated breakpoints ($p = 0.01$, Mann-Whitney U test). Interestingly, a subset of rearrangements from each of the three classes described above (non-templated insertions, blunt fusions, and microhomology) often exhibited microhomology of 3 to 6 bp flanking the rearrangement junction; this occurred in 17 of 26 rearrangements (65%) (Figure 3A,D, Supplementary Table 4). The high rate of microhomology occurring at or near the rearrangement junction and the presence of short non-templated insertions, which together account for 23 out of 26 rearrangements (88%), is highly indicative of NHEJ and/or MMEJ pathways. These pathways are known to be error-prone and to occur at regions of microhomology and thus may be involved in the generation of *TMPRSS2-ERG* rearrangements in prostate cancer [27, 28].

DISCUSSION

We have developed a highly sensitive and specific approach, termed geBACS, for identifying rearrangement breakpoints, and applied this pipeline to examine the genomic anatomy of *TMPRSS2* and *ERG* rearrangements in prostate cancer. We resolved *TMPRSS2-ERG* rearrangements at nucleotide resolution in 25 of 83 primary prostate cancer samples (30%), a prevalence similar to that previously reported (30–70%) [2]. Further analyses revealed several interesting features of *TMPRSS2-ERG* rearrangements.

We found a striking cluster of breakpoints within introns 1 and 2 of *TMPRSS2* and in a 30 kbp region in intron 3 of *ERG*, suggesting that rearrangements involving *TMPRSS2* and *ERG* arise non-randomly. Recent reports have shown that introns 1 and 2 of *TMPRSS2* and intron 3 of *ERG* are "hotspots" for androgen-induced double strand breaks (DSB) mediated by enzymes such as TOP2B and/or other nucleases [6, 29], suggesting that androgen receptor signaling may be integrally involved in generating *TMPRSS2* and *ERG* rearrangements [4].

We also observed that several *TMPRSS2-ERG* rearrangements occurred as part of complex genomic alterations involving multiple intra- and inter-chromosomal rearrangements. This finding is consistent with a recent report of whole genome sequencing of a small number of prostate cancer genomes showing that genomic rearrangements, including those involving *TMPRSS2* and *ERG*, often involved the joining of several discontiguous genomic segments in a "daisy-chaining" architecture [1]. We can speculate that such a complex rearrangement architecture may arise from the co-localization of several genes in the 3-dimensional space of the nucleus [10, 30–33] followed by generation of DSB at two or more of these loci through genotoxic stress (e.g. replication errors, telomere dysfunction or exogenous DSB inducing stress) or hormone-triggered transcription induced DSB [4]. Importantly, transcription factors, in particular the androgen receptor, have been shown to induce proximity between numerous gene loci in *cis* and *trans* [4, 29, 31, 34]. Concomitantly, initiation of transcription can induce transient DSB, likely mediated by TOP2B [6, 35–37]. Importantly, TOP2B mediated androgen induced DSB can be detected near sites of genomic breakpoints in *TMPRSS2* and *ERG*. This suggests a potential transcription dependent mechanism for the simultaneous generation of multiple, co-localized DSB. Illegitimate repair of breaks could then lead to rearrangement of several genomic segments in a single event to produce the observed complex genomic rearrangements.

Identification and resolution of a large number of *TMPRSS2* and *ERG* rearrangements at the nucleotide level has yielded important hypotheses regarding the DNA repair mechanisms involved in generating such prostate cancer-specific rearrangements. A significant fraction of the rearrangements observed at *TMPRSS2* and *ERG* appeared to occur at or near regions of microhomology, and often involved insertions at the breakpoint junction. This observation suggests that NHEJ/MMEJ pathways, which can be seeded by regions of microhomology and can be error-prone, might be involved in "stitching together" *TMPRSS2-ERG* fusions in prostate cancer. Indeed, components of the NHEJ DSB repair pathway were required for *de novo* generation of *TMPRSS2-ERG* fusions *in vitro* [6, 29]. Several aspects of the data from our current study also provide evidence against the involvement of homologous recombination in the generation of prostate cancer recurrent rearrangements. None of the *TMPRSS2-ERG* rearrangements we observed involved regions of long homology, suggesting that non-allelic homologous recombination (NAHR) mechanisms, which are thought to require the presence of >200 bp of homology [38, 39], are unlikely to play a major role in such rearrangements.

Our work has implications for personalized DNA-based biomarker development. DNA-based biomarkers may offer an advantage over their RNA- and protein-based counterparts in certain situations. For example, androgen deprivation therapy may directly lead to reductions in serum PSA and *TMPRSS2-ERG* transcript levels independent of reductions in the tumor burden, and would therefore be suboptimal for monitoring true treatment response. On the other hand, detection of personalized *TMPRSS2-ERG* genomic rearrangement breakpoints in cell-free DNA or circulating-tumor-cells could reflect the true tumor burden without being confounded by the abrogation of androgen signaling. Generally speaking, circulating biomarkers of all three types (DNA, RNA, and protein) will invariably be affected by pharmaceutical, radiological or surgical interventions; however, DNA biomarkers could, in principle, maintain a stoichiometric relationship to tumor cell number despite intervention-induced molecular signaling alterations that reduce transcript/protein levels independent of changes in tumor mass. Two proof-of-principal studies recentely demonstrated the feasibility of using DNA-based biomarkers (including mutations and structural rearrangements) to monitor tumor dynamics [12, 40]. Another benefit of DNA-based biomarkers is their potential utility for following clonal and subclonal evolution of the tumor burden [41–44].

In practice, a two-tier strategy could be deployed. First, with further technical optimization to allow analysis of even scant input DNA derived from clinical specimens, such as FFPE blocks/biopsies or bodily fluids, the geBACS approach could be used to identify cancer-specific recurrent rearrangement breakpoints from an individual's tumor in a costeffective manner. Indeed, recent reports have demonstrated that libraries for capture sequencing can be generated from very limited sample material, FFPE tissues and even directly from circulating cell-free DNA from peripheral blood. [45–47] Second, highly quantitative PCR assays can then be designed that would allow the detection of these personalized tumor-specific rearrangement junctions in cell-free DNA or in shed tumor cells in blood or other biospecimens (urine, sputum, etc) to follow disease burden longitudinally. The general feasibility of such an approach was recently demonstrated by Leary et al. [12, 47] highlighting the potential of genomic rearrangement-based DNA biomarkers in monitoring treatment response. For prostate cancer, such a genomic rearrangement based DNA biomarker could be very useful for monitoring treatment response following radiation therapy for primary cancer or systemic therapies for advanced cancers, where serum PSA-based monitoring has faced limitations [48–51]. Further refinements in the geBACS approach would allow highly automated and sensitive identification of *TMPRSS2-ERG* and other recurrent rearrangements for use in such a DNA-based biomarker strategy. Further study will be needed to evaluate whether the biomarker strategy outlined above can stoichiometrically reflect true tumor burden and/or augment existing monitoring tools for following prostate cancer treatment response.

DATA ACCESS

The sequence data from this study have been submitted to the NCBI Sequence Read Archive database (http://trace.ncbi.nlm.nih.gov/Traces/sra/) under accession no. SRX143861.3.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to acknowledge the Next Generation Sequencing Center at the Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins for bioinformatics support; The Johns Hopkins Brady Urological Research Institute Prostate Specimen Repository supported in part by the NCI/Prostate SPORE P50CA58236 for

help with providing prostate tissue specimens; Marcella Sutherland and Bonnie Gambichler from the Johns Hopkins TMA core facility for help with tissue processing; and David Moore from the Johns Hopkins University High-throughput sequencing core facility for library sequencing. We also thank Dr. Berrak Gümüşkaya, Dr. Martin Aryee, and Dr. Alan Meeker for insightful comments. Sources of funding: National Institutes of Health/National Cancer Institute grants P50CA058236, R01CA070196; Department of Defense Congressionally Directed Medical Reseach Program's Prostate Cancer Research Program grant W81XWH-08-1-0049; The Protate Cancer Foundation research awards (to S.Y. and W.G.N.); The Prostate Cancer Foundation Young Investigator Award (to M.C.H.); The V Foundation for Cancer Research Martin D. Abeloff V Scholar Award (to S.Y.); The Helen B. Masenhimer Fellowship (to S.Y.); Generous philanthropic support from Mr. David H. Koch and the Irving A. Hansen Memorial Foundation.

ABBREVIATIONS

DSB	Double Strand Breaks
geBACS	genomic Breakpoint Analysis by Capture Sequencing
FISH	Fluorescence In-Situ Hybridization
NHEJ	Non-Homologous End Joining
MMEJ	Microhomology Mediated End Joining
Array CGH	Array Comparative Genomic Hybridization
FFPE	Fresh Frozen Paraffin Embedded
P1,P2	refers to the individual reads of a read pair from paired end sequencing data
BLAT	BLAST Like Alignment Tool
NAHR	Non-Allelic Homologous Recombination

REFERENCES

1. Berger MF, Lawrence MS, Demichelis F, et al. The genomic complexity of primary human prostate cancer. Nature. 2011; 470:214–220. [PubMed: 21307934]
2. Rubin MA, Maher CA, Chinnaiyan AM. Common gene rearrangements in prostate cancer. J Clin Oncol. 2011; 29:3659–3668. [PubMed: 21859993]
3. 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. [PubMed: 16254181]
4. Haffner MC, De Marzo AM, Meeker AK, et al. Transcription-induced DNA double strand breaks: both oncogenic force and potential therapeutic target? Clin Cancer Res. 2011; 17:3858–3864. [PubMed: 21385925]
5. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nat Rev Cancer. 2008; 8:497–511. [PubMed: 18563191]
6. Haffner MC, Aryee MJ, Toubaji A, et al. Androgen-induced TOP2B–mediated double-strand breaks and prostate cancer gene rearrangements. Nat Genet. 2010; 42:668–675. [PubMed: 20601956]
7. Demichelis F, Setlur SR, Beroukhim R, et al. Distinct genomic aberrations associated with ERG rearranged prostate cancer. Genes Chromosomes Cancer. 2009; 48:366–380. [PubMed: 19156837]
8. Esgueva R, Perner S, C JL, et al. Prevalence of TMPRSS2-ERG and SLC45A3-ERG gene fusions in a large prostatectomy cohort. Mod Pathol. 2010; 23:539–546. [PubMed: 20118910]
9. Liu W, Laitinen S, Khan S, et al. Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. Nat Med. 2009; 15:559–565. [PubMed: 19363497]
10. Mani RS, Chinnaiyan AM. Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. Nat Rev Genet. 2010; 11:819–829. [PubMed: 21045868]
11. Scheble VJ, Braun M, Beroukhim R, et al. ERG rearrangement is specific to prostate cancer and does not occur in any other common tumor. Mod Pathol. 2010; 23:1061–1067. [PubMed: 20473283]

12. Leary RJ, Kinde I, Diehl F, et al. Development of personalized tumor biomarkers using massively parallel sequencing. Sci Transl Med. 2010; 2:20ra14.
13. He J, Wu J, Jiao Y, et al. IgH gene rearrangements as plasma biomarkers in Non-Hodgkin's lymphoma patients. Oncotarget. 2011; 2:178–185. [PubMed: 21399237]
14. Conrad DF, Bird C, Blackburne B, et al. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nat Genet. 2010; 42:385–391. [PubMed: 20364136]
15. Sobreira NL, Gnanakkan V, Walsh M, et al. Characterization of complex chromosomal rearrangements by targeted capture and next-generation sequencing. Genome Res. 2011; 21:1720–1727. [PubMed: 21890680]
16. Yegnasubramanian S, Kowalski J, Gonzalgo ML, et al. Hypermethylation of CpG islands in primary and metastatic human prostate cancer. Cancer Res. 2004; 64:1975–1986. [PubMed: 15026333]
17. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357–359. [PubMed: 22388286]
18. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002; 12:656–664. [PubMed: 11932250]
19. Lotan TL, Gupta NS, Wang W, et al. ERG gene rearrangements are common in prostatic small cell carcinomas. Mod Pathol. 2011; 24:820–828. [PubMed: 21336263]
20. Qu X, Randhawa G, Friedman C, et al. A novel four-color fluorescence in situ hybridization assay for the detection of TMPRSS2 and ERG rearrangements in prostate cancer. Cancer Genet. 2013
21. Chaux A, Albadine R, Toubaji A, et al. Immunohistochemistry for ERG expression as a surrogate for TMPRSS2-ERG fusion detection in prostatic adenocarcinomas. Am J Surg Pathol. 2011; 35:1014–1020. [PubMed: 21677539]
22. Furusato B, Tan SH, Young D, et al. ERG oncoprotein expression in prostate cancer: clonal progression of ERG-positive tumor cells and potential for ERG-based stratification. Prostate Cancer Prostatic Dis. 2010; 13:228–237. [PubMed: 20585344]
23. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009; 27:182–189. [PubMed: 19182786]
24. Park K, Tomlins SA, Mudaliar KM, et al. Antibody-based detection of ERG rearrangement-positive prostate cancer. Neoplasia. 2010; 12:590–598. [PubMed: 20651988]
25. Tomlins SA, Mehra R, Rhodes DR, et al. TMPRSS2:ETV4 gene fusions define a third molecular subtype of prostate cancer. Cancer Res. 2006; 66:3396–3400. [PubMed: 16585160]
26. 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. [PubMed: 18172298]
27. Nambiar M, Raghavan SC. How does DNA break during chromosomal translocations? Nucleic Acids Res. 2011; 39:5813–5825. [PubMed: 21498543]
28. McVey M, Lee SE. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. 2008; 24:529–538. [PubMed: 18809224]
29. Lin C, Yang L, Tanasa B, et al. Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. Cell. 2009; 139:1069–1083. [PubMed: 19962179]
30. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289–293. [PubMed: 19815776]
31. Mani RS, Tomlins SA, Callahan K, et al. Induced chromosomal proximity and gene fusions in prostate cancer. Science. 2009; 326:1230. [PubMed: 19933109]
32. Zhang Y, McCord RP, Ho YJ, et al. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell. 2012; 148:908–921. [PubMed: 22341456]
33. Chakalova L, Fraser P. Organization of transcription. Cold Spring Harb Perspect Biol. 2010; 2:a000729. [PubMed: 20668006]
34. Rickman DS, Soong TD, Moss B, et al. Oncogene-mediated alterations in chromatin conformation. Proc Natl Acad Sci U S A. 2012; 109:9083–9088. [PubMed: 22615383]

35. Ju BG, Lunyak VV, Perissi V, et al. A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription. Science. 2006; 312:1798–1802. [PubMed: 16794079]
36. Wong RH, Chang I, Hudak CS, et al. A role of DNA-PK for the metabolic gene regulation in response to insulin. Cell. 2009; 136:1056–1072. [PubMed: 19303849]
37. Brenner JC, Ateeq B, Li Y, et al. Mechanistic rationale for inhibition of poly(ADPribose) polymerase in ETS gene fusion-positive prostate cancer. Cancer Cell. 2011; 19:664–678. [PubMed: 21575865]
38. Rubnitz J, Subramani S. The minimum amount of homology required for homologous recombination in mammalian cells. Mol Cell Biol. 1984; 4:2253–2258. [PubMed: 6096689]
39. Reiter LT, Hastings PJ, Nelis E, et al. Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. Am J Hum Genet. 1998; 62:1023–1033. [PubMed: 9545397]
40. Diehl F, Schmidt K, Choti MA, et al. Circulating mutant DNA to assess tumor dynamics. Nat Med. 2008; 14:985–990. [PubMed: 18670422]
41. Sowalsky A, Ye H, Bubley GJ, et al. Clonal Progression of Prostate Cancers from Gleason Grade 3 to Grade 4. Cancer Res. 2012
42. Misale S, Yaeger R, Hobor S, et al. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. Nature. 2012; 486:532–536. [PubMed: 22722830]
43. Diaz LA Jr. Williams RT, Wu J, et al. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. Nature. 2012; 486:537–540. [PubMed: 22722843]
44. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012; 366:883–892. [PubMed: 22397650]
45. Beltran H, Yelensky R, Frampton GM, et al. Targeted Next-generation Sequencing of Advanced Prostate Cancer Identifies Potential Therapeutic Targets and Disease Heterogeneity. Eur Urol. 2012
46. Wagle N, Berger MF, Davis MJ, et al. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer Discov. 2012; 2:82–93. [PubMed: 22585170]
47. Leary RJ, Sausen M, Kinde I, et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Sci Transl Med. 2012; 4 162ra154.
48. Scher HI, Morris MJ, Basch E, et al. End points and outcomes in castration-resistant prostate cancer: from clinical trials to clinical practice. J Clin Oncol. 2011; 29:3695–3704. [PubMed: 21859988]
49. Attard G, de Bono JS. Prostate cancer: PSA as an intermediate end point in clinical trials. Nat Rev Urol. 2009; 6:473–475. [PubMed: 19727145]
50. Caloglu M, Ciezki J. Prostate-specific antigen bounce after prostate brachytherapy: review of a confusing phenomenon. Urology. 2009; 74:1183–1190. [PubMed: 19428077]
51. Roach M 3rd, Hanks G, Thames H Jr. et al. Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: recommendations of the RTOG-ASTRO Phoenix Consensus Conference. Int J Radiat Oncol Biol Phys. 2006; 65:965–974. [PubMed: 16798415]

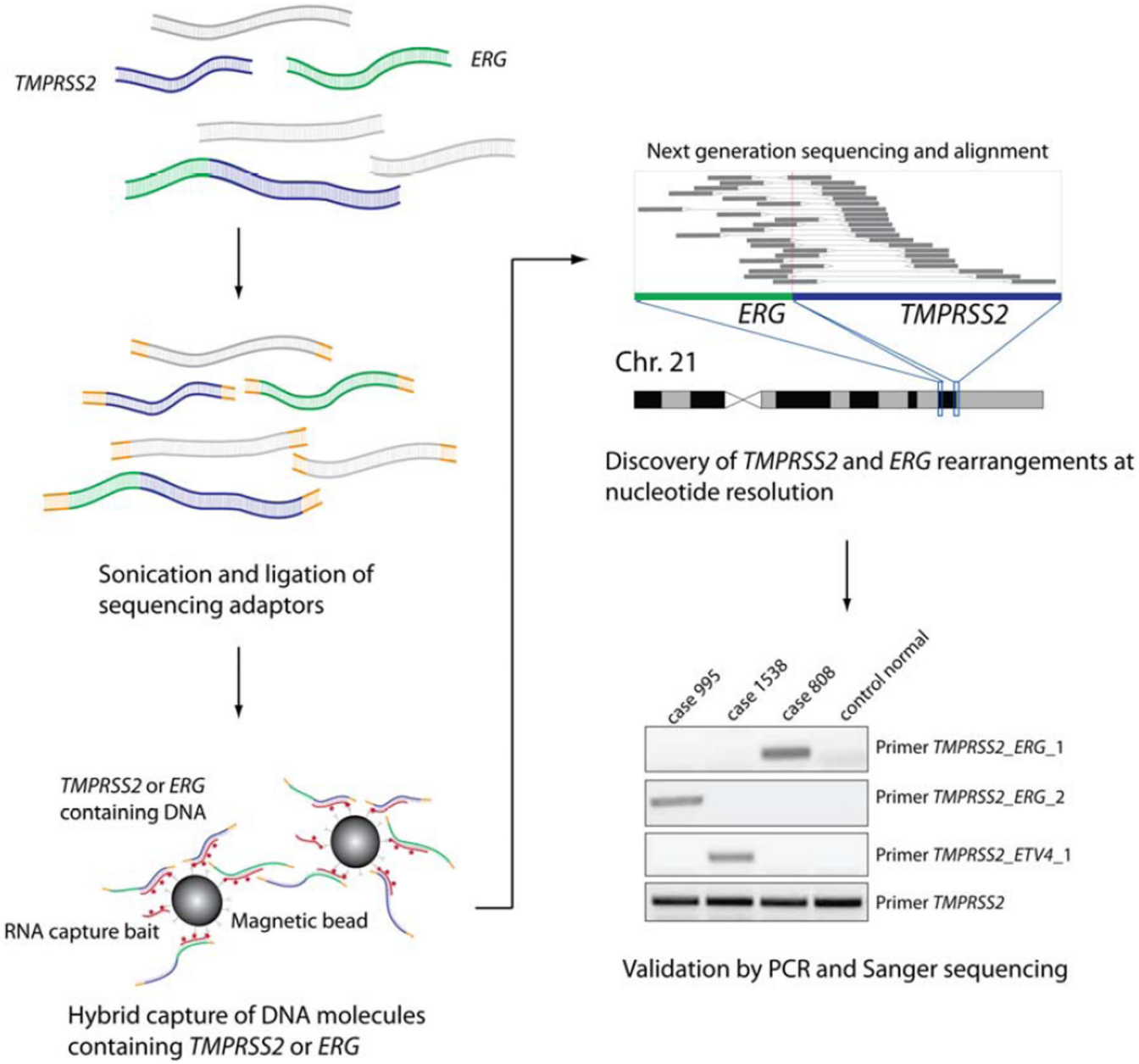

FIGURE 1. Overview of genomic breakpoint analysis by capture sequencing (geBACS)

Pooled samples of subject, reference, and cell line genomic DNA were fragmented by sonication and prepared for paired-end sequencing on the Illumina Genome Analyzer IIx platform. Pools were then enriched for *TMPRSS2* or *ERG* containing fragments through hybridization to a custom designed RNA capture library. Captured fragments were isolated, amplified, and subjected to paired-end next-generation sequencing. Sequenced read-pairs were aligned to the reference genome and reads flanking or spanning breakpoints in *TMPRSS2 ERG* and other fusion partners were determined. Nominated rearrangement breakpoints were then confirmed by PCR and Sanger sequencing.

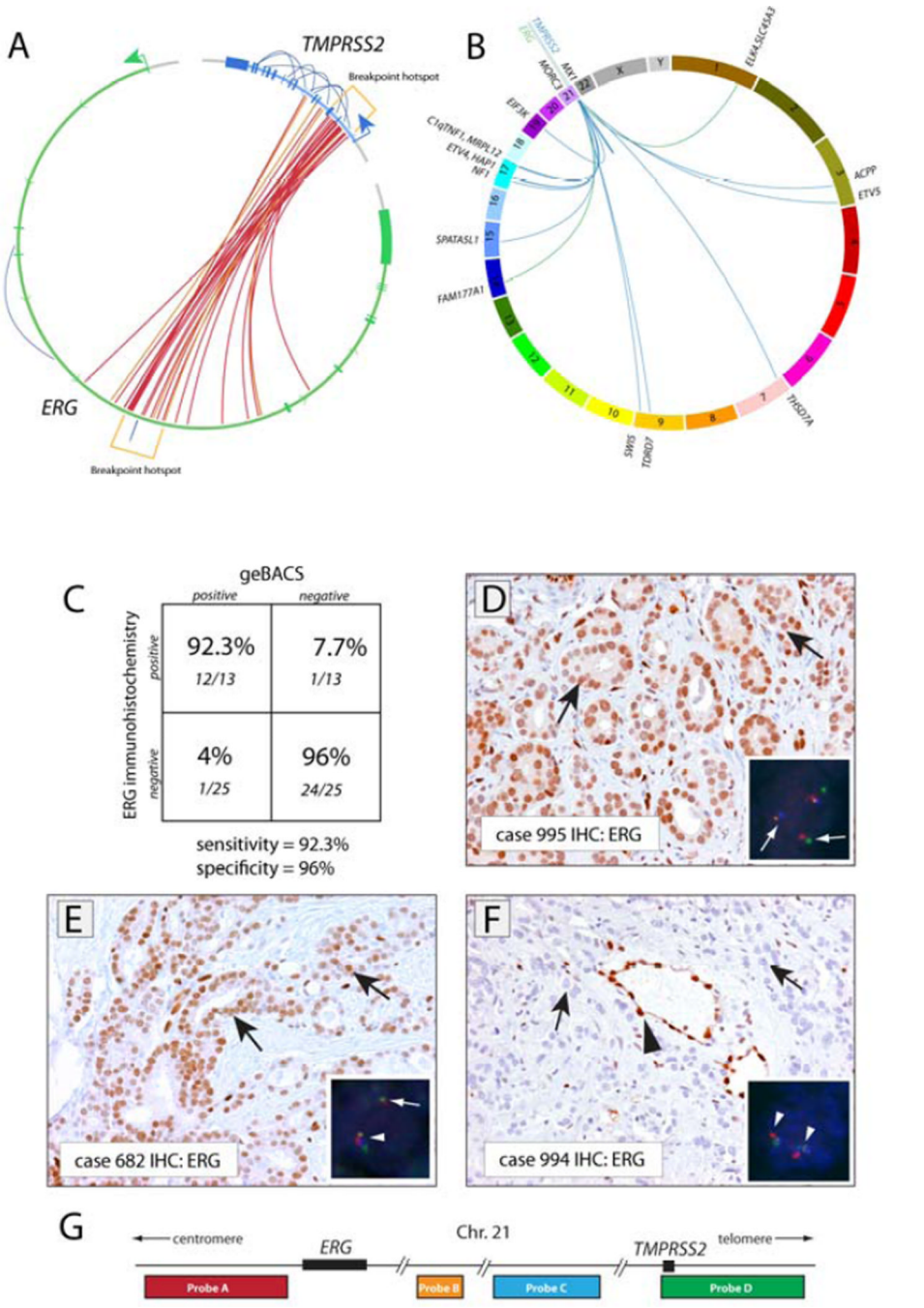

FIGURE 2. Genomic architecture and *in situ* confirmation of rearrangements involving *TMPRSS2* and/or *ERG*

(A) Circos plot depicting 26 rearrangements between *TMPRSS2* and *ERG* identified in 25 out of 83 prostate cancer tissues analyzed. Each inner line represents a *TMPRSS2-ERG* rearrangement. Outer lines represent intragenic rearrangements within the *TMPRSS2* or *ERG* loci. Red links represent rearrangements identified with standard orientation (5' sense fragment of *TMPRSS2* fused to a 3' sense fragment of *ERG*). Orange links represent nonstandard rearrangements as noted in the text. Rearrangement hotspots are noted in brackets. (B) Circos plot showing inter-chromosomal rearrangements involving *TMPRSS2* or *ERG* with other genes. The gene name of the overlapping or closest fusion partner gene is

indicated. Blue and green lines indicate rearrangements involving *TMPRSS2* or *ERG*, respectively. (C) Specificity and sensitivity of the geBACS pipeline in identifying *TMPRSS2-ERG* rearrangements, determined by using ERG IHC. (D) Strong nuclear staining for *ERG* in tumor cells in a representative prostate cancer case with FISH-confirmed *TMPRSS2-ERG* rearrangement by translocation, for which the *TMPRSS2-ERG* rearrangement junction was identified by geBACS. Insert shows four-color FISH for *TMPRSS2* and *ERG*. In inserts, white arrows indicate rearranged alleles, arrowheads normal alleles. (E) Positive ERG staining in a case with FISH-confirmed *TMPRSS2-ERG* rearrangement associated with deletion (see insert) for which the *TMPRSS2-ERG* rearrangement junction was identified by geBACS. (F) Absence of ERG staining in a case that did not show *TMPRSS2-ERG* rearrangement by FISH or geBACS. (G) Scheme of FISH probe localization on chromosome 21. (D–F) Black arrows indicate representative cancer cells; Black arrowhead indicates positive ERG staining in normal endothelial cells.

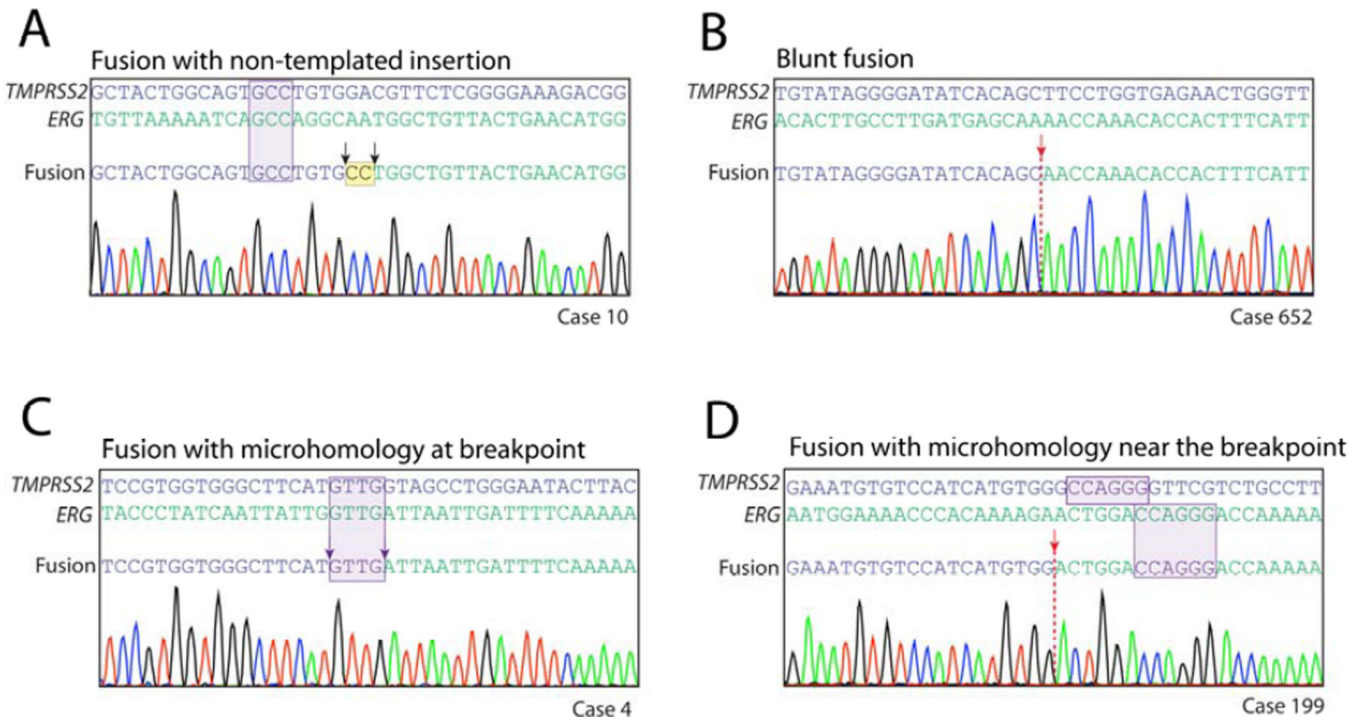

FIGURE 3. Nucleotide architecture of representative *TMPRSS2-ERG* genomic rearrangements

(A) Rearrangement sequence from a case showing non-templated insertions occurring at the junction in addition to flanking microhomology near the junction. (B) Rearrangement sequence from a case showing blunt fusion of rearranged genes. (C) Rearrangement sequence showing microhomology at the junction. Microhomology at a junction was defined as one or more nucleotides that could be assigned to either gene in the rearrangement. (B, D) Rearrangement sequences from cases 10 and 199 showed flanking microhomology adjacent to an insertion-associated and blunt junction respectively. Flanking microhomology was defined as three or more nucleotides of microhomology within 20 bp of the junction. This could occur in junctions that showed blunt fusion, fusion with inserted nucleotides, or fusions with microhomology at the junction.

Table 1

Characteristics of rearrangement junctions identified in patient, reference, and cell line samples
Patient Junctions

$A_{\text{Patient Bank \#}}$	$B_{\text{Junctions}}$	$C_{\text{Gene 1}}$	$D_{\text{Gene 2}}$	$E_{\text{Orientation}}$	F_{Intron}	$G_{\text{Reading Frame}}$	H_{FISH}	$I_{\text{ERG Stain}}$	J_{Reads}	$K_{\text{Junction Reads}}$	$L_{\text{Sequence Flanking Breakpoint}}$
4	*TMPRSS2-ERG*	chr21:41794326	chr21:38753597	s/s	1–3	IF	D	+	2	1	CGTGGTGGGCTTCATGTTGATTAATTGATTTTCAAAAATTATCTGTGGCT
10	*TMPRSS2-ERG*	chr21:41792775	chr21:38792757	s/s	1–3	IF	n/a	n/a	6	1	GGGTGGCTACTGGCAGTGCCTGTGccTGGCTGTTACTGAACATGGCCATC
70	*TMPRSS2-MPRL12*	chr21:41783383	chr17:77283404	s/a		ND	n/a	n/a	6	3	CCTCCTTCTTCCAGAGACTAAGCAATCCAGTTTACACAACAGGCTCTACAGT
	HAP1-TMPRSS2*	chr17:37144837	chr21:41779769	a/s		ND			19	5	GCCAGTGCATTAGGGCAGCGCTGACTGGTTTCCTGTTAAGCCCTCGC
	*TMPRSS2-HAP1**	chr21:41794230	chr17:37149531	s/s		ND			12	4	GAAGCCGGCCCCACTCTACTTCAGGATACCTAACACAGGTGAGCCCCACC
	TMPRSS2-SWI5	chr21:41779767	chr9:130085490	s/s		ND			5	2	CCCCTCTCCTAAGGGCCATCCTTGGTG AGGACGCTGAGGCAaGAGAATTA
82	*EIF3K–TMPRSS2*	chr19:43815389	chr21:41788584	a/s		ND	n/a	n/a	5	2	TTTACTGCTGCACGCCCAGTTTAGCTGTAGTACCTTTTTAATATCTTGAACCCG
	NF1-TMPRSS2	chr17:26590306	chr21:41774186	a/s		ND			13	4	GTCCATTTTCCATGAAAGCCCTTAGCATATCGGCACATGAGTGCTGGGGA
134	*TMPRSS2-ERG*	chr21:41795069	chr21:38803391	s/s	1–3	IF	n/a	n/a	8	2	CTCTGGTCTTCCATTCCAAAGTCCATGTTAGGGATGGGAGAGGAGGGTA
	ACPP-TMPRSS2	chr3:133565489	chr21:41795096	a/s		ND			6	2	CCCACTCTGTGTTTATTTTATTTTTACTCTGGTCTTCCATTCCAAAGTCCATGG
144	*TMPRSS2-ERG*	chr21:41779065	chr21:38796877	s/s	5–3	ND	n/a	n/a	7	5	TTATGTTACCTGAAGAGTTGCTCTCTATATTTTTACTTTGCCTTTTTATT
	TMPRSS2-TMPRSS2	chr21:41789960	chr21:41781176	s/s		ND			5	0	AAAGACTGGAGGAACTAAAAGTtcgagcaAAGATTTTGTGACCTCTGACC
199	*TMPRSS2-ERG*	chr21:41786586	chr21:38748712	s/a	3–3	ND	n/a	n/a	6	3	CGTGGAAATGTGTCCATCATGTGGACTGGACCAGGGACCAAAAAGGCTATAA
206	*TMPRSS2-ERG*	chr21:41793812	chr21:38763138	s/s	1–3	IF	n/a	n/a	9	0	TGGGAAATATTCTTCTGTGGtTAGACTACTCATGTCTTGTAAGAGGCGTT
214	*TMPRSS2-ERG*	chr21:41797467	chr21:38749676	s/s	1–3	IF	n/a	n/a	3	1	TTCGGTACATTCTAAGGTAGCTCCAGTGGGGAGGGGGGGGACCGTACCAATTA
341	*TMPRSS2-ERG*	chr21:41785537	chr21:38798302	s/s	3–3	ND	n/a	n/a	4	1	ACTCTCCTGGTGAGAACTGTACAGATTGTTTAAAAGTTATTTTAAGCTAA
353	*TMPRSS2-ERG*	chr21:41789358	chr21:38799621	s/s	2–3	IF	n/a	n/a	1	1	AGTGACGTGATCTCCATTCACTGCAACACGACTCATGTCTGGTTTCTAGCTGGCA
357	*TMPRSS2-ERG*	chr21:41794131	chr21:38797829	s/s	1–3	IF	n/a	n/a	4	2	TCTTTTCCTTCAGCAGGTTGTTAAATACACACCCTCATACACACACTCACACACTCA

$A_{\text{Patient Bank \#}}$	$B_{\text{Junctions}}$	$C_{\text{Gene 1}}$	$D_{\text{Gene 2}}$	$E_{\text{Orientation}}$	F_{Intron}	$G_{\text{Reading Frame}}$	H_{FISH}	$I_{\text{ERG Stain}}$	J_{Reads}	$K_{\text{Junction Reads}}$	$L_{\text{Sequence Flanking Breakpoint}}$
372	*TMPRSS2-ERG*	chr21:41797622	chr21:38784847	s/s	1–3	IF	−	−	1	0	AGTTCACTGGTGATGATGCATGGTTTTGCATTGCGTAGTCCTCTTTGG
	S LC 45A 3-ERG	chr1:203908669	chr21:38798718	s/s		IF			1	1	TGAACATGAACCCTTCCTGAATGTCATGTATAATTCATAAAATAGTGACAGAT
535	*TMPRSS2-ERG*	chr21:41790462	chr21:38748584	s/s	2–3	IF	D	+	1	1	AGATGTGGGCTGGTGGGGCCCAAGTGATAATTTGCACACAAGAGACTGTG
558	*TDRD7**-*TMPRSS2*	chr9:99213679	chr21:41794880	s/s		ND	n/a	n/a	2	0	GTAAAATACATATAGTGTGCTAGATATGAGTATTGATTCTTCACTCCTTT
	TMPRSS2-TMPRSS2	chr21:41771407	chr21:41780399	s/s		ND			7	1	TACAATAATAGGTCCTTCCTGACACCT/AATAAGTTTTAAAGGAAGAGGAA
	MX1-TMPRSS2	chr21:41752881	chr21:41775109	s/s		ND			6	3	CTAGAAACTGACACATGCTGAACATaaagGAGCATAAGGTCTTCAGCACT
580	*MX1-TMPRSS2*	chr21:41752819	chr21:41775553	s/s		ND	n/a	n/a	2	1	AAGGGATTTTCAGCCCTCAGACTTATTGAGATCTAATTTATGTACCATAA
652	*TMPRSS2-ERG*	chr21:41792424	chr21:38775810	s/s	1–3	IF	T	+	4	1	TGTTCTGTATAGGGGATATCACAGCAACCAAACACCACTTTCATTTTTAT
	ERG-ERG	chr21:38830023	chr21:38872708	s/s		ND			2	1	TCTCCAACCAAACCTGACTTATACTTAG/CCAGTGATAAAAGGAGACTCAAAC
675	*TMPRSS2-ERG*	chr21:41798800	chr21:38789416	s/s	1–3	IF	D	+	13	0	TGAAGGCCCAGGTGCATTTCTGCTCTCTGGTTCCCCGGACCTTTTAAGGAG
682	*TMPRSS2-ERG*	chr21:41796887	chr21:38791968	s/s	1–3	IF	D	+	6	1	GCAAATTGCATTGTATTTCAAATTTtAAACTTCAAACTCAGAAAAAAATT
733	*TMPRSS2-ERG*	chr21:41794661	chr21:38816717	s/s	1–3	IF	n/a	n/a	7	3	TCTGTCCTGCTATGAGACAAGAATGCGGACTTTGTTTTTGTTATTCTGTCTCCA
738	*THSD7A–TMPRSS2*	chr7:11743984	chr21:41788390	s/s		ND	n/a	−	4	2	ATTAGTACAATATTTTTGAAAATTTTCAACTGTTTAGGGGTCACCACCAG
780	*TMPRSS2-ERG*	chr21:41792243	chr21:38787029	s/s	1–3	IF	n/a	n/a	1	0	CCTGGCCGCTGCACTTACAATTGCACaGCTGCTTGCGATGTTTCACTCAT
808	*TMPRSS2-ERG*	chr21:41790468	chr21:38780232	s/s	2–3	IF	T	+	38	13	AGAAGGGGAAGATGTGGGCTGGTGGGGCCCTGCAAACATCACAAAGAGCACT
	TMPRSS2-ERG	chr21:41783369	chr21:38805184	a/a	Exon4-Intron3	ND			10	5	GAGGAAGGTCCCCAGGGTCAAGGTGAGATGTTTAATACCTACAAATACAG
	*SPATA5L1**-*TMPRSS2*	chr15:43481074	chr21:41771055	a/s		ND			9	1	TATCTACGTTATTAAGACAACTGCCTGACGCTCAGTGAAATAATTCAGGT
814	*TMPRSS2-ERG*	chr21:41792783	chr21:38784111	a/s	1–3	ND	T+		9	0	TCCCCGAGAACGTCCACAGGCACATTTTTCTTCTAAGACTAAAAT
	MORC3-TMPRSS2	chr21:36673099	chr21:41762272	s/a		ND			12	0	TGTAAGGACTGAGACAAATGTCAGCgtGCAGATATGTCTATGACAACCTG
816	*TMPRSS2-ERG*	chr21:41791718	chr21:38797360	s/s	2–3	IF	T	+	10	4	ACACAGCTGCCCAGGTGAGTCGCAAGAGCAGACTGAGATAGGCTTCCCGA
	C1qTNF1-TMPRSS2	chr17:74540999	chr21:41791286	s/s		ND			6	2	GTGGCTCGTAGCCACTGTACTGACTCAGGAATTTTCAGGGACAAACCTGC

$A_{\text{Patient Bank \#}}$	$B_{\text{Junctions}}$	$C_{\text{Gene 1}}$	$D_{\text{Gene 2}}$	$E_{\text{Orientation}}$	F_{Intron}	$G_{\text{Reading Frame}}$	H_{FISH}	$I_{\text{ERG Stain}}$	J_{Reads}	$K_{\text{Junction Reads}}$	$L_{\text{Sequence Flanking Breakpoint}}$
	C1qTNF1-TMPRSS2	chr17:74540985	chr21:41764084	a/s		ND			9	2	AAATGTTTAAGTCAGTACAGTGGCcTGAGGACGGGCATCCTGGACCCATG
	TMPRSS2-TMPRSS2	chr21:41764044	chr21:41796946	s/a		ND			13	3	TGGACCCATGGTGGCCACATCTAAGC/AAATGAGTAAGATGAAATTTAGCA
981	*TMPRSS2-ERG*	chr21:41795091	chr21:38796469	s/s	1–3	IF	D	+	15	3	GAATGTTACTGGCAGCATCACTCTGAAACAGGCCCTTTGGAGAGGGG
989	*FAM177A1-ERG*	chr14:34584342	chr21:38756847	a/s		ND	D	+	5	1	GCCAAAGTTCTTCTCCTTACCCGGACACACTGGGAGATAGGCAGGGAGTC
995	*TMPRSS2-ERG*	chr21:41780434	chr21:38793491	a/a	5–3	ND	T+		5	0	TAAAGGAAGAGGAAGAAAAGTAAAAGGAGGAGAAGGAGAAAGATAAGGGG
	TMPRSS2-TMPRSS2	chr21:41762098	chr21:41773713	a/s		ND			12	3	CCGAGCTGGCTCCGTGTCACTGA/CAGTGATTCCTAACGGGGGTGGGGCGG
	TMPRSS2-TMPRSS2	chr21:41780464	chr21:41790293	s/a		ND			10	3	GTACGGAATGGGGCTCTGCAGATGGCAG/TGTGGCCATACATAAATGTGAA
	TMPRSS2-TMPRSS2	chr21:41786495	chr21:41773659	s/s		ND			2	1	GCGGTGGTTACCATTTACCAGCACTGCCT/TGCTTTTATAAGGGAAGAATATGA
	ERG-ERG	chr21:38793472	chr21:38793604	a/s		ND			5	0	TTCTGGCTTTTATTATATACCTCTGATTT/TT/CTGTCCTTCCT TTCTCCTTCTCCAT
1164	*TMPRSS2-TV5*	chr21:41797113	chr3:187278460	s/s	1–7	IF	T, TMPRSS2	+ for ETV5	11	1	GCCAGTGGCTTGTGTGTGGTGATTCCTCCTTGAAGTCTGCCTGTGCATTTTGGA
1273	*ELK4-RG*	chr1:203858916	chr21:38746745	a/s		ND	T,ERG	−	7	0	CAGAGTAAGACTCCGTCTCAAAAAAAACCAACAAGCATTTGAGGTTAAGAAATT
1355	*TMPRSS2-RG*	chr21:41788584	chr21:38799738	s/s	2–3	IF	n/a	+	1	0	TACAGGCATGAGCCACCGCACCCAGCCTTCAAGGAATTTAAATCAAGAATTAGG
1422	*TMPRSS2-RG*	chr21:41795553	chr21:38788443	a/s	1–3	IF	n/a	+	7	2	ACCAGGAAATCACATTCAGTCTCATAAAGTTTTTGTTTTGCAAAACAGAT
1538	*TMPRSS2-TV4*	chr21:41795742	chr17:38978855	s/s	1–3	IF	T,TMPRSS2	+ for ETV4	17	0	CTTTCACACTTAAAGGAAAATTTCCCCACACACACACACACACACACACGTT
1665	*TMPRSS2-RG*	chr21:41791298	chr21:38729732	s/s	2–4	IF	n/a	n/a	1	0	ATATTCATTTTTTACTGGGTCTGTGTTTAGGGCTCAATCCTTTCTTCTCTGAAAA
1863	*TMPRSS2-MPRSS2*	chr21:41761944	chr21:41777674	a/s		ND	n/a	n/a	5	0	TTATGGTACATAAATTAGATCTCAATAAG/TCTGAGGGCTGAAAATCCCTT
VCAP Cell Line Junctions											
VCAP	*TMPRSS2-RG*	chr21:41779893	chr21:38798223	a/s	5–3	ND	T+		28	4	CTCCAGGAGGTTAGGACTGCATACATTGACCTATTTGGAGTCCTTGATAA
	TMPRSS2-MPRSS2	chr21:41793823	chr21:41779387	s/a		ND			19	7	GCTTTGTGCCAGCAACTGGGAAATA/AGTGAGCCATGGTCATGCCTCTGCA
Reference Case Junctions											

A Patient Bank #	B Junctions	C Gene 1	D Gene 2	E Orientation	F Intron	G Reading Frame	H FISH	I ERG Stain	J Reads	K Junction Reads	L Sequence Flanking Breakpoint
Case 45	*TMPRSS2-ERG*	chr21:41800806	chr21:38776561	s/s	1–3	IF	n/a	+	18	5	CGAGGAATCCTGGTGGCTTGTTTGGGAGAGGCAGAGGTTGCAGTCAGTCGAG
Case 66	*TMPRSS2-ERG*	chr21:41795687	chr21:38804491	s/s	1–3	IF	n/a	+	24	11	TGAGAGCCATCATCCCGGTCCTTTAAGCCTCAACTTCCATGGGTCACTAC
Case 77	*TMPRSS2-ERG*	chr21:41799367	chr21:38790118	s/s	1–3	IF	n/a	+	21	12	TATGACATGGCCAGAGTGCCTGGAACTATTTCCAGTCTCAATGCACTCTGTTTTT

A Patient Bank # indicates the subject identification number used for each patient specimen.

B Genes involved in rearrangements with 5' rearrangement partner listed first.

C Genomic position of the rearrangement breakpoint for the 5' partner.

D Genomic position of the rearrangement breakpoint for the 3' partner.

E Refers to the strand orientation of Gene1/Gene2 in the rearrangement; s = sense and a = antisense.

F Indicates the introns involved in a TMPRSS2-ETS gene rearrangement.

G Reading frame denoted as being In-Frame (IF) or Non-determinable (ND)

H Rearrangement status by FISH; Rearrangement with Deletion (D) or with Translocation (T)

I ERG staining status by IHC; additional ETS rearrangement staining by CISH

J Number of paired-end reads flanking or directly overlapping the rearrangement junction.

K Number of reads directly overlapping the rearrangement junction.

L Confirmed nucleotide resolution of rearrangemnt junctions with relevant architecture;
TMPRSS2, ERG,

and
Partner Genes

are shown with
microhomologies

and
insertions

at or near the junction. Microhomologies were defined as being one or more ambiguously assignable nucleotides at the junction and three or more nucleotides in sequences flanking the junction that had identity in both partner sequences.

* Indicates nearest gene when rearrangement occurs in intergenic space