Exome sequencing identified six copy number variations as a prediction model for recurrence of primary prostate cancers with distinctive prognosis

Jie Liu\textsuperscript{1,2}, Jiajun Yan\textsuperscript{3}, Ruifang Mao\textsuperscript{2}, Guoping Ren\textsuperscript{4}, Xiaoyan Liu\textsuperscript{4}, Yanling Zhang\textsuperscript{4,5}, Jili Wang\textsuperscript{4}, Yan Wang\textsuperscript{4}, Meiling Li\textsuperscript{6}, Qingchong Qiu\textsuperscript{2}, Lin Wang\textsuperscript{2}, Guanfeng Liu\textsuperscript{2}, Shanshan Jin\textsuperscript{2}, Liang Ma\textsuperscript{2}, Yingying Ma\textsuperscript{2}, Na Zhao\textsuperscript{3}, Hongwei Zhang\textsuperscript{6}, Biaoyang Lin\textsuperscript{1,2,7,8}

\textsuperscript{1}College of Life Science, Zhejiang University, Hangzhou 310027, China; \textsuperscript{2}Systems Biology Division, Zhejiang-California International NanoSystems Institute (ZCNI), Zhejiang University, Hangzhou 310027, China; \textsuperscript{3}Department of Urology, Shaoxing People's Hospital, Shaoxing Hospital of Zhejiang University, Shaoxing 312000, China; \textsuperscript{4}Department of Pathology, The First Affiliated Hospital, Zhejiang University Medical College, Hangzhou 310001, China; \textsuperscript{5}Department of Gynecology and Obstetrics, Sir Run Run Shaw Hospital, Zhejiang University Medical College, Hangzhou 310016, China; \textsuperscript{6}Department of Epidemiology, Second Military Medical University, Shanghai 200433, China; \textsuperscript{7}Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou 310027, China; \textsuperscript{8}Department of Urology, University of Washington, Seattle, WA, USA

Contributions: (I) Conception and design: B Lin, H Zhang, J Yan; (II) Administrative support: N Zhao, Y Ma, L Ma; (III) Provision of study materials or patients: G Ren, X Liu, Y Zhang, J Wang, Y Wang, M Li; (IV) Collection and assembly of data: J Liu, R Mao, Q Qiu, L Wang, G Liu, S Jin; (V) Data analysis and interpretation: J Liu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Biaoyang Lin, PhD. Director, Systems Biology, Room 243, Zhejiang-California International NanoSystems Institute (ZCNI), Zhejiang University, 866 Yu Hang Tang Road, Zi Jin Gang Campus Zhejiang University, Hangzhou 310058, China. Email: biaoylin@gmail.com.

Background: Prostate cancer (PCa) is a common type of malignancy, which represents one of the leading causes of death among men worldwide. Copy number variations (CNVs) and gene fusions play important roles in PCa and may serve as markers for the prognosis of this condition.

Methods: We have presently conducted an analysis of CNVs and gene fusions in PCa, using whole exome sequencing (WES) data of primary tumors. For this, a cohort of 74 PCa patients, including 30 recurrent and 44 non-recurrent cases, were assessed during 5 years of follow-up.

Results: We have identified 66 CNVs that were specific to the primary tumor tissues from the recurrent PCa group. Most of duplicated genomic regions were located in 8q2, suggesting that this chromosomal region could be important for the prognosis of PCa. Meanwhile, we have developed a random forest model, using six selected CNVs, with an accuracy near 90% for predicting PCa recurrence according to a 10-fold cross validation. In addition, we have detected 16 recurrent oncogenic gene fusions in PCa. Among these, ALK (ALK receptor tyrosine kinase)-involved fusions were the most common type of gene fusion (n=7). Four of these fusions (i.e., EML4-ALK, STRN-ALK, CLTC-ALK, ETV6-ALK) were previously identified in other cancer types, while the remaining three gene fusions (FRYL-ALK, ABL1-ALK, and BCR-ALK) were here identified.

Conclusions: Our findings expand the current understanding in regard to prostate carcinogenesis. Current data might be further used for assay development as well as to predict PCa recurrence, using primary tissues.

Keywords: Cancer recurrence; copy number variation (CNV); gene fusion; random forest; whole exome sequencing (WES)
Introduction

Second-generation sequencing is an efficient technique to identify gross genomic changes in cancer, including gene fusion events and copy number variations (CNVs). In this context, whole genome sequencing (WGS) would be the ideal application but, due to the cost and the complexity of data analysis, its applicability in many clinical settings has reduced. Instead, exome sequencing is a more cost effective approach which has been largely used to identify single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels). Moreover, several studies have successfully utilized whole exome sequencing (WES) as a platform for the identification of CNVs (1), and a number of algorithms have been developed for this purpose, including ExomeCNV, VarScan2, exomeCopy, CoNIFER, ExomeDepth and XHMM (2). Lastly, exome sequencing can also be used to identify subsets of genomic rearrangements whose breakpoints are located in or near exon regions (3).

CNVs can play important roles in prostate carcinogenesis. Using an array comparative genomic hybridization (aCGH) method, Saramäki and colleagues have confirmed common gains and losses along the genome of prostate cancer (PCa) samples, including gains at 1q, 7, 8q, 16p and 17q, and losses at 2q, 4p/q, 6q, 8p, 13q, 16q, 17p and 18q, in addition to a novel recurrent gain at 9p13-q21 (4). CNVs have been also related to the prognosis of PCa. According to the current literature, a CNV-based model can correctly predict 73% of relapsed cases and 75% of the cases with short PSA doubling time (PSADT, <4 months) (5). In another WGS study, Camacho and colleagues detected 64 recurrent regions of genomic loss and gain, and confirmed that the burden of SCNs (somatic copy number alteration) was predictive for biochemical recurrence (BCR). At the same time, nine individual regions were also associated with relapse, including two deletions and seven gains (6).

Gene fusion events also play important roles in PCas. Indeed, the first gene fusion event occurring in solid tumors was discovered in PCa, which involved the identification of the fused TMPRSS2-ETS gene (7). Another example relates to TMPRSS2:ERG gene fusions, which are present in ~50% of PCa cases. TMPRSS2:ERG is capable of inducing TGF-β signaling and epithelial to mesenchymal transition (EMT) in human PCa cells (8). Moreover, the incidence of TMPRSS2:ERG gene fusions have been also related to PCa prognosis. As reported, Kulda and colleagues showed that a combination between high level of prostate specific antigen (PSA) in preoperative serum and increased expression of TMPRSS2-ERG in tumor tissues (P=0.0001) was the most promising marker of recurrence risk after radical prostatectomy (9). Consistently, Nam and colleagues showed that, in a 5-year period, patients with TMPRSS2:ERG fusion had a significantly higher risk of recurrence than patients without this gene fusion (58.4% versus 8.1%, respectively; P<0.0001) (10). Still, TMPRSS2-ERG fusion was not a frequent genomic alteration among PCa patients from Asian countries such as China and Turkey and, in this case, it appears to have a limited significance in the clinical practice (11,12).

In order to predict PCa recurrence after radical prostatectomy (RP), a variety of prediction models and biomarkers have been developed. For instance, Lalonde and colleagues developed a 100-locus genomic classifier (later refined to 31 functional loci) which could identify patients with elevated BCR rates (hazard ratio =2.73, P<0.0001) (13). Oh and colleagues selected 16 significant predictive SNPs of BCR, from a large-scale exome wide association study, and indicated that the built GRS (genetic risk score) had additional predictive gain for BCR after RP (14). Furthermore, Campbell and colleagues performed a systematic review of prediction tools and suggested different tools for different purpose (15). For example, they recommended Stephenson nomograms for BCR, the CAPRA nomogram for aggressive BCR and metastasis for preoperative prognosis. For postoperative prognosis, different prediction models were suggested (15). For example, they suggested the CAPRA-Surgery (CAPRA-S), Stephenson, Kattan, Duke prostate cancer (DPC), and the Suardi nomograms for the prediction of BCR, the DPC nomogram for aggressive BCR (15).

In the present study, we have conducted the WES of primary PCa samples with known follow-up outcomes (defined as recurrent or non-recurrent) after 5 or more years of clinical follow-up. Specifically, we sequenced the exome of 30 recurrent and 44 non-recurrent PCa patients and then analyzed the presence of CNVs and fusion genes specific to recurrent/non-recurrent PCa or common to both outcomes. These data were then used to develop a prediction model for PCa recurrence. Clinical assays using genes associated with this model might be further utilized for prediction of recurrence of PCa.

Methods

Patients and study design

This study has been conducted in accordance with the ethical standards, stated by the Declaration of Helsinki.
as well as by National and International guidelines. This work was approved by the authors’ institutional review board. Several thousands of PCa patients who underwent radical prostatectomy, between 2003 and 2010, at the First Affiliated Hospital of Zhejiang University were screened and classified as recurrent or non-recurrent PCa patients. Non-recurrent PCa patients were defined as those with Gleason score lower than 7 at diagnosis, and with no BCR within 5 years of follow-up. Recurrent PCa patients were those with Gleason score lower than 7 at diagnosis but the presence of BCR within 5 years of follow-up. Paraffin-fixed tissues sections were retrieved from the pathological archives. Tumor contents were sectioned and then analyzed by IHC (immunohistochemistry) staining. Tissue blocks with high tumor content were eventually used for DNA extraction. A total of 44 non-recurrent PCa samples and 30 recurrent PCa samples were involved in this study.

DNA library preparation and sequencing

QIAamp DNA FFPE Tissue Kit (Qiagen, USA) was used to extract DNA from FFPE (formalin-fixed, paraffin-embedded) tissue sections. The quality and concentration of the extracted DNA were evaluated using Nanodrop 2000 spectrophotometer and Invitrogen Qubit 2.0 fluorometer. Thereafter, a sequencing system covering about 180,000 exons of the human genome (SureSelect Human All Exon V5 & V5 + UTRs, Agilent) was used to capture the respective exomes, using 2 μg DNA per sample. A 2x76 bp paired-end sequencing was conducted using the Illumina HiSeq2500 machine.

CNV calling and annotation

Figure 1 presents a flowchart for the analysis of CNVs and gene fusion events originated from WES data. The XHMM (eXome-Hidden Markov Model) software suite (16) was used to annotate CNVs in the exome sequencing data. At first, the “best practice” pipeline (i.e., BWA + Picard + GATK) was used to generate processed alignment files. After running the cnv calling pipeline, the result file (xvcn) was analyzed using the bedtools (17), which computed both the depth and breadth of the coverage of CNV events along the whole genome. After that, duplicated (copy number gain) or deleted (copy number loss) regions (>10 kb), specific to the recurrent group, were filtered accordingly. For this, respective duplicated and deleted regions should be solely present in at least 10% samples of the recurrent group (i.e., absent in the non-recurrent group). In addition, these CNVs should be consistently present in all samples from the recurrent group. The Annovar software tool was further used to annotate these CNV events with the refgene and cytoband databases.

Building of random forest model with CNV events

Using the output file (vcf) of XHMM, the genotypes (DIP,
DEL, DUP) related to the CNV events in every sample were picked. Thereafter, the frequency of three genotypes per CNV event in recurrent group and/or non-recurrent group were counted. The chi-square test was used to evaluate the correlation between CNV events and PCa recurrence (CNV events with P value <0.1 were selected). After that, CNV events which could not be genotyped in more than 10% of PCa samples were removed. The remaining CNV events were further used to build a random forest model, with missing genotypes replaced by the most common genotype (DIP, Diploidy). Meanwhile, the importance of predictors (CNV events) was assessed. Lastly, the function tool rfcv of the randomForest package was used to show the cross-validated prediction performance of models with sequentially reduced number of predictors (ranked by variable importance). Based on this modelling performance, a final panel of CNV events was established to build the prediction model (random forest) for PCa recurrence.

**Gene fusion detection and annotation**

To detect gene fusions in respective PCa samples, we used the software FACTERA (Fusion And Chromosomal Enumeration and Recovery Algorithm) and the GeneFuse program by adopting different algorithms. FACTERA directly processed alignment files (BAM) in three phases: identification of discordant read clusters, detection of breakpoints at nucleotide resolution and *in silico* validation of candidate fusions (18). By contrast, GeneFuse searched for reads that could be mapped into two different genes (left and right sides), but could not be entirely mapped to any position of the whole reference genome (19). These reads were called “supporting reads for potential gene fusions”. GeneFuse started to run with raw sequences (FASTQ) rather than alignment files (BAM). So, compared with FACTERA, GeneFuse could detect more potential gene fusions but required longer computing time. Accordingly, FACTERA aimed at whole genome to search for fusion events involved with all genes, while GeneFuse aimed at all COSMIC (Catalogue of Somatic Mutations in Cancer) curated fusion genes and druggable fusion genes to search for fusion events involved with specific genes. Afterwards, Oncofuse was used to predict the oncogenic potential of all gene fusions with a naive Bayes Network Classifier, trained and tested using the Weka machine learning package (20).

**Results**

*WES of primary PCa, with different prognosis status, after a 5-year clinical follow-up*

We sequenced a total of 44 non-recurrent and 30 recurrent samples by WES. The raw data was submitted to the SRA (Sequence Read Archive) database accordingly [BioProject ID PRJNA496568 (SUB4629515) and SRA submission ID SUB4637601]. After alignment of raw sequences to the human genome (hg19), both coverage rate and average depth were calculated. In most of PCa samples, coverage rate exceeded 99% and the mean coverage rate was 99.4%. The average sequence depth ranged from 20× to 100×, and the average depth of the targeted regions for all samples was 44.8. The detailed description of this dataset has been recently reported (21).

*Copy number gains in the chromosome 8q2 and centromere protein F (CENPF) locus is detected in the recurrent PCa samples*

Using XHMM, a total of 3,484 CNV events have been identified in the PCa samples. Specifically, 764 duplications and 603 deletions were detected in 76.7% (23/30) of recurrent PCa samples, with a median length of 33.29 and 55.94 kb, respectively. A total of 1,302 duplications and 815 deletions were detected in 68.2% (30/44) of the non-recurrent PCa samples, with a median length of 37.44 and 48.41 kb, respectively.

To identify the CNVs specifically related to the recurrent group, we applied a filter (see Methods) and identified a total of 66 specific CNVs (2 deleted and 64 duplicated regions) specific to the recurrent group. Strikingly, most of the specific duplicated regions in the recurrent group were located on chromosome 8q2, including 8q21, 8q22, 8q23 and 8q24, which were previously shown to be associated with PCa progression and recurrence (22). Except for the well-known gene *MYC* (MYC Proto-Oncogene, BHLH Transcription Factor), some PCa-related genes were present in duplicated regions of 8q2, such as *PTK2* (Protein Tyrosine Kinase 2), *RAD21* (RAD21 Cohesin Complex Component), *KIAA0196* (WASH Complex Subunit 5) and *EIF3H* (Eukaryotic Translation Initiation Factor 3 Subunit H). According to the COSMIC database (release v89, 15th May 2019), *RAD21* was duplicated in 158 and deleted in 18 PCa samples. *KIAA0196* was duplicated in 166 and deleted in 6 PCa samples, and *EIF3H* was duplicated in 158
and deleted in 26 PCa samples. No copy number variants involving the \( \text{PTK2} \) gene were reported. In addition, we have found copy number gains of \( \text{CENPF} \) (centromere protein F) on chromosome 1q41 and \( \text{RFPL4A} \) (Ret Finger Protein Like 4A) on 19q13.42, which was reported to be duplicated in 31 and 16 PCa samples according to the COSMIC database, respectively.

Although some of these specific CNVs had been reported in PCa, our findings suggest that these specific duplication or deletion events in recurrent PCa might affect the expression of genes in these chromosomal regions and play roles in the recurrence of PCa. For instance, duplication of 8q24 has been associated with \( \text{MYC} \) overexpression and characterized as an independent predictor of recurrence after RP (22).

### Identification of common CNV events in both recurrent and non-recurrent PCa

In addition to CNV events specific to recurrent and non-recurrent PCa groups, we have also identified 7 duplications and 3 deletions, commonly present in both groups, using the criterion that the specified regions (over 10 kb) should be either duplicated or deleted in, at least, 2 recurrent and 2 non-recurrent PCa samples, with no deletion or duplication in any PCa samples, respectively (Table 1). We noted that one particular CNV resulted in \( \text{CHD1} \) (Chromodomain Helicase DNA Binding Protein 1) gene deletion at 5q21.1. According to the COSMIC database, \( \text{CHD1} \) deletion was detected in 135 recurrent PCa samples. As previously reported, \( \text{CHD1} \) deletion might increase cell invasiveness in PCa, indicating a possible novel role of altered chromatin remodeling during prostate tumorigenesis (23).

### Random forest modelling to predict PCa recurrence

To build a prediction model for PCa recurrence, the output file (vcf) of XHMM software which contained a total of 2,866 CNV events, derived from the genome of all PCa samples, were utilized. After statistical analysis of respective CNV genotypes in every PCa sample, a chi-square test indicated that 51 CNV events were correlated with PCa recurrence. Among these, 34 CNVs could not be genotyped in more than 10% of the PCa samples, and 2 CNV events were too small (DNA length less than 1 kb). As a result, the remaining 15 CNVs were used to build a random forest model. Based on a 10-fold cross validation, a prediction model was built with the six most important CNV events (Table 2) and the cross-validated error rate was 11.3% (Figure 2A). In addition, we conducted a receiver operating characteristic (ROC) analysis of the prediction model, based on these same events, resulting in an AUC (area under curve) of 0.857 for predicting PCa recurrence (Figure 2B).

### Identification of gene fusion events in PCa

We identified a total of 324 gene fusion events by FACTERA. Furthermore, 67 gene fusion events involving COSMIC curated fusion genes and 150 gene fusion

<table>
<thead>
<tr>
<th>Region</th>
<th>Gene</th>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Band</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exonic</td>
<td>( \text{RPTN, TCHH} )</td>
<td>chr1</td>
<td>152086333</td>
<td>152127241</td>
<td>1q21.3</td>
<td>Gain</td>
</tr>
<tr>
<td>Exonic</td>
<td>( \text{OR4C11, OR4C16} )</td>
<td>chr11</td>
<td>55339551</td>
<td>55370953</td>
<td>11q11</td>
<td>Gain</td>
</tr>
<tr>
<td>Exonic</td>
<td>( \text{NOMO2} )</td>
<td>chr16</td>
<td>18570167</td>
<td>18604037</td>
<td>16p12.3</td>
<td>Gain</td>
</tr>
<tr>
<td>Exonic</td>
<td>( \text{TMPRSS11E, UGT2B17} )</td>
<td>chr4</td>
<td>69344564</td>
<td>69416377</td>
<td>4q13.2</td>
<td>Gain</td>
</tr>
<tr>
<td>Exonic</td>
<td>( \text{ZFHX4} )</td>
<td>chr8</td>
<td>77763117</td>
<td>77776784</td>
<td>8q21.11</td>
<td>Gain</td>
</tr>
<tr>
<td>Exonic</td>
<td>( \text{GLI4, ZNF696} )</td>
<td>chr8</td>
<td>144358035</td>
<td>144378371</td>
<td>8q24.3</td>
<td>Gain</td>
</tr>
<tr>
<td>Exonic</td>
<td>( \text{SPANXB1} )</td>
<td>chrX</td>
<td>140084812</td>
<td>140097840</td>
<td>Xq27.1</td>
<td>Gain</td>
</tr>
<tr>
<td>Exonic</td>
<td>( \text{RHD} )</td>
<td>chr1</td>
<td>25599065</td>
<td>25634302</td>
<td>1p36.11</td>
<td>Loss</td>
</tr>
<tr>
<td>Exonic</td>
<td>( \text{KIR2DL1, KIR2DL3, KIR3DL3} )</td>
<td>chr19</td>
<td>55246682</td>
<td>55286924</td>
<td>19q13.42</td>
<td>Loss</td>
</tr>
<tr>
<td>Exonic</td>
<td>( \text{CHD1} )</td>
<td>chr5</td>
<td>98204186</td>
<td>98240822</td>
<td>5q21.1</td>
<td>Loss</td>
</tr>
</tbody>
</table>

Note: gain and loss represent copy number gain (duplication) and copy number loss (deletion) respectively. CNV, copy number variation.
events involving druggable fusion genes were identified by GeneFuse (Table 3). According to the COSMIC database, many of these gene fusions had been previously identified in different types of cancers, including (I) BCR-ABL1, CLTC-ALK, ETV6-ABL1 and RANBP2-ALK in haematopoietic and lymphoid-related cancers, (II) EML4-ALK in lung and thyroid cancers, (III) STRN-ALK in lung, peritoneum and thyroid cancers, (IV) RANBP2-ALK, TPM3-ALK and ATIC-ALK in soft, haematopoietic and lymphoid-related cancers, and (V) NTRK1-TPM3 in thyroid and colorectal cancers. Among these nine fusion genes, seven involved the ALK (ALK receptor tyrosine kinase) gene.

After Oncofuse analysis, we discovered that 18 gene fusions (i.e., 16 unique events) involving COSMIC curated fusion genes and 59 gene fusions (i.e., 33 unique events) related to druggable fusion genes were potentially oncogenic (corrected P value <0.05) (Figure 3). Among these potentially oncogenic gene fusions, 16 events were detected in at least 2 PCa samples which were called as “recurrent gene fusions” (Table 4). From these, 5 gene fusions (BCR-ABL1, EML4-ALK, STRN-ALK, CLTC-ALK and ETV6-ABL1) were previously reported in the COSMIC database, while ETV6-ALK was identified in epithelioid fibrous histiocytoma (24). Ten recurrent oncogenic gene fusions were novel, among which three new fusion partners of oncogenic ALK were discovered (i.e., FRYL-ALK, ABL1-
Table 3 The statistics of gene fusions detected by FACTERA and GeneFuse

<table>
<thead>
<tr>
<th>Type</th>
<th>Total</th>
<th>Putative oncogenic</th>
<th>New</th>
<th>New + putative oncogenic</th>
</tr>
</thead>
<tbody>
<tr>
<td>FACTERA</td>
<td>324</td>
<td>0</td>
<td>324</td>
<td>0</td>
</tr>
<tr>
<td>GeneFuse_COSMIC</td>
<td>67</td>
<td>18</td>
<td>67</td>
<td>18</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>150</td>
<td>59</td>
<td>129</td>
<td>40</td>
</tr>
</tbody>
</table>

Note: The “New” category covers those that were not included in COSMIC database, and oncogenic gene fusions were predicted by Oncofuse. The statistical figures did not eliminate duplicates, which meant that the same fusion events detected in many samples would be re-counted. So the unique gene fusion events would be less than the statistics shown in the figure.

Figure 3 A circos plot of potential oncogenic gene fusion events. A total of 49 unique potential oncogenic gene fusion events were mapped in and shown on this circos plot. The red lines represented intrachromosomal gene fusions and the blue lines represented interchromosomal gene fusions. The shade of the color reflects the frequency of gene fusions. For example, the color of the link between ETV6 and ABL1 was deeper than other links because ETV6-ABL1 fusion was detected in 7 prostate cancer samples.
**Table 4** The 16 putative oncogenic gene fusions that were detected in at least 2 prostate cancer samples

<table>
<thead>
<tr>
<th>Type</th>
<th>ID</th>
<th>bcr</th>
<th>no_bcr</th>
<th>Total (≥2)</th>
<th>P value</th>
<th>COSMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeneFuse_COSMIC</td>
<td>FRYL:ALK</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.001947</td>
<td>N</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>BCR:ABL1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1.3E-05</td>
<td>Y</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>ETV6:ALK</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5.57E-05</td>
<td>N</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>EML4:ALK</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0.000138</td>
<td>Y</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>BCR:ALK</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0.000161</td>
<td>N</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>STRN:ALK</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0.000747</td>
<td>Y</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>RANBP2:BCR</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0.00309</td>
<td>N</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>TPM4:BCR</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0.009371</td>
<td>N</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>BCR:EZR</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.037034</td>
<td>N</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>ETV6:EZR</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0.037034</td>
<td>N</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>CLTC:ALK</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0.000247</td>
<td>Y</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>EML4:NTRK3</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>8.99E-05</td>
<td>N</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>GOPC:NTRK3</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0.00241</td>
<td>N</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>ABL1:ALK</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>0.000138</td>
<td>N</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>ABL1:STRN</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>6.07E-05</td>
<td>N</td>
</tr>
<tr>
<td>GeneFuse_druggable</td>
<td>ETV6:ABL1</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>1.3E-05</td>
<td>Y</td>
</tr>
</tbody>
</table>

Note: “bcr” and “no_bcr” stand for recurrent prostate cancer group and non-recurrent prostate cancer group. The number below stands for the frequency of every gene fusion event in recurrent group and non-recurrent group. “P value” stems from OncoFuse and indicates whether this gene fusion is predicted to be oncogenic. “Y” and “N” in the last column indicate whether this gene fusion had been included in COSMIC database.

**Discussion**

We presently conducted an analysis of CNVs and gene fusion events in PCa, using WES data from primary tumors of 74 PCa patients. This genomic data was originated from 30 recurrent and 44 non-recurrent PCa cases, after 5 years of clinical follow-up. Finally, we identified 66 specific CNVs (2 deletions and 64 duplications) that were specific to primary tumor tissues from the recurrent PCa group. Most of the 64 duplication regions were located at 8q2, including 8q21, 8q22, 8q23, 8q24, indicating that this chromosomal region is an important duplication site for PCa. As previously reported, 8q24 duplication has been associated with MYC overexpression and PCa progression (22). In addition, this duplication event may serve an independent predictor for PCa recurrence after RP (22). The chromosomal 8q24 region has been also associated with certain aggressive forms of PCa (25). Our findings suggest that duplication regions may have occurred specifically in primary PCa tissues that were recurrent after 5-year follow-up. This observation indicates that the 8q2 duplication might be an early event that predisposes PCa to recurrence, and not solely an outcome of PCa recurrence. These duplication regions could be further used to develop assays to predict PCa recurrence, using primary cancer tissues.

Specifically, a high number of cancer-risk SNPs have been detected in the 8q24 region, which also harbors c-MYC (26,27). Another important gene located in the region is the prostate-specific noncoding RNA gene, PCaT-1, whose overexpression promotes proliferation, migration, invasion and inhibits apoptosis in PCa cells (28). In the recurrent PCa group, our data analyses have revealed that copy number gain of 8q2 includes other important genes, such as PTK2 (Protein Tyrosine Kinase 2), RAD21 (RAD21 Cohesin Complex Component) and KIAA0196 (WASHC5, WASH Complex Subunit 5). Based on a large validation cohort, PTK2 was duplicated in 1% of localized PCa and 35% of...
CRPC (castration resistant PCa). Interestingly, inhibition of PTK2 expression significantly affects cell proliferation and migration of PC3 PCa cells (29), suggesting that duplication of PTK2 might be associated with more aggressive PCa. Two additional genes from chromosomal 8q24—RAD21 and KIAA0196—have shown increased expression in PCa and were also duplicated in 30–40% of PCa xenografts and hormone-refractory tumors (30). Moreover, KIAA0196 expression appears to be significantly higher in tumors with gene duplication than those with no CNV. These data suggest that KIAA0196 and RAD21 are putative effector genes for the common duplication of 8q23-24 in PCa (30). The duplication of 8q2 region in recurrent PCa were consistent with our current understanding of the important roles of 8q2 in PCa. In addition, our findings also suggest a potential mechanism of action (MOA) of specific duplicated regions in PCa recurrence and prognosis, by enhancing the expression of genes in these particular chromosomal sites.

We have also identified a duplicated 1q41 region in the recurrent PCa group. This region harbors CENPF (Centromere Protein F). It has been reported that CENPF acts synergistically with FOXM1 to promote PCa growth, and co-expression of FOXM1 and CENPF is a robust prognostic indicator of poor survival and metastasis (31). Overexpression of CENPF has been related to higher Gleason grade, advanced pathological tumor stage, accelerated cell proliferation, and lymph node metastasis, which corroborate its application as a potential biomarker for PCa aggressiveness (32). In addition, up-regulation of CENPF is an independent predictor of poor BCR-free survival (33). Therefore, the copy number gain of CENPF in 1q41 might promote its expression and contribute to prostate recurrence.

Upon CNV analysis, we have also detected several common copy number gain/loss in both the recurrent and the non-recurrent PCa, including CHD1 deletion. As previously reported, CHD1 loss sensitizes PCa to DNA damaging therapy by promoting error-prone double-strand break repair (34). Loss of CHD1 causes defects on DNA repair and enhances the responsiveness to PCa therapy. In fact, CHD1 gene loss may serve as a marker for PCa patient stratification for targeted therapies, such as PARP inhibitors, which specifically affect tumors with HR (homologous recombination) defects (35). In summary, these common CNVs potentially shed light on the mechanism(s) of prostate carcinogenesis.

Remarkably, we have also discovered a total of 324 gene fusion events by FACTERA. Furthermore, we identified 67 gene fusions related to COSMIC curated fusion genes, and 150 gene fusions related to druggable fusion genes, respectively, by GeneFuse (Table 3). We have additionally found fusion events involving the ALK gene, which was the most common type of gene fusion in our data. In this case, a total of 7 recurrent oncogenic ALK fusion events were annotated, where four of these (EML4-ALK, STRN-ALK, CLTC-ALK, ETV6-ALK) have been previously identified in other types of cancers, and the remaining three fusions (FRYL-ALK, ABL1-ALK, and BCR-ALK) are novel events. These seven ALK fusion events were identified, for the first time, in PCa.

ALK is a receptor tyrosine kinase (RTK) that was first identified in anaplastic large-cell lymphoma (ALCL), due to a characteristic NPM-ALK fusion event (36,37). Since then, oncogenic ALK fusions have been identified in several types of cancer. Therefore, treatment with ALK inhibitors has been further considered in cancer therapy. ALK gene fusions have also been implicated in the pathogenesis of many kinds of tumors, including anaplastic large cell lymphomas, non-small cell lung cancer (38), melanomas (39), thyroid cancer (40), renal cell carcinoma (41), colorectal adenocarcinoma (42) and others (43). For instance, EML4-ALK translocation can lead to constitutive ALK kinase activity and represents an oncogenic addiction pathway in lung cancer (38). Intriguingly, ALK fusions have not been previously reported in PCa.

So far, multiple fusion partners of ALK have been identified in anaplastic large cell lymphomas (ALCLs) and inflammatory myofibroblastic tumors (IMTs), including nucleophosmin (NPM), 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase (ATIC), tropomyosin 3 (TPM3), tropomyosin 4 (TPM4), clathrin heavy chain (CLTC), myosin heavy chain 9 (MYH), and the TRK-fused gene (TFG) (44). Renal cell carcinoma (RCC) also contains multiple ALK fusion partners, such as vinculin (VCL), tropomyosin 3 (TPM3), echinoderm microtubule associated protein-like 4 (EML4), hook microtubule tethering protein 1 (HOOK1) and striatin (STRN) genes (41). In particular, we have previously reported that the ALK gene is truncated (mostly by 5’ deletion) in 18 of 281 (6.4%) Chinese PCa cases (45), however, no ALK fusion partners were identified.

In our current study, we have detected three novel ALK fusion events, including FRYL:ALK. FRYL-like transcription coactivator (FRYL) is a putative target of miR-1205, which is encoded by the long non-coding PVT1 gene located...
at 8q24, a PCa susceptibility chromosomal region (46). Indeed, transfection of oligonucleotide mimics of miR-1205 into androgen-independent PC3 cells can lead to a significant decrease in the expression of FRYL, suggesting that FRYL may act as a direct target of miR-1205. Loss of miR-1205 promotes a tumorigenic phenotype in PCa (46). Interestingly, we have observed that the FRYL:ALK fusion event negatively affects the binding site of miR-1205 in 3’UTR of the FRYL gene. Therefore, we concluded that FRYL:ALK gene fusion might protect FRYL expression from miR-1205-dependent inhibition, contributing to PCa tumorigenesis due to elevated FRYL expression.

**Conclusions**

Here we demonstrated that WES is a cost effective approach to identify CNVs and gene fusion events in PCa. Our findings that 8q2 duplication specifically occurs in primary PCa samples that were recurrent, after 5-year follow-up, suggests that this particular duplication is an early genetic event that predispose PCa to recurrence. Meanwhile, a random forest model with most informative CNV events (n=6) were developed. This model could effectively predict PCa recurrence, with an accuracy close to 90%. Furthermore, we found a high frequency of ALK gene fusion events in our dataset. Interestingly, all seven recurrent oncogenic ALK fusion events were here identified, for the first time, in PCa. Altogether, these findings expand our understanding of prostate carcinogenesis and may lead to the development of more specific assays focused on the prediction of PCa recurrence.

**Acknowledgments**

**Funding:** The work was funded by the National Science and Technology Major Project of China (2018ZX10302205), National Key R&D Program of China (2016YFC1303401), National Natural Science Foundation of China (81572909) and a grant from the Medical Science and Technology Project of Zhejiang Province (2017ZD028).

**References**


36. Elmberger PG, Lozano MD, Weisenburger DD, et al. Transcripts of the npm-alk fusion gene in anaplastic large...


