








# NearTrans Can Identify Correlated Expression Changes Between Retrotransposons and Surrounding Genes in Human Cancer

Rafael Larrosa<sup>1</sup>, Macarena Arroyo<sup>2,3</sup>, Rocío Bautista<sup>4</sup>,  
Carmen María López-Rodríguez<sup>3</sup>, and M. Gonzalo Claros<sup>3</sup>

<sup>1</sup> Departamento de Arquitectura de Computadores,  
Universidad de Málaga, 29071 Malaga, Spain  
rlarrosa@uma.es

<sup>2</sup> Unidad de Gestión Clínica de Enfermedades Respiratorias,  
Hospital Regional Universitario de Málaga, Avda Carlos Haya s/n, Malaga, Spain

<sup>3</sup> Departamento de Biología Molecular y Bioquímica,  
Universidad de Málaga, 29071 Malaga, Spain  
{macarroyo,b12loroc,claros}@uma.es

<sup>4</sup> Plataforma Andaluza de Bioinformática,  
Universidad de Málaga, 29590 Malaga, Spain  
rociobm@uma.es

**Abstract.** Recent studies using high-throughput sequencing technologies have demonstrated that transposable elements (TEs) seem to be involved not only in some cancer onset but also in cancer development. New dedicated tools have been recently designed to quantify the global expression of the different families of TEs from RNA-seq data, but the identification of the particular, differentially expressed TEs would provide more profitable results. To fill the gap, here it is presented *NearTrans*, a bioinformatic workflow that takes advantage of gEVE (a database of endogenous viral elements) to determine differentially expressed TEs as well as the activity of genes surrounding them to study if changes in TE expression is correlated with nearby genes. An especial requirement is that input RNA-seq reads must derive from normal and cancerous tissue from the same patient. *NearTrans* has been tested using RNA-seq data from 14 patients with prostate cancer, where two HERVs (HERVH-int and HERV17-int) and three LINE-1 (L1PA3, L1PA4 and L1PA7) were over-expressed in separate positions of the genome. Only one of the nearby genes (ACSM1) is over-expressed in prostate cancer, in agreement with the literature. Three (PLA2G5, UBE2MP1 and MIR4675) change their expression between normal and tumor cell, although the change is not statistically significant. The fifth (LOC101928437) is highly distant to the L1PA7 and their correlation is unlikely. These results are supporting that, in some cases such as the HERVs, TE expression can be governed by the genome context related with cancer, while in others, such as the LINES, their expression is less related with the genome context, even though they are surrounded by

genes potentially involved in cancer. Therefore, *NearTrans* seems to be a suitable and useful workflow to discover or corroborate genes involved in cancer that might be used as specific biomarkers for the diagnosis, prognosis or treatment of cancer.

**Keywords:** Transposon · Transposable element · Cancer  
Mobile element · RNA-seq · Workflow · Human

## 1 Introduction

Currently, cancer is one of the leading causes of morbidity and mortality with variable survival rates depending on the type of cancer. Recent studies have demonstrated that, besides the specific somatic or germinal mutations that drive tumor growth, mobile elements, also known as transposable elements (TEs) are involved in the onset of many human diseases, as well as in the development of established cancers. For example, in epithelial cancer, activation of TEs correlates with their mobilisation and genomic drift [15]. This is due to the fact that TEs are DNA molecules with the ability to move from one place to another in the genome, contributing to genomic instability and causing genetic disorders. Since nearly 50% of the human genome is composed of TEs, cells try to avoid the deleterious consequences of TE activity inducing the inactivation of most TEs by large deletions, stop codons, and frameshift mutations within their open reading frames. It has been recently shown that some human endogenous viral elements (HEVEs) are still active and play a crucial role in placental development in various mammalian species [20].

The study of TEs using high-throughput technologies has been relegated due to the complexity of its measurement and processing, since there is a large number of copies of TEs present throughout the genome. Earlier efforts drove to tools such as *RepEnrich* [9] or *TEtranscript* [14] that were designed to accurately quantify the global expression of the different families of TEs from RNA-seq data, the TE evaluation being based on *RepBase*. Another one, *Lions* [4], has been developed to quantitatively measure and compare the contribution of TEs promoters to their expression in cancer. Recently, *TEtools* [16] has been designed to analyse the TE expression using non-annotated and non-assembled genomes. But better than knowing the activity of a specific family of TEs, the identification of the particular, differentially expressed TEs would provide more profitable results. Our main objective is not related to the detection of TE jumps that can explain a disease, but to design a tool that can identify which copy of the different TEs in human genome presents differential expression when the normal cell becomes a cancer cell. To elucidate this problem, *gEVE* [20], the database of endogenous viral elements (EVEs) including endogenous retrovirus that was developed to investigate the function and evolution of the TEs in mammalian genomes, seems to be more appropriate than *RepBase*. The great advantage of *gEVE* is that it provides nucleotide and amino acid sequences, genomic loci and functional annotations of all EVEs. Particularly, this database describes 33 966

EVEs, 1782 *gag* elements, 1482 *pro* elements, 29 120 *pol* elements, and 1731 *env* elements in human genome. As a result, the bioinformatic workflow *NearTrans*, that is able to determine (i) differentially expressed TEs and (ii) the activity of genes surrounding them to study whether changes in TE expression are related to nearby genes. As a biological model, prostate cancer was elected, a cancer where it was already known that LINE-1 was over-expressed [9].

## 2 Materials and Methods

### 2.1 Input Data

Control (healthy prostate cells) and treatment (prostate cancer) RNA-seq reads from 14 patients from Shanghai Hospital were publicly available from BioProject PRJEB2449 [24]. The main feature of these data is that prostate cancer and nearby normal tissues were paired, since they were sequenced from the same individual.

Information about EVEs in gEVE was downloaded from <http://geve.med.u-tokai.ac.jp/> for the Hg38 human genome in GTF format. Estructural information about human genome Hg38 was downloaded from UCSC web portal (<http://genome.ucsc.edu/cgi-bin/hgTables>). The sequences of the human genome assembly Hg38 were downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/assembly?term=GRCh38>).

### 2.2 Implementation

The double task of *NearTrans* related to differential expression of TEs and expression level of their nearby genes was carried out as follows (Fig. 1), using the same tools for genes and TEs, normal and tumoral prostate, whenever is possible:

1. Data quality control using *SeqTrimNext* (STN) [11] with the specific NGS Illumina configuration parameters to remove low quality, ambiguous and low complexity stretches, adaptors, organelle DNA, polyA/polyT tails, and contaminated sequences while keeping the longest (at least > 20 bp) informative part of the read.
2. Mapping the pre-processed, useful reads to human genome hg38 using *STAR* v2.5 [10] with the following parameters (see the *STAR* help for the meaning of each parameter):

```
STAR --genomeLoad NoSharedMemory --runThreadN 16 $arg_read
--outSAMstrandField intronMotif
--sjdbGTFfile $REF/Annotation/Genes/genes.gtf
--genomeDir $REF/Sequence/STARIndex/index_genome_STAR/
--readFilesIn $file1 $file2 --outFilterMismatchNmax 6
--outFileNamePrefix align_STAR_sorted
--outSAMtype BAM SortedByCoordinate
--twopassMode Basic --outReadsUnmapped None
```

```

--chimSegmentMin 12 --chimJunctionOverhangMin 12
--alignSJDBoverhangMin 10 --alignMatesGapMax 200000
--alignIntronMax 200000 --chimSegmentReadGapMax parameter 3
--alignSJstitchMismatchNmax 5 -1 5 5.

```

- Use the GFFs of hg38 and gEVE with *Cufflinks* (v.2.2.1) [25] followed by *Cuffquant* and then *Cuffdiff*, for assessing expression levels of genes and TEs, respectively, between matched normal and cancer tissues, as described in [13]. *cummeRbund* v3.6 is then pipelined to analyse, explore, manipulate and plot (visualise) the results.
- Selection of differentially expressed TEs using as filters an adjusted  $P < 0.05$  and a  $|\log_2 FC| > 1$ .
- Location of nearby genes using *BEDTools* (v.2.26.0) [22], with the command

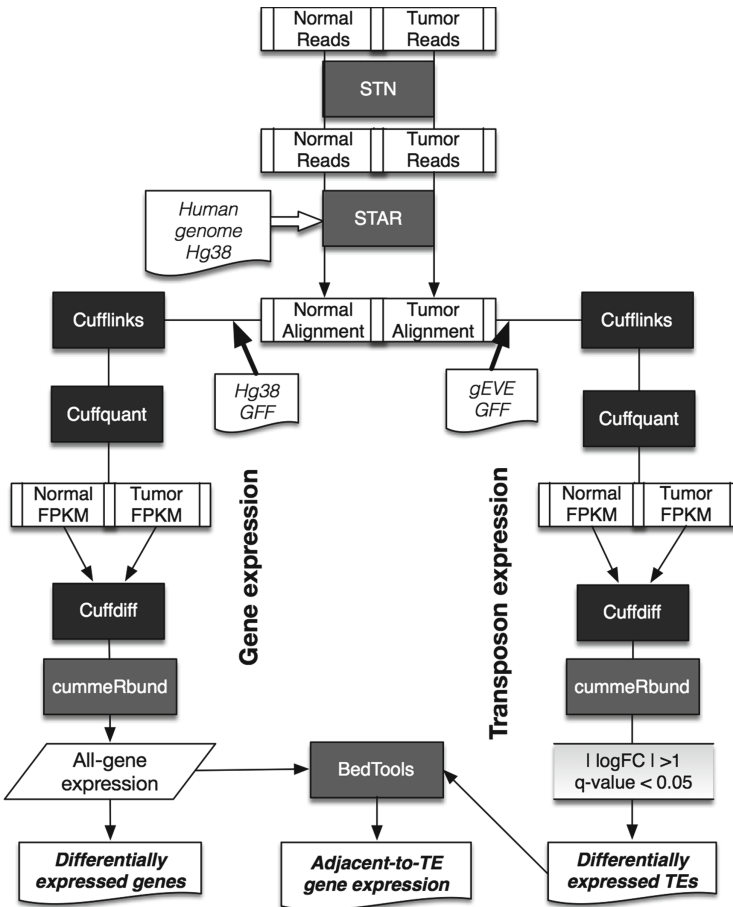


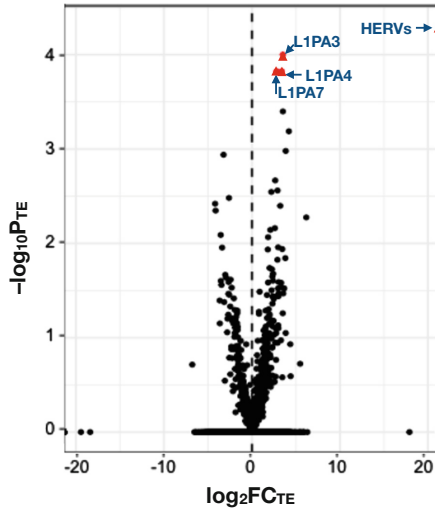
Fig. 1. Flowchart illustrating tools and datasets provided and obtained by *NearTrans* workflow.

```
bedtools closest -a TEs_file.bed -b genes_file.gtf -D a >
nearest_genes.bed.
```

Where the file *TEs\_file.bed* contains the location of the differentially expressed TEs in the human genome and *genes\_file.gtf* contains the location of all genes in the human genome.

### 3 Results

After preprocessing raw RNA-seq datasets data from the 14 prostate cancer patients from PRJEB2449, the percentage of useful reads is in the range of 93.54% for patient ERR031029 to 96.16% for patient ERR031025. This clearly shows the high quality of those sequence reads, and that further analyses will not be affected by read quality. Mapping useful reads resulted in a global 98.18% of the reads mapped on the human genome. Again, the high mapping rate confirms that results will not be affected by inadequate sequencing.



**Fig. 2.** Volcano plot where each TE is defined by its  $\log_2$  of fold-change ( $\log_2 FC_{TE}$ ) vs  $-\log_{10}$  of adjusted  $P$ -value ( $\log_{10} P_{TE}$ ). Dots highlighted in red are those presenting a significant over-expression in prostate cancer cells. The TE corresponding to each red dot is indicated. (Color figure online)

The differentially expressed TEs are shown as red dots in Fig. 2. The three red dots having the  $\log_2 FC_{TE}$  closer to 0 are LINES (L1PA3, L1PA4 and L1PA7), while the upper-right point is for the two HERVs (HERVH-int and HERV17-int). All TEs were found to be over-expressed in prostate cancer: HERVs were not expressed at all on normal cells, but expressed only on cancer cells (this is why they appear at the right border of Fig. 2 and as “Inf” in Table 1). On the contrary, LINES (as many other TEs) were expressed in normal cells and their

expressions were significantly increased in tumor cells. The advantage of using *gEVE* is that now we know that from the 20 699 described positions of LINE-1 in Hg38, 946 were strongly (although not significantly, adjusted  $P > 0.05$ ) repressed, while 3 829 were over-expressed (but only three positions exhibit significant over-expression). The remaining 15 924 positions of LINES can be considered unchanged, since they show a  $\log_2 FC_{TE}$  of  $-0.06$  with a standard deviation of 1.58. These results are highly compatible with the reported over-expression of LINE-1 already described in prostate cancer [9], the main innovation of *NearTrans* being the positions of the LINE-1 copies whose over-expression is significant.

Taking in mind the idea that a TE can only be expressed if its genomic context is not supercoiled (silenced), the chromosome region where each differentially expressed TE is located was screened for the closer gene. It can be seen that distances between genes and TEs is highly variable irrespective of the TE (Table 1). The stronger correlation was observed between the expression of HERV17-int and *ACSM1*, while LINES present the less significant correlation (adjusted  $P > 0.5$ ). Interestingly, expression of MIR4675 (close to L1PA4) that has not been found in the samples analysed. It seems that those HERVs are more dependent on the genetic context than LINES.

**Table 1.** Summary of differentially expressed retrotransposons in prostate cancer and their nearby genes

TE <sup>a</sup>	Chr <sup>a</sup>	$\log_2 FC_{TE}$	$P_{TE}^b$	Gene	$\log_2 FC_g$	$P_g^b$	Distance <sup>c</sup>
HERV17-int	16	Inf	$5.0 \times 10^{-5}$	ACSM1	3.67	$5.0 \times 10^{-5}$	-5 045
HERVH-int	1	Inf	$5.0 \times 10^{-5}$	PLA2G5	0.33	0.25805	-16 585
L1PA3	16	3.51	$1.0 \times 10^{-4}$	UBE2MP1	-0.32	0.50	51 219
L1PA4	10	3.38	$1.5 \times 10^{-4}$	MIR4675	0	1	-5 728
L1PA7	X	2.81	$1.5 \times 10^{-4}$	LOC101928437	3.31	1	211 321

<sup>a</sup>TE: Transposable element. Chr: chromosome.

<sup>b</sup>The  $P$  refers to adjusted  $P$ -value of TEs ( $P_{TE}$ ) and genes ( $P_g$ ).

<sup>c</sup>Distance, in nucleotides, from the TE to te nearby gene; negative values indicate upstream and positive values indicate downstream.

## 4 Discussion

The capabilities of *NearTrans* workflow (Fig. 1) allowed the identification of five TEs (HERVH-int, HERV17-int, L1PA3, L1PA4 and L1PA7) with differential expression in separate positions of the human genome in prostate cancer (Fig. 2 and Table 1). In some cases (HERV17-int and L1PA7), TE over-expression appears to be correlated with high gene expression of their nearby genes (*ACSM1* and *LOC101928437*, respectively). In most cases, the gene is not highly expressed or the correlation is not significant. Even though the statistic significance of these correlations between genes and TEs is significant only in the case of HERV17-int/*ACSM1*, we will examine if nearby genes are related to prostate cancer to

know which TEs are over-expressed due to their proximity to expressed genes that have a role in the development of cancer.

Investigating the roles of the genes identified by *NearTrans* in prostate cancer close to the differentially expressed TEs, (Table 1) we found that:

- *ACSM1* has already been described as highly expressed when compared with the normal prostate tissue [1–3,26], while its expression was decreased when the patients underwent androgen deprivation and a chemotherapy antitumor treatment with docetaxel [23]. It has also been described that the silencing of *ACSM1* in breast cancer decreases the cellular invasion and progression, and therefore it is identified as a potential biomarker for the prognosis of cancer [7].
- *PLA2G5* has variable expression profile and is involved in diseases of immunological nature [5,8]. It was described as repressed in colon adenocarcinoma [19], acute myeloid leukemia [12] and in the leukemic cell line Jurkat [17]. It has been recently related to prostate as highly expressed in normal epithelial cells while repressed by methylation in diseased prostate [18]. In the analysis of *NearTrans*, *PLA2G5* has an adjusted  $P_g = 0.25$  and a  $\log_2 FC_g = 0.33$  (Table 1), indicating that its expression is not so high and not significant.
- *L1PA3* is close to two pseudogenes: *UBE2MP1* is the ubiquitin conjugating enzyme E2 M pseudogene 1 not apparently related with any disease, even though its upregulation was significantly involved in a pathway related to prostate cancer [21]. The HAVANA GTF for Hg38 predicts another closer pseudogene with unknown function, *VN1R68P*, only at 26 nt.
- *MIR4675* is a miRNA that has not been described in prostate cancer but is related with other types of tumors, including adenocarcinoma, colorectal carcinoma, non-small cell lung carcinoma and breast cancer, where its expression is inhibited with respect to normal tissue [6]. In our case it has not been found in the samples.
- We consider that the unknown nature of *LOC101928437*, its distance to *L1PA7* (211 321 nt) and the  $P_g = 1$  completely discard any influence on the expression of *L1PA7*.

In conclusion, *NearTrans* seem to be a suitable and useful workflow for detection of differentially expressed TEs and their nearby genes. It must be noted that *NearTrans* can be applied to any cancer or any other disease, provided that the same individual presents healthy and diseased tissues where the gene expression levels are different, and from which samples can be taken. The results presented regarding HERVs in prostate cancer suggest that they are expressed depending on the nature of the genome context. The over-expression of LINES is compatible with previous reports [9] but *NearTrans* offers more detail since it also indicates which genome copy of the TE is significantly over-expressed. Interestingly, the TEs belonging to LINE1 family appeared as the most genomic context independent, which supports the idea that this type of TE could be used to increase genome instability in cancer, even though the nearby genes could have a potential relation with cancer. We propose then that the study of TEs in cancer can

help in the discovery or corroboration of genes involved in cancer, and can be used as specific biomarkers for the diagnosis, prognosis or treatment of cancer.

**Acknowledgements.** This work was funded by the Neumosur grants 12/2015 and 14/2016, and was also co-funded by the European Union through the ERDF 2014-2020 “Programa Operativo de Crecimiento Inteligente” to the RTA2013-00068-C03-02 of the Spanish INIA and MINECO. The authors also thankfully acknowledge the computer resources and the technical support provided by the Plataforma Andaluza de Bioinformática of the University of Málaga.

## References

1. Alinezhad, S., Väänänen, R.M., Mattsson, J., Li, Y., Tallgrén, T., Tong Ochoa, N., Bjartell, A., Åkerfelt, M., Taimen, P., Boström, P.J., Pettersson, K., Nees, M.: Validation of novel biomarkers for prostate cancer progression by the combination of bioinformatics, clinical and functional studies. *PLoS ONE* **11**(5), e0155901 (2016)
2. Alinezhad, S., Väänänen, R.M., Ochoa, N.T., Vertosick, E.A., Bjartell, A., Boström, P.J., Taimen, P., Pettersson, K.: Global expression of AMACR transcripts predicts risk for prostate cancer - a systematic comparison of AMACR protein and mRNA expression in cancerous and noncancerous prostate. *BMC Urol.* **16**(1), 10 (2016)
3. Alinezhad, S., Väänänen, R.M., Tallgrén, T., Perez, I.M., Jambor, I., Aronen, H., Kähkönen, E., Ettala, O., Syvänen, K., Nees, M., Kallajoki, M., Taimen, P., Boström, P.J., Pettersson, K.: Stratification of aggressive prostate cancer from indolent disease—prospective controlled trial utilizing expression of 11 genes in apparently benign tissue. *Urol. Oncol.: Semin. Orig. Investig.* **34**(6), 255.e15–255.e22 (2016). Seminar on Preservation Strategies in Bladder Cancer
4. Babaian, A., Lever, J., Gagnier, L., Mager, D.L.: LIONS: analysis suite for detecting and quantifying transposable element initiated transcription from RNA-seq. *bioRxiv* (2017)
5. Balestrieri, B., Arm, J.P.: Group V sPLA2: classical and novel functions. *Biochimica et Biophysica Acta (BBA) - Mol. Cell Biol. Lipids* **1761**(11), 1280–1288 (2006)
6. Best, M.G., Sol, N., Kooi, I., Tannous, J., Westerman, B.A., Rustenburg, F., Schellen, P., Verschueren, H., Post, E., Koster, J., Ylstra, B., Ameziane, N., Dorsman, J., Smit, E.F., Verheul, H.M., Noske, D.P., Reijneveld, J.C., Nilsson, R.J.A., Tannous, B.A., Wesseling, P., Wurdinger, T.: RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell* **28**(5), 666–676 (2015)
7. Bockmayr, M., Klauschen, F., Györfy, B., Denkert, C., Budczies, J.: New network topology approaches reveal differential correlation patterns in breast cancer. *BMC Syst. Biol.* **7**, 78 (2013)
8. Boilard, E., Lai, Y., Larabee, K., Balestrieri, B., Ghomashchi, F., Fujioka, D., Gobezie, R., Coblyn, J.S., Weinblatt, M.E., Massarotti, E.M., Thornhill, T.S., Divangahi, M., Remold, H., Lambeau, G., Gelb, M.H., Arm, J.P., Lee, D.M.: A novel anti-inflammatory role for secretory phospholipase A2 in immune complex-mediated arthritis. *EMBO Mol. Med.* **2**(5), 172–187 (2010)
9. Criscione, S.W., Zhang, Y., Thompson, W., Sedivy, J.M., Neretti, N.: Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genom.* **15**(583), 1–17 (2014)



10. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1), 15–21 (2013)
11. Falgueras, J., Lara, A.J., Fernandez-Pozo, N., Canton, F.R., Perez-Trabado, G., Claros, M.G.: SeqTrim: a high-throughput pipeline for preprocessing any type of sequence reads. *BMC Bioinform.* **11**(1), 38 (2010)
12. Fiancette, R., Vincent, C., Donnard, M., Bordessoule, D., Turlure, P., Trimoreau, F., Denizot, Y.: Genes encoding multiple forms of phospholipase A2 are expressed in immature forms of human leukemic blasts. *Leukemia* **23**(6), 1196–1199 (2009)
13. Ghosh, S., Chan, C.K.K.: Analysis of RNA-seq data using TopHat and Cufflinks. *Methods Mol. Biol.* **1374**, 339–361 (2016)
14. Jin, Y., Tam, O.H., Paniagua, E., Hammell, M.: TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**(22), 3593–3599 (2015)
15. Kassiotis, G.: Endogenous retroviruses and the development of cancer. *J. Immunol.* (Baltim. Md.: 1950) **192**(4), 1343–1349 (2014)
16. Lerat, E., Fablet, M., Modolo, L., Lopez-Maestre, H., Vieira, C.: TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res.* **45**(4), e17 (2017)
17. Menschikowski, M., Hagelgans, A., Kostka, H., Eisenhofer, G., Siegert, G.: Involvement of epigenetic mechanisms in the regulation of secreted phospholipase A2 expressions in Jurkat leukemia cells. *Neoplasia* (N.Y.) **10**(11), 1195–1203 (2008)
18. Menschikowski, M., Hagelgans, A., Nacke, B., Jandeck, C., Mareninova, O.A., Asatryan, L., Siegert, G.: Epigenetic control of group V phospholipase A2 expression in human malignant cells. *Tumor Biol.* **37**(6), 8097–8105 (2016)
19. Mounier, C.M., Wendum, D., Greenspan, E., Fléjou, J.F., Rosenberg, D.W., Lambreau, G.: Distinct expression pattern of the full set of secreted phospholipases A2 in human colorectal adenocarcinomas: sPLA2-III as a biomarker candidate. *Br. J. Cancer* **98**(3), 587–595 (2008)
20. Nakagawa, S., Takahashi, M.U.: gEVE: a genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes. *Database* **2016**, baw087 (2016)
21. Ning, Q.Y., Wu, J.Z., Zang, N., Liang, J., Hu, Y.L., Mo, Z.N.: Key pathways involved in prostate cancer based on gene set enrichment analysis and meta analysis. *Genet. Mol. Res.* **10**(4), 3856–3887 (2011)
22. Quinlan, A.R., Hall, I.M.: BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6), 841–842 (2010)
23. Rajan, P., Stockley, J., Sudbery, I.M., Fleming, J.T., Hedley, A., Kalna, G., Sims, D., Ponting, C.P., Heger, A., Robson, C.N., McMenemin, R.M., Pedley, I.D., Leung, H.Y.: Identification of a candidate prognostic gene signature by transcriptome analysis of matched pre- and post-treatment prostatic biopsies from patients with advanced prostate cancer. *BMC Cancer* **14**(1), 977 (2014)
24. Ren, S., Peng, Z., Mao, J.H., Yu, Y., Yin, C., Gao, X., Cui, Z., Zhang, J., Yi, K., Xu, W., Chen, C., Wang, F., Guo, X., Lu, J., Yang, J., Wei, M., Tian, Z., Guan, Y., Tang, L., Xu, C., Wang, L., Gao, X., Tian, W., Wang, J., Yang, H., Wang, J., Sun, Y.: RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res.* **22**(5), 806–821 (2012)

25. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L.: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**(3), 562–578 (2012)
26. Väänänen, R.M., Lilja, H., Kauko, L., Helo, P., Kekki, H., Cronin, A.M., Vickers, A.J., Nurmi, M., Alanen, K., Bjartell, A., Pettersson, K.: Cancer-associated changes in expression of TMPRSS2-ERG, PCA3 and SPINK1 in histologically benign tissue from cancerous versus non-cancerous prostatectomy specimens. *Urology* **83**(2), 511.e1–511.e7 (2014)