Journal of Biomedical Informatics 60 (2016) 422–430

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A novel model used to detect differential splice junctions as biomarkers in prostate cancer from RNA-Seq data

Iman Rezaeian [a,*], Ahmad Tavakoli [a], Dora Cavallo-Medved [b], Lisa A. Porter [b], Luis Rueda [a]

[a] School of Computer Science, University of Windsor, 401 Sunset Ave., Windsor, Ontario N9B 3P4, Canada
[b] Department of Biological Sciences, University of Windsor, 401 Sunset Ave., Windsor, Ontario N9B 3P4, Canada

ARTICLE INFO

Article history:
Received 20 November 2015
Revised 10 February 2016
Accepted 15 March 2016
Available online 15 March 2016

Keywords:
Prostate cancer
RNA-Seq
Alternative splicing
Junction detection

ABSTRACT

Background: In cancer alternative RNA splicing represents one mechanism for flexible gene regulation, whereby protein isoforms can be created to promote cell growth, division and survival. Detecting novel splice junctions in the cancer transcriptome may reveal pathways driving tumorigenic events. In this regard, RNA-Seq, a high-throughput sequencing technology, has expanded the study of cancer transcriptomics in the areas of gene expression, chimeric events and alternative splicing in search of novel biomarkers for the disease.

Results: In this study, we propose a new two-dimensional peak finding method for detecting differential splice junctions in prostate cancer using RNA-Seq data. We have designed an integrative process that involves a new two-dimensional peak finding algorithm to combine junctions and then remove irrelevant introns across different samples within a population. We have also designed a scoring mechanism to select the most common junctions.

Conclusions: Our computational analysis on three independent datasets collected from patients diagnosed with prostate cancer reveals a small subset of junctions that may potentially serve as biomarkers for prostate cancer.

Availability: The pipeline, along with their corresponding algorithms, are available upon request.

1. Background

Prostate cancer is a complex disease, and diagnosis is becoming progressively more prevalent. Worldwide, prostate cancer is the second most common cancer in men with more than one million men diagnosed in 2012, resulting in an estimated 307,000 deaths [1]. As with all cancers, the study of prostate cancer at the molecular level uncovers the regulatory and transcriptional mechanisms of the tumor biology. Currently a top priority in the prostate cancer biology field is to discover biomarkers to differentiate between clinically significant disease with a high risk of progression and clinically insignificant disease with low risk of progression. The lack of such biomarkers is a major obstacle in guiding treatment decisions for prostate cancer patients.

The advent of RNA-Seq has revolutionized the way in which genomic and transcriptomics studies are conducted. RNA-Seq allows for the reading of the transcriptome at a single-nucleotide resolution and reveals unexplored genomic territories [2,3]. This has led to a better understanding of the unknown regulatory mechanisms of transcription and discovery of novel transcripts undetected by conventional tools. This high-throughput technique is currently being used to identify non-conventional biomarkers, such as noncoding RNA, alternative splicing, and gene fusion [2,3]. Of particular interest is alternative splicing of RNA that produces protein isoforms with potentially differing functions. For example, in ovarian and breast tumors, around half of all active alternative splicing events are changed [4]. RNA-Seq can also be used to measure transcriptomic activity and transcriptome assembly [5–8] to better understand the mechanism of alternative splicing and the regulation of corresponding protein isoforms. Since studies using RNA-Seq data for prostate cancer are in early days, there are no standard protocols using RNA-Seq to determine the role of alternate splicing events in initiation, progression and invasion of this disease. Using machine learning approaches for RNA-Seq data analysis, researchers are able to remove redundant and less significant information and provide a selection of potential biomarkers for biological validation using conventional laboratory analysis. Nonetheless, a typical RNA-Seq experiment produces a

* Corresponding author.
E-mail addresses: rezaeia@uwindsor.ca (I. Rezaeian), ahmadtavakkoli@gmail.com (A. Tavakoli), dcavallo@uwindsor.ca (D. Cavallo-Medved), lporter@uwindsor.ca (L.A. Porter), lrueda@uwindsor.ca (L. Rueda).

http://dx.doi.org/10.1016/j.jbi.2016.03.010
1532-0464/

large amount of data, thus demanding significant computational resources in both time and space.

In 2012, Feng et al. presented a comprehensive review of the most recent studies on alternative splicing in cancer using RNA-Seq data [9], including an overview of several publically available RNA-Seq datasets and the most recent open source bioinformatics tools for RNA-Seq data analysis. Recent studies using RNA-Seq for prostate cancer analysis include genome wide association and variation studies, non-coding RNAs (e.g., microRNA, lincRNA and siRNA), somatic mutations, chimeric RNA and gene fusion [10]. Kannan et al. reported on new chimeric RNA events in prostate cancer using RNA-Seq [10]. Their study was conducted on 20 human prostate cancer and 10 matched benign prostate tissues from patients who had received no preoperative therapy prior to radical prostatectomy. They found a small group of 27 novel highly recurrent chimeric RNA events within the prostate cancer tissues only, suggesting a link between increased chimeric RNA events and prostate cancer.

Pflueger et al. [11] used RNA-Seq in 25 human prostate cancer samples to identify novel gene fusions. They reported seven new gene fusions related to prostate cancer, including TMPRSS2-ERG. TMPRSS2-ERG gene fusion is present in 50–90% of human prostate cancers and has been identified as an early molecular event associated with invasion of the disease [12]. Ren et al. also conducted a study to identify recurrent gene fusions in 14 primary prostate tumors from a Chinese population. Although they reported that TRMPRSS2-ERG fusion occurs at very low frequency, they also identified two novel gene fusions, CTAGE5-KHDRBS3 and USP9Y-TTTY15, which occur frequently in the Chinese cohort, [13]. These conflicting reports also illustrate the disparities among prostate cancer patients of different ethnic backgrounds.

Xu et al. identified 92 new genes with somatic mutations in human prostate cancer [14]. Their study used RNA-Seq data from five cancer patients to detect variants of chromosomal rearrangements, insertions and deletions. Of particular significance, they identified a frame shift mutation in the coding region of *TNFSF10* that disrupts its ability to induce apoptosis, and hence, promotes abnormal tumor progression. Prensner et al. [15] focused on new noncoding RNA, finding an unannotated lincRNA, PCAT-1, discovered to be a prostate specific regulator of cell proliferation.

Many recent studies use methods to find genes (related to splice junction or chimeric events) that act as drivers of cancer, they fail however to exploit the high-resolution features of RNA-Seq. Reconstructing the transcriptome, inferring protein isoforms, and the corresponding protein functions and interactions as participants in transcriptional and regulatory pathways, offer an integrative approach worth implementing.

There are different methods for finding alternative splicing events in RNA-Seq data. PASSion [16] is a pattern growth algorithm-based pipeline for splice site detection in paired-end RNA-Seq reads with the ability of detecting junctions that do not have known splicing motifs, and which cannot be found by other tools. TopHat2 [17] aligns reads with various lengths and allows for variable-length insertions and deletions with respect to the reference genome. It also has the ability to identify novel splice sites with direct mapping to known transcripts. rMAT [18] is a statistical tool for detecting differential alternative splicing events from replicate RNA-Seq data. rMAT has the ability to analyze both unpaired replicates and paired replicates, such as case-control matched pairs in clinical RNA-Seq datasets. spliceR [19] is another method for detecting single or multiple exon skipping, intron retention and mutually exclusive exon events. spliceR is also able to annotate the genomic coordinates of the differentially spliced elements and facilitate the downstream sequence analysis. One of the drawbacks of these methods is that they tend to find the spliced sites in each sample separately and there is no procedure for tracking those sites across different samples within a dataset and identify those that are consistently present in different samples.

In this paper, we propose a novel model for detecting differential splice junctions in prostate cancer using RNA-Seq data. The model considers an integrative approach that includes a new two-dimensional peak finding algorithm to combine and remove irrelevant junctions and a scoring mechanism to select the most informative junctions. Our analysis on three independent datasets reveals a small subset of 12 junctions that could be used as potential biomarkers for prostate cancer. The main contributions of this study are: (i) developing a model for combining and filtering out splice junctions on large scale data using peak-finding in 2-D histograms, and (ii) designing a method used to identify splice junctions as biomarkers based on transcriptomic measures among cancer samples.

2. Results

This study involves computational experiments on three independent datasets (see Methods for descriptions of the datasets). We used Kannan's and Ren's datasets for obtaining the most significant junctions, and Rajan's dataset as an independent set to validate our findings. In the first computational experiment, we used PASSion with default parameters [16] to obtain a total of 2,325,352 splice junctions across all chromosomes from the 20 cancer samples in Kannan's dataset. From Ren's dataset, 2,032,719 junctions across 14 cancer samples were obtained.

Fig. 1 shows the distribution of the junctions across different scores found in both datasets. The *x*-axis represents the score while the *y*-axis contains the number of junctions. As seen in the figure, most junctions fall around the middle of the score spectrum (-2, -1, 0, 1, 2), while only a fraction of junctions at the maximum ends of the spectrum (-20, 20). This suggests that only a small percent of junctions (0.11% and 0.34%, in Kannan's and Ren's datasets respectively) exist in all cancer samples. Fig. 2 shows the number of junctions common to both datasets. As seen in the figure, 12 junctions are present in at least all but one cancer sample in both datasets. Out of these 12 junctions, 4 of them are found in all cancer samples for both datasets. The small number of common junctions between these datasets may be due to the demographic differences between the samples. While patients in Kannan's dataset are Caucasian, Ren's dataset contains cancer samples from a Chinese population.

We validated the detected junctions using 4 samples from Rajan's dataset corresponding to prostate cancer patients prior to any treatment. Table 1 shows the 12 junctions common in both datasets along with their corresponding scores in Rajan's dataset. The first three columns show the genomic positions of the junctions, while the next three columns show the number of samples that contain each particular junction. The final score column unifies the results of all three datasets in one measure, yielding the percentage of tumor samples that contain each particular junction across all three datasets. For example, junction 1 (Table 1, Row 1), was found in 19 out of 20 tumor samples in Kannan's dataset and in all 14 tumor samples in Ren's dataset. This particular junction was only found in 2 of the 4 samples in Rajan validation dataset. Hence, the total number of samples containing junction 1 is 35 out of a total of 38 samples among the three datasets, or equivalently 92.1% of all samples. The last column shows the genes corresponding to each junction, which have been obtained using the BioMart tool [20].

As seen in Table 1, 10 out of 12 junctions have perfect score (4 out of 4) in Rajan's dataset. This shows the power of generalization of the proposed method to identify junctions across different prostate cancer datasets.

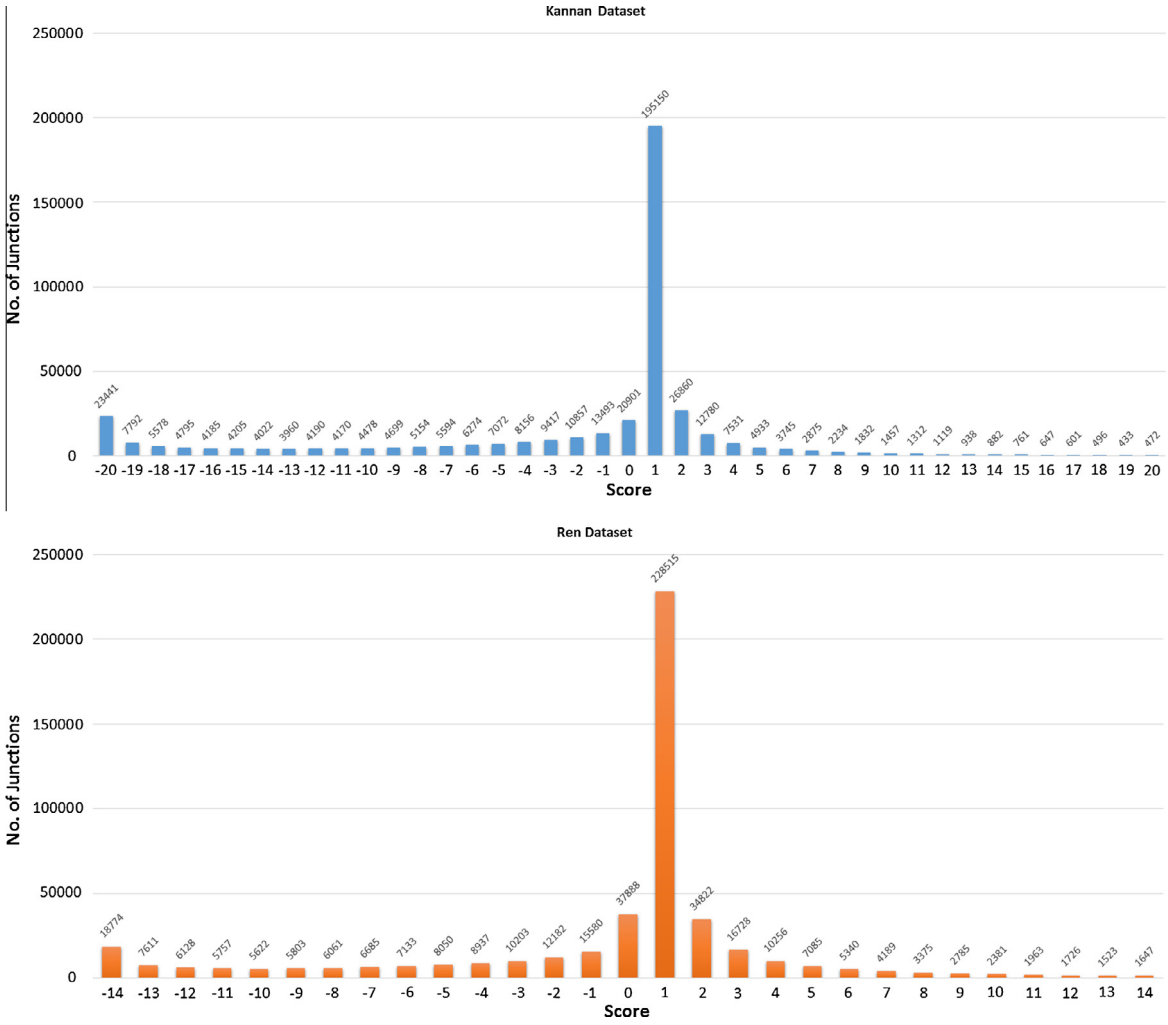

Fig. 1. Distribution of junctions across different scores in Kannan's dataset (top) and Ren's dataset (bottom).

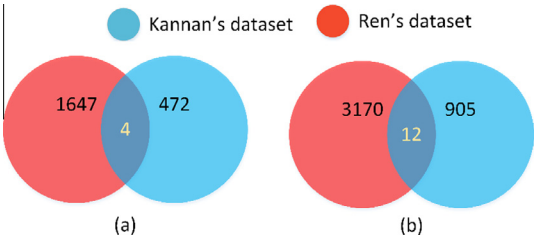

Fig. 2. Venn diagram illustrating the number of junctions commonly identified in both Kannan's and Ren's datasets. (a) Junctions present in all samples and (b) junctions present in at least all but one sample.

In the next step, the Human Protein Atlas [21] was used to study previous annotations of these genes and their association with prostate cancer. As shown in Table 2, prostate cancer tissue staining for protein products of these genes was estimated at four different levels, including *high*, *medium*, *low* and *not detected*. The last column of the table, Normal Tissue Staining, represents the level of staining for the protein product of that particular gene in noncancerous tissue.

Prostate cancer-staining information was found for 11 out of the 12 studied genes. Although, there was no information available regarding prostate cancer-staining of PMEPA1, previous study showed an association between this gene and prostate cancer progression [22]. Among the genes found, GALNT3 appears to have high staining in prostate cancer tumor cells, while only medium staining in normal prostate tissues, which is interesting considering its link to prostate cancer [23]. CUL9, on the other hand, has medium staining in normal cells, while there was no detectable staining in the majority of the prostate cancer samples. IL17RA also stains low in normal prostate cells, while there was no detectable staining in the majority of the cancer samples. NIPAL3, as another example, had medium staining in the majority of cancer samples, while in the normal samples, the staining is usually high. It needs to be mentioned that, sometimes, it is possible that the degree of staining does not correlate directly with RNA splicing events, because the staining may not recognize or distinguish different protein isoforms depending on epitope for the antibody.

The circos plot of Fig. 3 shows the histograms of the junctions in Kannan's dataset (red) and in Ren's dataset (blue), as well as the genomic position of the genes containing detected junctions. The height of the histogram shows the score of the junctions at each locus (genomic position).

We also analyzed the complexity of our proposed model based on the input size (number of junctions). The overall complexity of the model is of $O(n^2)$, which shows that the elapsed time for finding the significant junctions is proportional to the square of the number of junctions used as input of the model. More details about the complexity analysis of the proposed model can be found in Appendix A.

Table 1
Identified junctions that are in common in all datasets with the highest score.

No.	Chr.	Start	End	Kannan's dataset score	Ren's dataset score	Rajan's dataset score	Final score (%)	Gene
1	1	24,790,608	24,792,492	19	14	2	92.1	NIPAL3
2	1	206,765,177	206,766,959	20	14	4	100.0	EIF2D
3	2	166,621,566	166,626,686	19	13	4	94.7	GALNT3
4	3	120,347,375	120,351,983	19	14	4	97.4	HGD
5	3	131,677,802	131,678,144	20	14	4	100.0	CPNE4
6	4	141,310,458	141,311,773	19	14	4	97.4	CLGN
7	6	43,188,643	43,188,886	20	13	4	97.4	CUL9
8	15	64,373,349	64,380,885	19	14	4	97.4	FAM96A
9	19	35,612,012	35,612,124	20	14	4	100.0	FXYD3
10	20	56,234,753	56,284,530	20	13	4	97.4	PMEPA1
11	22	17,583,191	17,584,378	20	14	4	100.0	IL17RA
12	X	21,995,354	21,996,078	19	14	3	94.7	SMS

Table 2
Cancer-staining information of the genes that contain the selected junctions with prostate cancer.

Gene	Chromosome	Junc. Start	Junc. End	Final Score	Prostate Cancer Tissue Staining				Normal Tissue Staining
					High	Medium	Low	Not detected	
NIPAL3	1	24790608	24792492	92.1%	36.4%	63.6%	0.0%	0.0%	High
EIF2D	1	206765177	206766959	100.0%	0.0%	18.2%	27.3%	54.5%	Not Detected
GALNT3	2	166621566	166626686	94.7%	66.7%	25.0%	8.3%	0.0%	Medium
HGD	3	120347375	120351983	97.4%	0.0%	45.4%	27.3%	27.3%	Medium
CPNE4	3	131677802	131678144	100.0%	0.0%	0.0%	0.0%	100.0%	Not Detected
CLGN	4	141310458	141311773	97.4%	10.0%	40.0%	20.0%	30.0%	Medium
CUL9	6	43188643	43188886	97.4%	0.0%	25.0%	8.3%	66.7%	Medium
FAM96A	15	64373349	64380885	97.4%	0.0%	0.0%	27.3%	72.7%	Not Detected
FXYD3	19	35612012	35612124	100.0%	8.3%	66.7%	25.0%	0.0%	Medium
PMEPA1	20	56234753	56284530	97.4%	---	---	---	---	---
IL17RA	22	17583191	17584378	100.0%	0.0%	8.3%	8.3%	83.4%	Low
SMS	X	21995354	21996078	94.7%	0.0%	16.7%	8.3%	75.0%	Not Detected

3. Discussion and conclusion

This study uses a novel method for detecting differential splice junctions in prostate cancer using RNA-Seq data. One of the main differences between the proposed model and other existing methods is that instead of analyzing spliced events for each sample separately, we unify all junctions corresponding to different samples and identify those that are present in most of the cancer samples but not in normal samples or vice versa. Computational analysis of three independent datasets from prostate cancer patients revealed a small subset of junctions that may serve as biomarkers for prostate cancer. Of the 12 junctions isolated, several had predicted roles as tumor suppressors; for example PMEPA1 has been implicated as a tumor suppressor functioning downstream of TGF-B signaling during the progression of prostate cancer [24]. CUL9 functions as a p53 binding protein and induces rapid tumorigenesis when deleted in mouse models [25] and FAM96A is a pro-apoptotic tumor suppressor deleted in gastrointestinal stromal tumors [26]. It will be an important next step to determine how alternate splicing of these proteins may impact their tumor suppressor functions, and how this may participate in the progression of prostate cancer. Other isolated candidates may actively promote tumorigenesis via the alteration in splicing – one such example is the cytokine receptor IL17RA. Interleukin-17 (IL-17) works though its receptor to promote the pathogenesis of many inflammatory disorders and elevated levels of IL-17 are associated with risk of tumor progression [27,28]. Alternate splice variants of the receptor have been noted to both facilitate and to antagonize signaling [29], hence determining the biological role of this novel predicted splice site and the implications of this in the progression of prostate cancer is an exciting and important next step. There are also splice sites detected in candidates not previously implicated in cancer, representing avenues that may provide completely novel insight into prostate cancer initiation and/or progression.

4. Methods

The proposed method consists of various steps that include detecting junctions, unifying junctions with almost identical start and end positions, filtering out less common junctions using a proposed scoring model, and finally, using a SVM-based classifier to identify the most significant junctions. A diagram of the entire

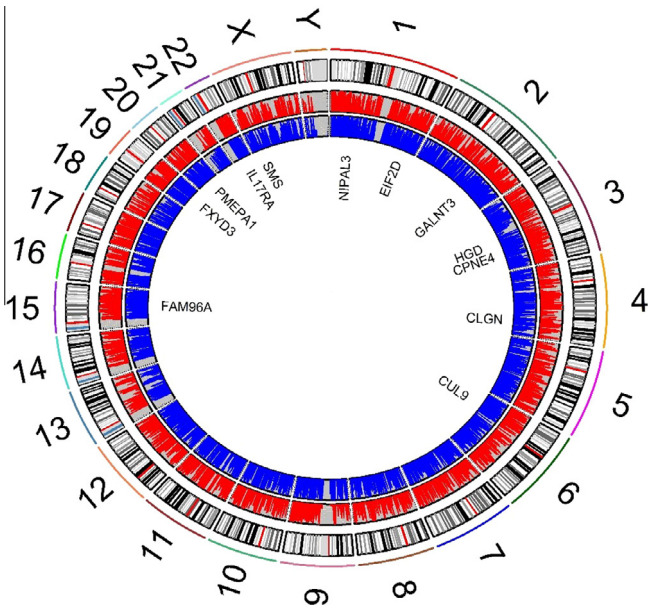

Fig. 3. The most significant junctions detected using samples in both Kannan's and Ren's datasets.

process is shown in Fig. 4. Each step of the process is discussed in more detail in the following sections.

4.1. Dataset

In the computational experiments, we have used three datasets consisting of raw RNA-Seq data obtained from patients diagnosed with prostate adenocarcinoma. The first dataset contains data from a previously published study by Kannan et al. [10], which is publicly available in the GEO repository under accession number GSE22260. The dataset contains 20 samples from patients who did not receive any preoperative therapy prior to radical prostatectomy. The dataset contains more than 667 million paired-end RNA-Seq reads that have been acquired using the Illumina Genome Analyzer II platform, and which are stored in SRA format. Each SRA file contains short reads of 36 bp in length for both forward and reverse strands. The insert size for the prostate cancer dataset is 150 bp.

The second dataset contains data published by Ren et al. [13], and is publicly available in the SRA repository under accession number ERP00550. The dataset contains 14 prostate cancer samples from 14 different patients with various stages of prostate cancer, from Stage T1 to T4. Each sample contains an average of more than 66 million reads around 6 Giga bases of sequenced nucleotides.

We also used a third independent dataset to validate the results. This publicly available dataset, which has been published by Rajan et al. [30], can be obtained via the GEO repository under accession number GSE51005. We used four pre-treatment samples corresponding to four patients with newly diagnosed advanced/metastatic prostate cancer.

Moreover, as a reference, we used Illumina Body Map 2.0, which consists of 16 human tissue types, including prostate [31]. The raw reads corresponding to each tissue were aligned to the genome and then linked exons into tissue-specific transcript models using the reads that span an exon-exon boundary.

4.2. RNA-Seq preprocessing

Since most of the well-known software packages, as well as all packages that we use in this study, are only compatible with FASTA/FASTQ format, we used SRAToolkit [32], developed by the National Center for Biotechnology Information (NCBI), with

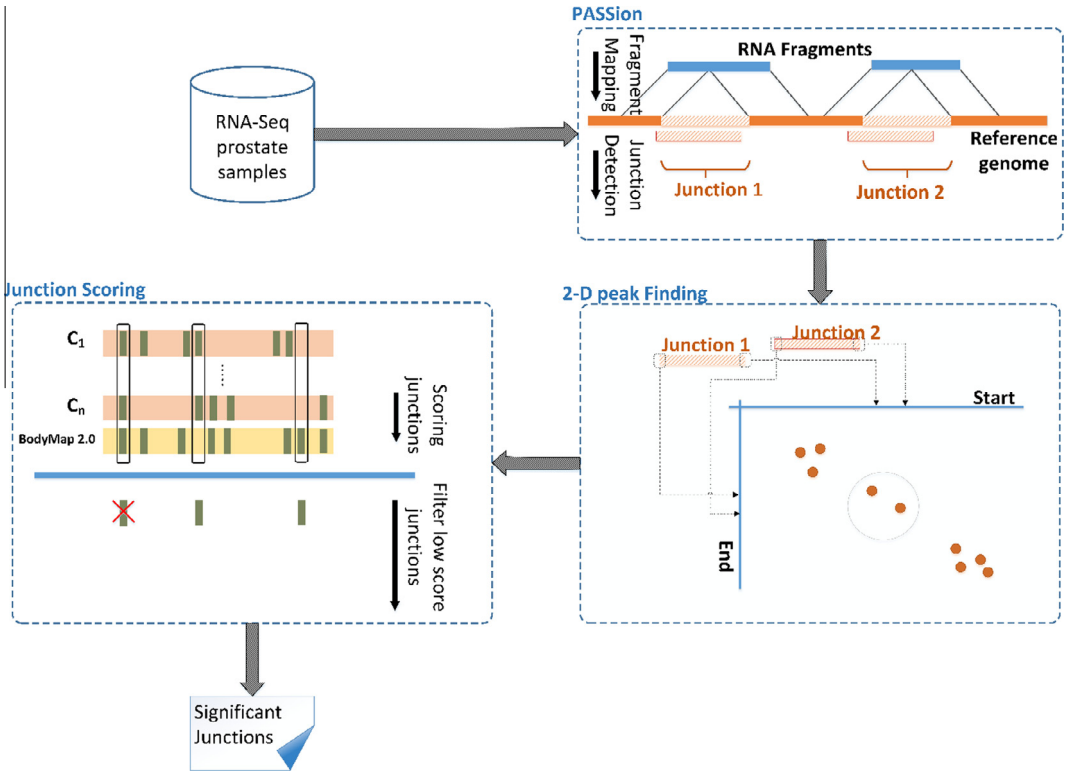

Fig. 4. Schematic view of the entire process for the proposed model.

"split-3" as parameter in order to obtain the FASTQ files in paired-end format and feed them into the PASSion.

4.3. Splice junction detection

4.3.1. Reference genome alignment

In order to align the reads to the reference genome, we used SMALT and SAMTools packages [16]. We also used the Human Genome, Build 37 (GRCh37.p10) from the Genome Reference Consortium [33], as the reference genome, which constituted the input for PASSion and SAMTools.

4.3.2. Finding significant junctions

PASSion has been used for this step [16]. All optional arguments have been set to default values as recommended by Zhang et al. [16]. One of the important parameters, the cut-off limit, was set to 0.1. This parameter implies that any junction where its cut-off score falls short of this limit will be discarded.

Other important parameters include maximum number of SNPs allowed, which was set to 2, while minimum intron size was set to 20, and the sequence error rate was fixed at 0.05. PASSion includes details about the mapping of the split reads across each exon-exon junctions in its output files.

PASSion stores the found splice junctions using BED format, which supplies the means to store data for an annotation track as standardized by the UCSC genome browser [31]. This format requires three fields (chromosome name, the start and end position of the desired feature in chromosome) as required. The start and end positions of a junction can be calculated using the following equations [34]:

$$\text{Junction}_{\text{start position}} = \text{chromosome start position} + \text{block start} \quad (1)$$

$$\text{Junction}_{\text{end position}} = \text{chromosome end position} - \text{block end} + 1 \quad (2)$$

4.4. Selecting junctions

Kannan's and Ren's datasets contain 20 and 14 cancer samples, respectively. Due to these low sample numbers and also the high probability of base pair errors in the mapping process, we identified the same junctions across all different samples with high accuracy. This step is necessary because a single base pair error introduced in mapping, in either the start position or the end position of a junction, could compromise the robustness of the whole process for that junction. We developed a method to filter out less relevant junctions to improve the accuracy of our method in finding more meaningful junctions.

This issue can be modeled as a peak-finding problem in a three-dimensional space. The solution to the problem of finding splice junctions is based on alternative splicing, which means that in some cancer samples, mRNA has been spliced differently than in normal samples. We designed a fast and efficient method to account for the inherent differences that have been introduced to the system by running peak finding separately for start and end positions.

Since start and end positions of each junction are the boundaries between exons and introns, we expect that in case of alternative splicing in each sample the same position would happen in other junctions but with different start or end positions. It is important to note that studying alternative splicing is possible because we are mapping the reads against the reference genome.

4.4.1. 2-D peak finding algorithm

Each detected junction has a specific start and end position, which are independent from each other. One way to find coincident junctions across different samples is to map the start and end positions onto the two-dimensional space as in Fig. 5. The *x*-axis corresponds to the start position, while the *y*-axis corresponds to the end position. Once all junctions in all samples are mapped onto the two-dimensional space, we construct a two-dimensional histogram using the number of samples that contain that junction as "frequency". Due to slight misalignments (insertions, deletions and/or substitutions) a junction in one sample could appear in the vicinity as the same junction for another sample. Then, once the histogram is constructed, significant peaks or "clusters" are found by a new procedure described below. Although the problem of finding centers or clusters in a two dimensional space is computationally intractable (indeed, the *k*-center problem is NP-complete [35]), the advantage here is that the histogram is rather sparse – the peaks will tend to spread along the main diagonal (Fig. 6). This is important, as there are a significantly large number of peaks, in the order of 400,000.

The peak finding module in the two-dimensional histogram proceeds as follows. First, we process the histogram for "start" positions by splitting the table into several smaller windows, transform the data into a full matrix for each of them, and then process the data in each window separately. For each start point, only end points are deemed fit if they have their start position in the vicinity of our unified start positions, and hence these points act as a mean to limit the searching space to find a local maximum. To obtain the final junctions, we run our peak finding algorithm on the end points. Finally, we merge the results into a new sparse matrix structure. We also implemented a safety mechanism to ensure that no peak occurs within the boundaries of a window; for junctions occurring at the edge of a window, the window is adjusted to reach a length of at least 5 bp. We have developed a module, which uses MATLAB to find the rough peaks [36], and identified junction along the whole chromosome using the sparse matrix as described.

We define a parameter called margin to be passed to this module as a minimum peak distance variable, which defines the minimum distance between two peaks. After the peak finding process on start positions finishes, if position *a* is found as a peak, we

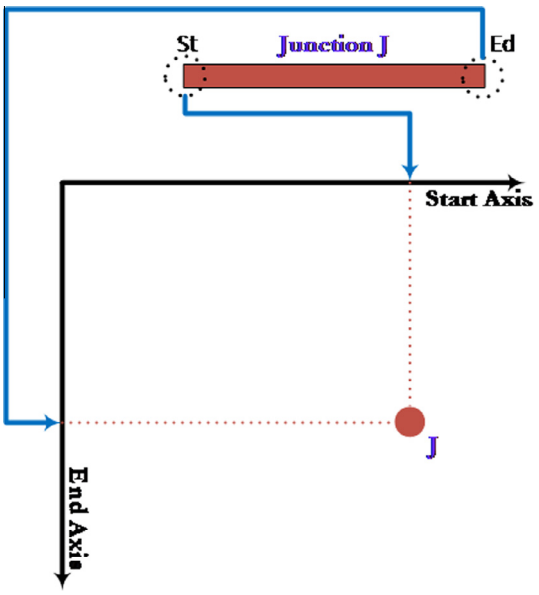

Fig. 5. Transforming each junction into a point.

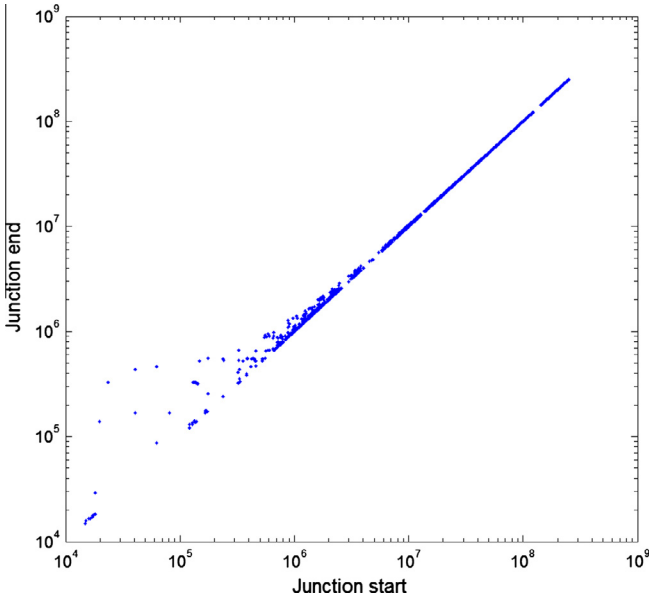

Fig. 6. Distribution of all junctions in chromosome 1 based on their start and end positions.

Table A2
Running time of different steps in our proposed model. The times are given in terms of the input size or number of junctions.

Step	Running time
Combine junctions of all samples and separate them based on their chromosome	n
Project junctions to the x-axis based on their start position	n
Create the histogram of the projected junctions on x-axis	n
Run FindPeaks matlab function on start histogram	Cn
Finding corresponding end positions for each junction	n
Project junctions to the y-axis based on their end position	n
Run FindPeaks matlab function on end histogram for each start position	Dn^2

search for all junctions that have a start position in the neighborhood of $[a - \text{margin}, a + \text{margin}]$, yielding a set of end points for the selected start points. Since the peak finding module discards some start points near some peak, the margin parameter is used to account for them when analyzing end points in the next step. Considering the minimum length of a junction as found in both datasets, which is 19 bp, 5 bp is selected as the vicinity for combining junctions. This gives a safe margin near a quarter of the size of the minimum intron. We chose margin of 2 to cover a 5 bp area of the genome for each junction position. Algorithm A1 described in Appendix A shows the pseudo-code of our proposed 2-D peak finding method.

4.4.2. Scoring junctions and thresholding

Each junction is scored by following a scheme based on the number of samples in which the junction is present for each particular class. Since we use Illumina Body Map 2.0 prostate sample as our control, the ratio between cancer and control sample would be 20 to 1 in Kannan's dataset and 14 to 1 in Ren's dataset.

To compensate for the imbalance in the number of cancer versus control samples, our model considers a +1 score for each sample that belongs to the cancer group, and score C, as a compensation parameter for each junction that is present in the control sample. C is set to the number of samples in each cancer dataset. By adding this parameter, we can equalize the overall score for both cancer and control samples irrespective of their imbalance sample size. The scoring formula is defined in Eq. (3). This scoring scheme accounts for the imbalance that exist between the two classes.

$$\text{Score}_{\text{junction}} = (\#\ \text{of Junctions}_{\text{control}} \times C) + (\#\ \text{of Junctions}_{\text{cancer}} \times 1) \quad (3)$$

where the C parameter is -20 and -14 for Kannan's and Ren's datasets respectively.

Then, we limit the number of junctions reported by the filtering mechanism based on a defined minimum score. As high scoring junctions have occurred more frequently in only one of the groups, they are expected to be more relevant as features for separating cancer and normal samples.

Competing interests

The authors declare that they have no competing interests.

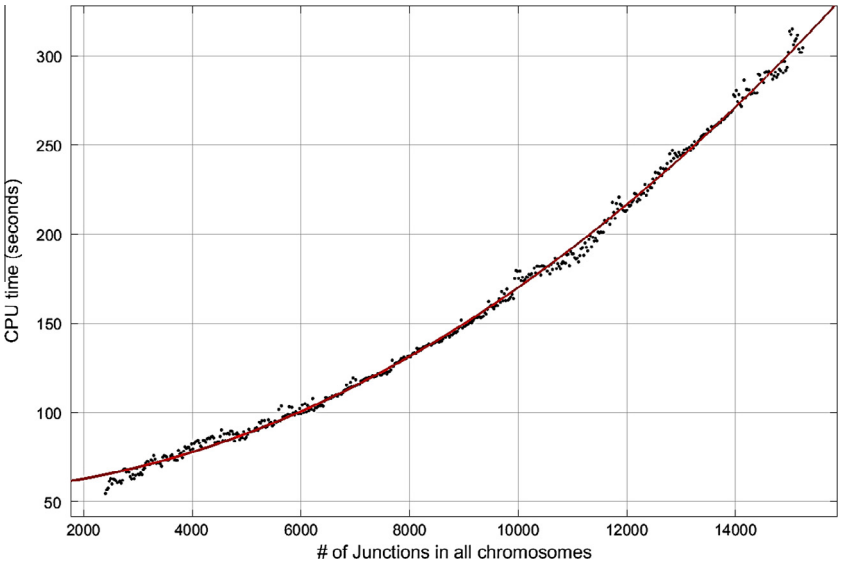

Fig. A3. The elapsed time for finding the significant junctions using different number of junctions as input size.

Authors' contributions

LR and IR conceived the model. IR and AT implemented the algorithms and conducted the experiments. DC and LP conducted the biological validation. All authors have read and approved the final manuscript.

Acknowledgments

This work has been supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), a Seeds4Hope grant from the Windsor Essex County Cancer Centre Foundation (WECCCF), and funding from the University of Windsor.

Appendix A

A.1. The 2D peak finding algorithm

Algorithm A1 shows the pseudo-code of our proposed two dimensional peak finding method, *PeakFinding2D*. The algorithm receives junctions from all samples as input. In the first step, the junctions from all samples are separated based on their corresponding chromosome. In our proposed method, each junction j will be treated as a point in the two-dimensional space, where the x-axis represents the start positions and the y-axis represents the end positions of the junctions.

Algorithm A1. Two-dimensional peak finding method for unifying junctions across different samples.

```
Algorithm PeakFinding2D
Input: Junctions table of size samples × chromosomes
Output: Set of uni-junctions {i, endPosition(i)} per
   chromosome
margin ← 2 bp;
for each chromosome C do
   Positions ← combine junctions of all samples corresponding to
   C;
   Project Positions to the x-axis;
   Peaks ← FindPeaks(Positions,margin);
   For each startPosition i in Peaks do
      for j ← i−margin to i+margin do
         endPositions (j) ← end points for all junctions starting
    on j;
      end for
      endPosition(i) ← FindPeaks(endPositions(j),margin);
   end for
end for
```

Next, we project the junctions onto the x-axis and create a histogram, where the height of the histogram shows the number of junctions corresponding to each locus in the chromosome. In this step, we use the *FindPeaks()* function from the Matlab Signal Processing Toolbox to find the local maxima (peaks) of the junction's starting position. The parameter *margin* defines the minimum distance between two peaks. After the peak finding process on start positions finishes, if position a is found as a peak, a search for all junctions that have a start position in the neighborhood of [a − margin, a + margin] is performed, yielding a set of end points for the selected start points. Since the peak finding module discards some start points near some peak, the margin parameter is used to account for them when analyzing end points in the next step. Considering the minimum length of a junction as found in both datasets is 19 bp, 5 bp is selected as the vicinity for combining junctions. This gives a safe margin near a quarter of the size of the minimum intron. The margin is set to 2 in order to cover a 5 bp area of the genome for each junction position. In the next step, the same procedure is repeated for end positions, again using the *FindPeaks()* function from the Matlab Signal Processing Toolbox to unify the endpoints. Finally, a set of (start, end) positions corresponding to those unified junctions across different samples are returned as the output.

A.2. Complexity of the 2D peak finding algorithm

The input to the algorithm is the set of detected junctions obtained from PASSion. If the number of junctions is, n, the running time of each different step of the process is listed in Table A2.

Here, C and D are constants. Thus, since we have a set of sequential processes for which the highest running time is proportional to the square of the input size (number of junctions), the overall complexity of the model is $O(n^2)$. We also measured the elapsed time for finding significant junctions by using different number of junctions as input size. The experiments were run on a desktop computer with Windows 10 operating system, Intel i7-4770 processor and 32 GB of RAM. We also used Matlab 2015a for these experiments. Fig. A3 shows the CPU times for the whole process, where the x-axis represents the number of junctions and the y-axis shows the measured time for obtaining the output of the model (significant junctions). To make sure the comparison is not biased by the junctions selected from different chromosomes, we used 100 junctions per chromosome as the starting point and extend it to 600 junctions per chromosome. As we observe in the plot, the CPU time is proportional to the square of the number of junctions used as input of the proposed model. A quadratic function $y = 9.913e{-}07x^2 + 0.001506x + 55.97$ was fitted to the points in the plot (solid line) with $R2 = 0.9979$, where x is the number of junctions and y is the CPU time in seconds. As shown in the formula, the coefficient of x^2 is very small (9.913e−07). Considering the total number of junctions for all patients in all chromosomes in the dataset, which was around 3 million junctions, the complexity of the model is very close to linear.

References

[1] World Health Organization, GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012, <http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx>, 2012 (accessed July 10, 2015).
[2] M. Garber, M.G. Grabherr, M. Guttman, C. Trapnell, Computational methods for transcriptome annotation and quantification using RNA-seq, Nat. Meth. 8 (6) (2011) 469–477.
[3] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (1) (2009) 57–63.
[4] J.P. Venables, R. Klinck, C. Koh, J. Gervais-Bird, A. Bramard, L. Inkel, et al., Cancer-associated regulation of alternative splicing, Nat. Struct. Mol. Biol. 16 (6) (2009) 670–676.
[5] T. Steijger, J.F. Abril, P.G. Engström, F. Kokocinski, T.J. Hubbard, R. Guigó, et al., Assessment of transcript reconstruction methods for RNA-seq, Nat. Meth. 10 (12) (2013) 1177–1184.
[6] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. Van Baren, et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat. Biotechnol. 28 (5) (2010) 511–515.
[7] A.M. Mezlini, E.J. Smith, M. Fiume, O. Buske, G.L. Savich, S. Shah, et al., IReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data, Genome Res. 23 (3) (2013) 519–529.
[8] P.G. Engström, T. Steijger, B. Sipos, G.R. Grant, A. Kahles, G. Rätsch, et al., Systematic evaluation of spliced alignment programs for RNA-seq data, Nat. Meth. 10 (12) (2013) 1185–1191.
[9] H. Feng, Z. Qin, X. Zhang, Opportunities and methods for studying alternative splicing in cancer with RNA-Seq, Cancer Lett. 340 (2) (2013) 179–191.
[10] K. Kannan, L. Wang, J. Wang, M.M. Ittmann, W. Li, L. Yen, Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing, Proc. Natl. Acad. Sci. 108 (22) (2011) 9172–9177.

[11] D. Pflueger, S. Terry, A. Sboner, L. Habegger, R. Esgueva, P.C. Lin, et al., Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing, Genome Res. 21 (1) (2011) 56–67.
[12] S.A. Tomlins, B. Laxman, S. Varambally, X. Cao, J. Yu, B.E. Helgeson, et al., Role of the TMPRSS2-ERG gene fusion in prostate cancer, Neoplasia 10 (2) (2008) 177–188.
[13] S. Ren, Z. Peng, J.H. Mao, Y. Yu, C. Yin, X. Gao, et al., 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) (2012) 806–821.
[14] X. Xu, K. Zhu, F. Liu, Y. Wang, J. Shen, J. Jin, et al., Identification of somatic mutations in human prostate cancer by RNA-Seq, Gene 519 (2) (2013) 343–347.
[15] J.R. Prensner, M.K. Iyer, O.A. Balbin, S.M. Dhanasekaran, Q. Cao, J.C. Brenner, et al., Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression, Nat. Biotechnol. 29 (8) (2011) 742–749.
[16] Y. Zhang, E.W. Lameijer, P. AC't Hoen, Z. Ning, P.E. Slagboom, K. Ye, PASSion: a pattern growth algorithm-based pipeline for splice junction detection in paired-end RNA-Seq data, Bioinformatics 28 (4) (2012) 479–486.
[17] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S.L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, Genome Biol. 14 (4) (2013). R36.
[18] S. Shen, J.W. Park, Z.X. Lu, L. Lin, M.D. Henry, Y.N. Wu, Y. Xing, RMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data, Proc. Natl. Acad. Sci. 111 (51) (2014) E5593–E5601.
[19] K. Vitting-Seerup, B.T. Porse, A. Sandelin, J. Waage, SpliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data, BMC Bioinform. 15 (1) (2014) 81.
[20] A. Kasprzyk, BioMart: driving a paradigm change in biological data management, Database (2011) bar049.
[21] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, et al., Towards a knowledge-based human protein atlas, Nat. Biotechnol. 28 (12) (2010) 1248–1250.
[22] L.L. Xu, Y. Shi, G. Petrovics, C. Sun, M. Makarem, W. Zhang, et al., PMEPA1, an androgen-regulated NEDD4-binding protein, exhibits cell growth inhibitory function and decreased expression during prostate cancer progression, Cancer Res. 63 (15) (2003) 4299–4304.
[23] Y. Wang, V. Balan, X. Gao, P.G. Reddy, D. Kho, L. Tait, et al., The significance of galectin-3 as a new basal cell marker in prostate cancer, Cell Death Disease 4 (8) (2013) e753.
[24] P.G. Fournier, P. Juárez, G. Jiang, G.A. Clines, M. Niewolna, H.S. Kim, et al., The TGF-β signaling regulator PMEPA1 suppresses prostate cancer metastases to bone, Cancer Cell 27 (6) (2015) 809–821.
[25] X.H. Pei, F. Bai, Z. Li, M.D. Smith, G. Whitewolf, R. Jin, et al., Cytoplasmic CUL9/PARC ubiquitin ligase is a tumor suppressor and promotes p53-dependent apoptosis, Cancer Res. 71 (8) (2011) 2969–2977.
[26] B. Schwamb, R. Pick, S.B.M. Fernández, K. Völp, J. Heering, V. Dötsch, et al., FAM96A is a novel pro-apoptotic tumor suppressor in gastrointestinal stromal tumors, Int. J. Cancer 137 (6) (2015) 1318–1329.
[27] X. Chen, J. Wan, J. Liu, W. Xie, X. Diao, J. Xu, et al., Increased IL-17-producing cells correlate with poor survival and lymphangiogenesis in NSCLC patients, Lung cancer 69 (3) (2010) 348–354.
[28] M. Numasaki, M. Watanabe, T. Suzuki, H. Takahashi, A. Nakamura, F. McAllister, et al., IL-17 enhances the net angiogenic activity and in vivo growth of human non-small cell lung cancer in SCID mice through promoting CXCR-2-dependent angiogenesis, J. Immunol. 175 (9) (2005) 6177–6189.
[29] M. Sohda, Y. Misumi, K. Tashiro, M. Yamazaki, T. Saku, et al., Identification of a soluble isoform of human IL-17RA generated by alternative splicing, Cytokine 64 (3) (2013) 642–645.
[30] P. Rajan, J. Stockley, I.M. Sudbery, J.T. Fleming, A. Hedley, G. Kalna, et al., 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) (2014) 977.
[31] C.M. Farrell, N.A. O'Leary, R.A. Harte, J.E. Loveland, L.G. Wilming, C. Wallin, et al., Current status and new features of the consensus coding sequence database, Nucl. Acids Res. 42 (D1) (2014) D865–D872.
[32] National Center for Biotechnology Information, SRA Handbook, <http://www.ncbi.nlm.nih.gov/books/NBK47537>, 2010 (accessed July 10, 2015).
[33] Genome Reference Consortium, Genome reference consortium human build 37 patch release 10 (grch37.p10), <http://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.22>, 2012 (accessed July 10, 2015).
[34] NCBI Resource Coordinators, Database resources of the National Center for Biotechnology Information, Nucl. Acids Res. 42 (D1) (2014) D7–D17.
[35] N. Megiddo, K.J. Supowit, On the complexity of some common geometric location problems, SIAM J. Comput. 13 (1) (1984) 182–196.
[36] Matlab signal processing toolbox, <http://www.mathworks.com/help/signal/ref/findpeaks.html> (accessed July 10, 2015).